The Information Retrieval Series

Ryen W. White
Chirag Shah *Editors*

Information Access in the Era of Generative AI

Springer

The Information Retrieval Series

Volume 51

Information Retrieval (IR) deals with access to and search in mostly unstructured information, in text, audio, and/or video, either from one large file or spread over separate and diverse sources, in static storage devices as well as on streaming data. It is part of both computer and information science, and uses techniques from e.g. mathematics, statistics, machine learning, database management, or computational linguistics. Information Retrieval is often at the core of networked applications, web-based data management, or large-scale data analysis.

The Information Retrieval Series presents monographs, edited collections, and advanced text books on topics of interest for researchers in academia and industry alike. Its focus is on the timely publication of state-of-the-art results at the forefront of research and on theoretical foundations necessary to develop a deeper understanding of methods and approaches.

This series is abstracted/indexed in EI Compendex and Scopus.

Ryen W. White • Chirag Shah
Editors

Information Access in the Era of Generative AI

Springer

Editors
Ryen W. White
Microsoft Research
Redmond, WA, USA

Chirag Shah
University of Washington
Seattle, WA, USA

ISSN 1871-7500 ISSN 2730-6836 (electronic)
The Information Retrieval Series
ISBN 978-3-031-73146-4 ISBN 978-3-031-73147-1 (eBook)
https://doi.org/10.1007/978-3-031-73147-1

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

In recent years, Generative Artificial Intelligence (GenAI) has emerged as a groundbreaking technology that promises to revolutionize many industries and people's personal and professional lives. This book discusses GenAI and its role in information access, or more broadly, information *interaction*, both now and for decades to come. It is well known that information is the lifeblood of decision-making and action, and being able to help people find, understand, and use information is a central tenet of helping them attain their goals. The intersection of GenAI and information access is often referred to as Generative Information Retrieval (GenIR). We use "GenAI" and "GenIR", and variants thereof, where appropriate throughout the book, depending on the context.

The book is targeted to graduate students as well as advanced undergraduates and researchers interested in issues of information retrieval, access, and interactions, as well as applications of GenAI in various informational contexts. While some of the parts assume prior background in IR or AI, most others do not, making this book suitable for adoption in various classes as a primary source or as a supplementary material for a wide variety of curricula and training programs.

The role of GenAI in information access is complex and dynamic, with many dimensions. To address this, following our brief introduction to GenAI and GenIR (Chap. 1), we divide the remainder of the book into eight chapters, each targeting a different dimension or sub-topic. The chapters are described in more detail below. Each chapter is written by subject matter experts in the field and dives deep into the dimension at hand, presenting a general overview, a summary of the latest advances in that area, and a path forward.

Chapter 2: Foundations of Generative Information Retrieval—This chapter, authored by Qingyao Ai, Jingtao Zhan, and Yiqun Liu, delves into the core principles of GenIR, exploring the synthesis of results, validation and attribution processes, the phenomenon of hallucination in Large Language Models (LLMs), and the expansion beyond text to include multimedia and multimodal interactions with LLMs, Retrieval-Augmented Generation (RAG), and corpus understanding.

Chapter 3: Interactions with Generative Information Retrieval Systems—Authored by Mohammad Aliannejadim Jacek Gwizdka and Hamed Zamani, this chapter investigates the dynamics of user engagement with GenIR systems. It covers intent understanding, the art of querying, crafting new experiences, the intricacies of conversational systems, the role of agency and automation, the craft of prompt engineering, and the importance of explanations.

Chapter 4: Adapting Generative Information Retrieval Systems to Users, Tasks, and Scenarios—Johanne Trippas, Damiano Spina, and Falk Scholer discuss research on tailoring GenIR systems to fit individual users, specific tasks, and diverse scenarios. The chapter covers personalization, specialization, hybrid inference, and task understanding, reflecting the adaptability of GenIR systems.

Chapter 5: Improving Generative Information Retrieval Systems Based on User Feedback—Qingyao Ai, Zhicheng Dou, and Min Zhang take the helm in this chapter, focusing on the iterative improvement of GenIR systems through user feedback. It examines alignment, the integration of humans in the loop, continual learning, Reinforcement Learning from Human Feedback (RLHF), and the nuances of implicit and explicit feedback.

Chapter 6: Generative Information Retrieval Evaluation—Marwah Alaofi, Negar Arabzadeh, Charles Clarke, and Mark Sanderson provide a comprehensive overview of the evaluation metrics, methodologies, and the importance of reproducibility in GenIR systems. They also explore the connection to foundational IR evaluation frameworks such as Cranfield and TREC.

Chapter 7: Sociotechnical Implications of Generative Artificial Intelligence for Information Access—Bhaskar Mitra, Henriette Cramer, and Olya Gurevich discuss the societal implications of GenIR. They present an overview of systemic risks from the usage of GenAI in IR (e.g., negative information ecosystem impact, safety, and power dynamics), impact of evaluation methods, and ecosystem incentives around misuse.

Chapter 8: Recommendation in the Era of Generative Artificial Intelligence—Wenjie Wang, Yongfeng Zhang, and Tat-Seng Chua explore the realm of recommendations within GenIR. They delve into recommender systems, personalized recommendations, explainable recommendations, and the integration of LLMs with recommendation systems.

Chapter 9: Designing for the Future of Information Access with Generative Information Retrieval—Vanessa Murdock, Chia-Jung Lee, and William Hersh envision the future of information access through the lens of GenIR. They discuss new experiences such as proactive information access, emerging business models, applications beyond traditional information retrieval, the ubiquity of GenIR, and its application in specialized domains such as healthcare.

These chapter brief summaries provide a glimpse into the comprehensive coverage of GenIR topics that the book will offer. Each chapter promises to contribute valuable insights into the rapidly evolving field of GenAI, and GenIR in particular.

We hope that for you, the reader, this book creates clarity about the current state of the art in the fast-moving field of GenIR, piques your interest, is useful in your work or studies, and perhaps even inspires you to pursue your own scientific research in this important new area.

Redmond, WA, USA — Ryen W. White
Seattle, WA, USA — Chirag Shah

Contents

Chapter 1
Introduction

Ryen W. White and Chirag Shah

Abstract Information access systems, especially search engines and recommender systems, play a vital role in the access to information that is crucial for decision-making and action in the world. The emergence of Generative Artificial Intelligence (GenAI) has led to more advanced user experiences in these systems with natural user-system interactions and auto-generated answers and suggestions, potentially saving people time and cognitive effort, while improving task outcomes. This chapter explores the synergies between GenAI and information access and provides a framing for the rest of the book. GenAI technologies, such as transformers and large language models, have revolutionized various fields, including creative writing, software development, and multimodal content generation. We briefly discuss ongoing GenAI-related research in search and recommendation that is exploring areas such as generative document retrieval, grounded answer generation, generative recommendation, and generative knowledge graphs, enhancing the capabilities of information systems. We also cover other topics such as combining information interaction modalities (e.g., data types, interaction paradigms) in different ways to create unified, so-called "panmodal" GenAI-powered information experiences that leverage the strengths of different interaction modes and highlight the growing interest and collaboration in GenAI and its applications in information access. We conclude by discussing the ethical considerations and challenges that come from the rise of this new technology, emphasizing the need for responsible development and deployment to harness its potential while mitigating risks.

R. W. White (✉)
Microsoft Research, Redmond, WA, USA
e-mail: ryenw@microsoft.com

C. Shah
University of Washington, Seattle, WA, USA
e-mail: chirags@uw.edu

R. W. White, C. Shah (eds.), *Information Access in the Era of Generative AI*,
The Information Retrieval Series 51, https://doi.org/10.1007/978-3-031-73147-1_1

1.1 The Importance of Information Access

Information is critical for decision-making and action in the world. Information systems play an important role in facilitating access to that information or *information interaction* more broadly [1]. Search engines are the most commonly used means of information access, serving results for billions of queries on a daily basis. These systems have grown dramatically in scale and complexity given the advent of the Web, but many still resemble the original Information Retrieval (IR) systems, with an index of documents, a user-defined text query, and an algorithmic matching process [2]. The interaction model with search engines is well documented and studied: user queries submitted to the search engine, the retrieval of lists of results, result clicks, visits to landing pages, and subsequent query refinements, pagination through result pages, etc. Interactions with search systems have been used as a means of improving search systems over time via user feedback, at an individual, group, and population level. There are also other interaction models such as faceted search [3], which allows users to filter results based on different facets or attributes, e.g., in an e-commerce setting.

These interactions put the user in control of much of the search process, which also relies on their searching skills (e.g., search engine responses are highly dependent on how queries are formulated) and are also affected by cognitive and content biases of various types. Systems do not need to wait for user requests; they can also recommend content dynamically based on the available context (e.g., geolocation, Web browsing histories, application usage) or the behavior of other users in collaborative filtering scenarios.

Advances in neural IR in recent years have meant that queries and documents can now be represented semantically rather than syntactically, improving the quality of search engine responses. Advances in conversational IR have also made interactions with search systems more natural over time, more like a human conversation than direct engagement with a system. We are now seeing the emergence of a new wave of search advances powered by GenAI, where information systems can index content more efficiently (e.g., with differentiable search indices), respond directly to searches with comprehensive AI-generated answers, and can make real-time inferences from multimodal contexts such as video streams, among other data types (e.g., images, audio, sensor data). The intersection of GenAI and information access has been referred to as Generative Information Retrieval (GenIR) [4]. This type of AI-powered assistance can be especially useful for complex tasks, where people are often faced with multiple steps and multi-faceted intents, a need to engage with many different resources, and may struggle to be successful [5].

The era of GenAI is creating incredible new opportunities for information access. Before outlining some of these and future directions, we provide a brief overview of GenAI technologies and how this technology is reshaping the information landscape across many industries.

1.2 Emergence of Generative AI

GenAI has seen rapid advancements and holds massive potential for various applications. The technology is powered by a transformer architecture [6], a type of neural network architecture that has revolutionized natural language processing and a growing number of different aspects of information access. Transformers are designed to handle sequential data, like text, and are particularly effective for tasks that involve understanding the context and relationships between words in a sentence, within a paragraph, or even across documents.

Transformers use a mechanism called self-attention, which allows them to weigh the importance of different parts of the input data differently. This is crucial for tasks such as language translation, where the meaning of a word can depend heavily on the words around it. The architecture of transformers is highly parallelizable, making it efficient for training on large datasets.

Next token prediction is a task where a model is trained to predict the probability of the next word in a sequence, given the words that precede it. This is a fundamental task for language models such as GPT (Generative Pretrained Transformer) [7, 8], which learn to generate text by predicting one word at a time. This approach has led to the creation of models that can write essays, summarize text, translate languages, and even generate code.

The progress in this area has been swift, with models becoming increasingly large and complex. For instance, Large Language Models (LLMs) such as GPT-4 from OpenAI have been trained on vast amounts of text data and can perform a wide range of tasks without task-specific training. This demonstrates the flexibility and potential of GenAI to impact various fields, from creative writing to software development. ChatGPT from OpenAI provides a conversational interface to these LLMs, and SearchGPT, also from OpenAI, provides a combination of search engine features and GenAI capabilities to provide users with fast, timely answers from clear and relevant sources. Small Language Models (SLMs) such as Phi-3 from Microsoft [9] and Gemini Nano from Google [10] are also emerging that are more scalable than LLMs (they can be run client side and shipped on device) while preserving much of the accuracy of the larger models.

Beyond text, diffusion models such as DALL-E[1] are another exciting area of GenAI. These models simulate the process of spreading substances from areas of higher concentration to lower concentration, which can be applied to various fields beyond physics, such as finance and marketing. In AI, diffusion models have been used for tasks like image generation and enhancing the realism and compositionality of generated content. There is also the recent development of multimodal models such as GPT-4o[2] and Gemini[3] that integrate text, vision, and audio capabilities.

[1] https://openai.com/index/dall-e-3

[2] Hello GPT-4o | OpenAI

[3] Gemini - Google DeepMind

These models represent a significant leap forward in AI's ability to understand and generate content across different modalities. Multimodal models such as these can be used to understand the context around which information is requested by users (e.g., "what am I looking at [through my smart glasses] right now") or proffered by systems.

Generative Adversarial Networks (GANs) [11] are a class of artificial intelligence algorithms used in unsupervised machine learning. They consist of two neural networks, the generator and the discriminator, which are trained simultaneously through adversarial processes. The generator creates data that is indistinguishable from real data, while the discriminator evaluates the authenticity of the data, effectively learning to distinguish between real and generated data. This dynamic training process allows GANs to generate high-quality, realistic data, which has applications in various fields such as image and video enhancement, gaming, and even the creation of artificial human faces.

However, with great power comes great responsibility. The potential misuse of GenAI, such as in creating deep fakes or spreading misinformation, raises ethical concerns. There are also significant reliability concerns given GenAI's tendency to hallucinate. It is crucial to continue research and development with a focus on factuality, safety, transparency, and fairness to ensure that the benefits of GenAI are realized while minimizing its risks. There are other pertinent issues such as human control/agency, which are reduced with GenAI over some aspects of the search process, namely result inspection and answer synthesis, while users have more/different control over the specifications of their intent via longer text prompts and multi-turn dialog, and provenance, where creator attribution and direct access to the source must be provided to improve user trust and ensure the continued willingness of content creators and publishers to provide their material for GenAI training. This will help us to avoid the "paradox of reuse" [12], where fewer visits to online content results in less content being created and progressively worse models over time.

In summary, GenAI is a rapidly evolving field with the potential to significantly impact technology and society, and information access will certainly be part of this. GenAI's ability to understand and generate human-like text opens up new possibilities for information access, task automation, and creativity, but it also necessitates careful consideration of its implications. In the next section, we focus on synergies between GenAI and information access, and the growth in popularity of the intersections between these fields in the IR research community in particular.

1.3 Generative AI and Information Access Synergies

We strongly believe that AI should be used to augment, amplify, and empower people, not replace them. The domain of information access is no different. By being used in the right way, GenAI can supercharge information access and help more people get answers and complete complex tasks more effectively.

With great responsibility comes great power. If we deploy these GenAI systems in a responsible way, i.e., provide fairness, transparency, and accountability, we could make these systems much more powerful to help users with varying backgrounds, skills, and literacy. Responsible integration of GenAI technologies in information access systems also means a user-first approach – focusing on user needs and contexts and building solutions around them.

1.3.1 Ongoing Research

The community has been exploring a few ways that GenAI can affect information access, namely generative document retrieval, grounded answer generation, generative recommendation, and generative knowledge graphs. We will introduce each of these in turn.

Generative Document Retrieval (GDR) is an emerging paradigm in IR that has gained significant attention in recent years. It involves learning to build connections between documents and identifiers within a single model. This approach enables retrieval by directly generating relevant document identifiers without explicit indexing, offering more flexibility, efficiency, and creativity (boosting recall). So-called *differentiable search indices* [13] integrate the different stages of search indexing into a single, end-to-end trainable neural model. This allows for the entire retrieval process to be parameterized and optimized as part of the model's learning, making it a differentiable component that can be fine-tuned using gradient descent methods. GDR can be particularly useful in scenarios such as search engines, question answering, and recommendation systems, where traditional retrieval methods based on similarity matching may fall short.

Grounded Answer Generation (GAG) is a sophisticated approach within conversational AI that aims to provide accurate and contextually relevant answers by grounding responses in verified information sources. Systems such as ChatGPT, Gemini, and Copilot (the assistive AI agent from Microsoft) engage users in multi-turn dialog to help them complete their tasks. In conversational IR, the context of the dialog and the interdependencies between questions and answers play a significant role in understanding and generating responses. The Retrieval-Augmented Generation (RAG) [14] framework is a prominent example of GAG, where an LLM is augmented with external knowledge retrieved from databases and/or search engines to generate responses that are not only relevant but also grounded in source material. These sources can be shown to users as hyperlinked references for provenance purposes and to help build user trust in system responses. Retrieval-Augmented Fine Tuning (RAFT) [15, 16] combines the strengths of RAG and fine-tuning by integrating retrieval-based methods with generation to access external knowledge during response generation, and then further training the model on task-specific data to optimize its performance on the target task. In the enterprise context, grounded answer generation is being explored to enhance the capabilities of conversational agents and copilots. Google provides "AI Overviews," GenAI-

powered answers shown inline on the search engine result page as part of Google's so-called search generative experience.[4] Bing also now offers a generative search experience.[5]

Generative Recommendation represents a cutting-edge approach in the field of recommender systems, leveraging the power of GenAI to create personalized content that caters to the unique preferences and needs of users [17]. Methods such as GenRec [18] also emphasize the trustworthiness of the generated items through various fidelity checks. Unlike traditional methods, which retrieve existing relevant items from a corpus, this paradigm shift to generative techniques where new content is created for recommendation allows for a more dynamic and tailored user experience.

Generative Knowledge Graphs (KGs) are an innovative approach to enhancing AI models with structured world knowledge. They enable AI systems to generate new knowledge entities and relationships, enriching the existing data and potentially uncovering new insights. There are a range of different methods to do this, including MoKGE [19], which diversifies generative reasoning using a mixture of experts (MoE) strategy on commonsense knowledge graphs, and COMET [20], which generates rich and diverse commonsense descriptions in natural language. Generative KGs are being recognized for their potential to transform how we interact with and utilize large datasets.

All of this research is pivotal in designing intelligent information systems that can understand intent, personalize interactions, and provide generative responses that are both accurate and trustworthy.

1.3.2 Combining Modalities

To be useful to users, the technologies described above need to be used in interactive experiences. Information interaction modalities describe the different modes of interaction with information systems. These modes can be defined in many ways, including the interaction paradigm (e.g., query-response, multi-turn dialog, proactive suggestions), but also different input mechanisms (e.g., text, speech, touch) and different device types.

GenAI methods used in isolation to enhance existing experiences, for example, GDR can enhance search and GAG can enhance chatbots. We refer to staying within a single interaction modality (either because it is the only one available, the only one selected by the user, or it is the only one feasible in the current context) as *monomodal information interaction*. However, modalities can also be combined to create "better together," unified experiences that combine the strengths of the different modalities depending on the task and/or task stage. We refer to

[4] Google I/O 2024: New generative AI experiences in Search (blog.google)

[5] Introducing Bing generative search | Bing Search Blog

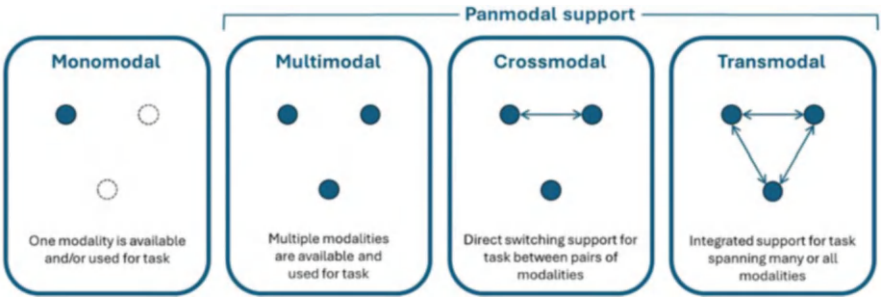

Fig. 1.1 Different aspects of panmodal support and its relation to monomodal support

these as *panmodal information interactions* [21], spanning at least three types: multimodal (make several modalities available), crossmodal (support transitions between modalities), and transmodal (use several modalities seamlessly to complete a task). Figure 1.1 illustrates different aspects of panmodality.

We are already seeing the genesis of multimodal experiences that combine modalities with search and chat coexisting within the same experience and GenAI-based answers (e.g., Google's AI Overviews, Bing's generative search) generated dynamically and embedded within search engine results pages. Retrieval systems have supported some aspects of panmodality since the 1980s (e.g., the I^3R system [22]), and recent research has examined the challenge of directing users to the search engine or computing device best placed to tackle their task. Recent work has also explored combining different devices simultaneously to take advantage of the capabilities of each device type, e.g., multi-device experiences comprising smart speakers (far-field audio capabilities) plus tablets (high-resolution displays), all mediated by services running in the cloud [23].

There are additional information interaction modalities beyond search and chat, e.g., bespoke interfaces generated natively by GenAI for the task at hand, interactive visualizations generated by GenAI akin to dynamic queries, proactive recommendations from GenAI based on audio and vision sensing, and moving from information access to information use in GenAI-based operation or suggestion of tools to support task completion. This will also expand beyond interaction paradigms into new modes of interaction (e.g., tactition, gesture, eye gaze, or eye contact with devices or application windows) and new device types (e.g., smart rings, smart glasses), where GenAI could help interpret signals and add intelligence. Identifying the most fitting type of panmodality support can act as helpful prompts for users, even if their direct use is not integrated within the application or if there's no seamless transition to other applications. For instance, in a scenario where transmodal support is ideal, GenAI could outline possible modalities and suggest a sequence of operations according to a generated plan. This approach could also serve as a means to increase user awareness of the various modalities at their disposal, thereby gradually shaping their usage patterns to more effectively leverage that support independently over time.

1.3.3 Community Engagement

GenAI has been the subject of many recent keynotes at premier conferences on search, recommendation, and knowledge management [24–26], highlighting the broad appeal and general interest in work in this area. Recent workshops and special sessions are indicative of the burgeoning interest in GenAI and its applications in information access. The first "GenIR" workshop at the ACM SIGIR Conference on Research and Development on Information Retrieval in July 2023 [27] and its subsequent second edition at the same conference in July 2024 highlight the IR community's ongoing engagement with GenAI. This workshop invited discussions on models, training, evaluation, and applications of retrieval-oriented applications of GenAI, reflecting the field's dynamic nature and its potential for innovation. A workshop at Microsoft Research in September 2023 [28] discussed the latest developments in GenAI with a focus on the challenges and opportunities in task-oriented applications. A special session on LLMs at the ACM CHIIR Conference on Human Information Interaction and Retrieval in March 2024 provided a platform for exploring the impact of large language models on human-computer interaction and IR, underscoring the importance of user-centered approaches in the development of AI technologies. The "LLM Day" at the ACM SIGIR Conference on Research and Development on Information Retrieval in July 2024 was a dedicated event to delve into the advancements and challenges associated with large language models, fostering a deeper understanding of their capabilities and limitations in search and retrieval contexts. These events serve not only as a testament to the rapid progress in GenAI but also as a catalyst for further research, collaboration, and knowledge sharing within the field. We expect that the number of such events will continue in the coming years, with GenAI becoming a mainstay of conferences where information access is discussed.

1.4 Emerging Trends

GenAI will democratize access to information and enable more people to tackle a broader range of tasks. There are some emerging trends in this area that we should not ignore. These span technological advances (covered in this section) and challenges (covered in the next section).

AI agents powered by GenAI will help users tackle their tasks and attain their goals. We also expect to see AI agents working together—in multi-agent systems such as CAMEL [29] and AutoGen [30]—and with humans (e.g., via conversation) to facilitate task completion, with users having full visibility into and control over the process. These agents will also take action on behalf of users, with user consent, and optionally, oversight and control/feedback, e.g., operating digital applications such as reservations systems. This move toward task automation also signifies

progress along the task lifecycle from systems supporting information access to systems now supporting both information access *and* information use.

We will also see more use of the contextual understanding and environment recognition facilitated by multimodal models such as GPT-4o and Gemini to better ground user requests. Applications such as Google Astra[6] are examples of future information access systems that possess sophisticated reasoning, planning, and memory skills, and leverage GenAI to help users complete tasks. Astra can process video and auto input in real time and use that to answer questions about the environment ("what is that type of plant"), a user's lost items ("where did I leave my keys"), and much more. This type of situational awareness fits well with emerging personal devices trends such as augmented reality and smart glasses.

Systems will be imbued with near-infinite memory, which will allow them to develop personalized intelligence (where deep knowledge of a user will become a significant differentiator for one assistant over another and engender user loyalty to a specific assistant), more accurately tailoring the answers they provide to the interests and intentions of the current user, and further personalize user experiences by learning from user interactions and preferences over time, leading to more intuitive and anticipatory support systems.

A more complete personal and contextual understanding will also allow future information access systems to offer proactive experiences and recommendations, e.g., during Web browsing, given an event trigger such as an incoming phone call/email or change in location, or as situated interactions while moving around the physical world.

These advances must consider the broader landscape of the need for energy-efficient computational methods, growing consumer and enterprise concerns about privacy and security, and the wide availability of specialized hardware for AI acceleration (such as graphics processing units (GPUs), neural processing units (NPUs), and tensor processing units (TPUs)) in mainstream devices, e.g., NPUs in Microsoft Copilot+ PCs and TPUs in Google Pixel smartphones. This may mean that most of these advancements will be implemented on-device, powered by SLMs not LLMs, or leverage a hybrid architecture where SLMs are used for simple tasks and LLMs in the cloud are used for more complex task workloads. LLM routing, directing user (or agent) requests to the most appropriate language model, whether small or large, given constraints, will emerge as an important direction in general.

1.5 Challenges

There are a range of sociotechnical implications that must be considered as the technology advances. To address these challenges, it is essential to develop ethical guidelines and governance frameworks that can guide the responsible use of GenAI.

[6] Project Astra - Google DeepMind

One of the primary concerns is the reliability of these systems, particularly the occurrence of hallucinations or the generation of unfaithful content, which has been seen in tools such as Google AI Overviews (e.g., recent press coverage about how these overviews suggest to users that they eat one rock per day to stay healthy[7]). These issues highlight the need for robust mechanisms to ensure the accuracy and trustworthiness of GenAI outputs.

Another significant challenge is the potential loss of human control and agency [31]. The human factors and information science communities have discussed the boundaries between humans and systems for many decades [32]. We must retain a continued focus on human-AI cooperation, where searchers stay in control while the degree of system support increases as needed [33]. As AI systems become more autonomous, there is a risk that users may become overly reliant on these technologies, leading to a decrease in human oversight and the ability to intervene when necessary. This is compounded by missed opportunities for serendipitous information encounters (since the system is synthesizing answers), which can lead to a homogenization of information and a reduction in the diversity of content encountered and generated by users. User education in fostering an understanding of AI capabilities and limitations is also essential for building trust and ensuring effective collaboration between humans and AI agents.

Finally, a potential side effect of relying on AI-generated answers for information needs is depriving the user of the opportunity to learn and discover [34]. As many scholars (e.g., [35–37]) have pointed out, searching is often more than simply finding information; it is also an opportunity to learn. "Searching as learning" subfield in IR has evolved over the years to explicitly acknowledge and support such possibilities. The learning here happens through the user actively engaging with formulating or reformulating queries, assessing results, and re-examining their needs. When we cut short that process, we may be taking away not only the user's ability to learn but also their opportunities for discovery, serendipity, as well as other key activities such as critical thinking. These may be desirable characteristics in many situations. The increasing push toward automating complex tasks via AI agents that can interact with applications and Web sites on behalf of humans (e.g., the UFO system [15, 16]) raises similar concerns about risks to human cognition (and control, etc.). Agents can provide cognitive scaffolding to gradually help people reflect and learn how to perform complex tasks independently over time, e.g., by offering hints and structured cues.

Evaluating the performance of GenAI systems is also a challenging task. Traditional metrics may not fully capture the nuances of generative content, necessitating the development of new evaluation frameworks that can account for the unique characteristics of AI-generated information and output/experiences that may also be non-deterministic. GenAI can also play a vital role in evaluating information systems, working together with human judges to assess the retrieval performance of these systems.

[7] Google AI Overviews Search Errors Cause Furor Online - The New York Times (nytimes.com)

Bias and toxicity in AI-generated content are additional concerns, as GenAI systems often reflect the biases present in their training data. This can lead to the perpetuation of stereotypes and discriminatory content, which is harmful and undermines the credibility of the AI system. Provenance, or the ability to trace the origin of AI-generated content, is another area that requires attention. Ensuring that users can identify the source of information and understand the process by which it was generated is crucial for transparency and trust (and, as mentioned earlier, as a way to drive traffic and revenue to content creators and publishers).

As AI agents become more autonomous and capable of taking actions on behalf of users, it is crucial to ensure that these systems operate within ethical boundaries and have mechanisms in place to prevent misuse. Furthermore, the integration of GenAI in digital applications such as reservation systems points to the need for interoperability standards to ensure seamless interaction between different systems and platforms. There are challenges of data privacy and security in the context of AI systems with near-infinite memory, emphasizing the need for secure data handling practices and user consent mechanisms.

The rapid integration of GenAI in search and recommendation systems also prompts a reassessment of IR's core research focus and the adoption of a design approach that aligns with user and societal needs for information access [38]. The advancement of GenAI will benefit from the collaboration of disciplines, including computer science, social sciences, and humanities. This interdisciplinary approach can provide a more holistic understanding of the impact of GenAI on society and contribute to the development of more effective human-centric systems, including information access systems.

References

1. White, R.W.: Interactions with search systems. Cambridge University Press (2016)
2. van Rijsbergen, C.J.: Information retrieval. Butterworth-Heinemann (1979)
3. Hearst, M.A.: Clustering versus faceted categories for information exploration. Commun. ACM. **49**(4), 59–61 (2006)
4. Metzler, D., Tay, Y., Bahri, D., Najork, M.: Rethinking search: Making domain experts out of dilettantes. ACM SIGIR Forum. **55**(1), 1–27 (2021)
5. White, R.W.: Advancing the search frontier with AI agents. Commun. ACM. **67**(9), 54–65 (2024)
6. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Adv. Neural Inf. Proces. Syst. **30** (2017)
7. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., et al.: Language models are few-shot learners. Adv. Neural Inf. Proces. Syst. **33**, 1877–1901 (2020)
8. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training. OpenAI (2018). https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf
9. Abdin, M., Jacobs, S. A., Awan, A. A., Aneja, J., Awadallah, A., Awadalla, H., Bach, N. et al. Phi-3 technical report: A highly capable language model locally on your phone. arXiv preprint arXiv:2404.14219 (2024)

10. Gemini Team, Anil, R., Borgeaud, S., Wu, Y., Alayrac, J-B., Yu, J., Soricut, R. et al.: Gemini: A family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
11. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Bing, X., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Adv. Neural Inf. Proces. Syst. **27**, 2672–2680 (2014)
12. Vincent, N.: The paradox of reuse, language models edition (2022). https://nmvg.mataroa.blog/blog/the-paradox-of-reuse-language-modelsedition/. Accessed: 2024-07-27
13. Tay, Y., Tran, V., Dehghani, M., Ni, J., Bahri, D., Mehta, H., Qin, Z., et al.: Transformer memory as a differentiable search index. Adv. Neural Inf. Proces. Syst. **35**, 21831–21843 (2022)
14. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. Adv. Neural Inf. Proces. Syst. **33**, 9459–9474 (2020)
15. Zhang, C., Li, L., He, S., Zhang, X., Qiao, B., Qin, S., Ma, M. et al.: Ufo: A ui-focused agent for windows os interaction. arXiv preprint arXiv:2402.07939 (2024a)
16. Zhang, T., Patil, S. G., Jain, N., Shen, S., Zaharia, M., Stoica, I., Gonzalez, J. E.: Raft: Adapting language model to domain specific rag. arXiv preprint arXiv:2403.10131 (2024b)
17. Rajput, S., Mehta, N., Singh, A., Keshavan, R.H., Trung, V., Heldt, L., Hong, L., et al.: Recommender systems with generative retrieval. Adv. Neural Inf. Proces. Syst. **36** (2024)
18. Ji, J., Li, Z., Xu, S., Hua, W., Ge, Y., Tan, J., Zhang, Y.: Genrec: Large language model for generative recommendation. In Proceedings of the European conference on information retrieval, 494–502 (2024)
19. Yu, W., Zhu, C., Qin, L., Zhang, Z., Zhao, T., Jiang, M.: Diversifying content generation for commonsense reasoning with mixture of knowledge graph experts. arXiv preprint arXiv:2203.07285 (2022)
20. Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Celikyilmaz, A., Choi, Y.: COMET: Commonsense transformers for automatic knowledge graph construction. arXiv preprint arXiv:1906.05317 (2019)
21. Shah, C., White, R. W.: Panmodal information interaction. Under review at CACM (2024)
22. Bruce Croft, W., Thompson, R.H.: I^3R: A new approach to the design of document retrieval systems. J. Am. Soc. Inf. Sci. **38**(6), 389–404 (1987)
23. White, R.W., Fourney, A., Herring, A., Bennett, P.N., Chandrasekaran, N., Sim, R., Nouri, E., Encarnación, M.J.: Multi-device digital assistance. Commun. ACM. **62**(10), 28–31 (2019)
24. Najork, M.: Generative information retrieval. In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, 1–1 (2023)
25. Shah, C.: Generative AI and the Future of information access. In Proceedings of the 32nd ACM international conference on information and knowledge management, 3–3 (2023)
26. White, R.W.: Tasks, copilots, and the future of search. In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, 5–6 (2023)
27. Bénédict, G., Zhang, R., Metzler, D.: Gen-IR@ SIGIR 2023: The first workshop on generative information retrieval. In Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval, pp. 3460–3463 (2023)
28. Shah, C., White, R.W.: Report on the 1st workshop on task focused IR in the era of generative AI. ACM SIGIR Forum. **57**(2), 1–8 (2023)
29. Li, G., Hammoud, H., Itani, H., Khizbullin, D., Ghanem, B.: Camel: Communicative agents for "mind" exploration of large language model society. Adv. Neural Inf. Proces. Syst. **36**, 51991–52008 (2023)
30. Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., Awadallah, A. H., White, R. W., Burger, D., Wang, C.: Autogen: Enabling next-gen llm applications via multi-agent conversation framework. arXiv preprint arXiv:2308.08155 (2023)
31. Shah, C., Bender, E.M.: Situating search. In Proceedings of the 2022 ACM CHIIR Conference on Human Information Interaction and Retrieval, 221–232 (2022)
32. Bates, M.J.: Where should the person stop and the information search interface start? Inf. Process. Manag. **26**(5), 575–591 (1990)

33. Shneiderman, B.: Human-centered AI. Oxford University Press (2022)
34. Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., Mariman, R.: Generative AI can harm learning (2024). Available at SSRN: https://ssrn.com/abstract=4895486 or doi:https://doi.org/10.2139/ssrn.4895486
35. Ghosh, S., Rath, M., Shah, C.: Searching as learning: Exploring search behavior and learning outcomes in learning-related tasks. Proceedings of the 2018 ACM CHIIR Conference on Human Information Interaction and Retrieval, 22–31 (2018)
36. Rieh, S.Y., Collins-Thompson, K., Hansen, P., Lee, H.-J.: Towards searching as a learning process: A review of current perspectives and future directions. J. Inf. Sci. **42**(1), 19–34 (2016)
37. Vakkari, P.: Searching as learning: A systematization based on literature. J. Inf. Sci. **42**(1), 7–18 (2016)
38. Shah, C., Bender, E.M.: Envisioning information access systems: What makes for good tools and a healthy web? ACM Trans. Web. **18**(3), 1–24 (2024)

Chapter 2
Foundations of Generative Information Retrieval

Qingyao Ai, Jingtao Zhan, and Yiqun Liu

Abstract The chapter discusses the foundational impact of modern generative Artificial Intelligence (AI) models on Information Access (IA) systems. In contrast to traditional AI, the large-scale training and superior data modeling of generative AI models enable them to produce high-quality, human-like responses, which bring brand new opportunities for the development of IA paradigms. In this chapter, we identify and introduce two of them in detail, i.e., information generation and information synthesis. Information generation allows AI to create tailored content addressing user needs directly, enhancing user experience with immediate, relevant outputs. Information synthesis leverages the ability of generative AI to integrate and reorganize existing information, providing grounded responses and mitigating issues like model hallucination, which is particularly valuable in scenarios requiring precision and external knowledge. This chapter delves into the foundational aspects of generative models, including architecture, scaling, and training, and discusses their applications in multi-modal scenarios. Additionally, it examines the retrieval-augmented generation paradigm and other methods for corpus modeling and understanding, demonstrating how generative AI can enhance information access systems. It also summarizes potential challenges and fruitful directions for future studies.

The primary distinction between modern generative models and traditional AI techniques lies in their capability to generate complicated and high-quality output based on human instructions. As shown by many studies [1–3], modern generative AI models possess remarkable abilities to generate responses that closely mimic human interaction. Generally speaking, such impressive performance comes from

Qingyao Ai and Jingtao Zhan contributed equally to this work.

Q. Ai (✉) · J. Zhan · Y. Liu
Department of Computer Science and Technology, Tsinghua University, Beijing, China
e-mail: aiqy@tsinghua.edu.cn; zhanjt20@mails.tsinghua.edu.cn; yiqunliu@tsinghua.edu.cn

R. W. White, C. Shah (eds.), *Information Access in the Era of Generative AI*,
The Information Retrieval Series 51, https://doi.org/10.1007/978-3-031-73147-1_2

their large-scale training collections and their advanced data modeling algorithms. Their superior data understanding ability can benefit almost every components of existing information access systems, from document encoding and index construction to query processing and relevance analysis, etc. However, when talking about new opportunities or paradigms that are uniquely brought by the generative AI to information access, they can be broadly categorized in two directions. The first one is to create content that directly addresses the user's information needs. By understanding and taking user queries as input instructions, generative AI models are able to generate specific answers or products tailored to the individual's request. This direct approach to information generation can significantly enhance user experience by providing immediate and relevant responses. The second direction is to leverage the advanced instruction-following capabilities of generative AI models to synthesize and recombine existing information in innovative ways. Generative AI such as Large Language Models (LLMs) can take existing data and transform it into new, coherent pieces of information that may not have been explicitly outlined before. This ability to reinterpret and organize information opens up new possibilities for retrieval system design and applications. Therefore, in this chapter, we discuss how generative AI models could help information access from two perspectives, namely, *information generation* and *information synthesis*.

2.1 Information Generation

Information needs are diverse and typically long-tail. Traditional information retrieval systems, such as search engines and recommendation platforms, are designed to present information that already exists. However, these systems often fall short when it comes to fulfilling the less common information needs. This is particularly evident in scenarios requiring creative creation, where users seek not just information but inspiration and novel ideas. The limitations of traditional information systems in addressing these unique demands have paved the way for the emergence of generative models, which hold the promise of creating new information that aligns closely with long-tail information needs.

In recent years, generative models have made significant developments. For instance, ChatGPT can respond to user questions, Bing enhances its responses with retrieval-augmented generation, Midjourney generates images based on user prompts, and recommendation systems generate personal contents for different users. The development is mainly driven by the capable model architectures, computational resources, and large-scale Internet data. These elements have facilitated the performance of generative models to new heights. With the continuous efforts on scaling up these elements, the model performance is still rapidly improving. Nowadays, generative models have gradually been integrated into various workflows and everyday life activities.

In this section, we present the foundation of generative models. This section is organized as follows: Sect. 2.1.1 shows the efforts on designing the model archi-

tectures for LLMs. Section 2.1.2 discusses how scaling facilitates the development of generative models and its potential future. Section 2.1.3 presents the different training stages of LLMs. Finally, Sect. 2.1.4 introduces how LLMs are used in multimodal scenarios.

2.1.1 Model Architecture

In different generation scenarios like ChatGPT or SoRA, the transformer [4] has emerged as the predominant model structure. It starts with an embedding layer, followed by multiple neural layers. Within each layer, an attention mechanism models the interactions between words, creating contextualized embeddings. The final decision on word generation probabilities is derived by comparing the output embedding with the vocabulary embeddings. We illustrate the model architecture in Fig. 2.1. Unlike traditional recurrent neural networks [5], Transformers are capable of modeling long-distance interactions between words directly, which provides a more powerful representational capability. Numerous enhancements to the transformer architecture have been proposed. In the following, we will explore various modifications to each component of the Transformer, highlighting the advancements that have further improved its efficacy and efficiency.

2.1.1.1 Word Embedding

Word embedding module is at the bottom of the Transformer architecture. Initially, a tokenizer breaks down a sentence into tokens, which the Word embedding module then maps into embeddings. These are combined with position embeddings and fed into subsequent neural layers. Recent research on large-scale language models has

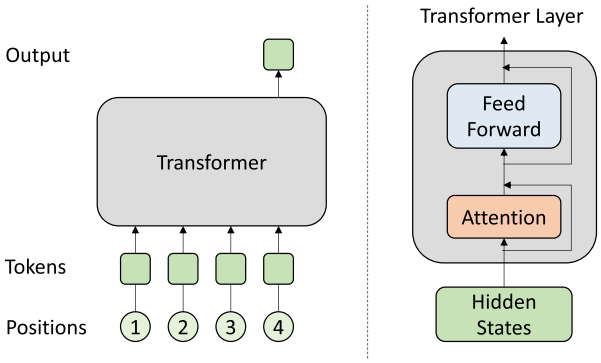

Fig. 2.1 Transformer architecture: overview on the left and the illustration of one layer on the right [4]

identified word embeddings as one of the main sources to training instability [6]. Particularly in the early stages of training, the gradients of word embeddings are often orders of magnitude larger than those of other parameters. To address this issue, [7] introduced a Layer Normalization (LN) immediately after the word embedding layer, stabilizing the distribution effectively. Besides, [6] opted to scale down the gradients of the word embeddings by an order of magnitude to prevent substantial updates. Both approaches have been proven effective in stabilizing the training of language models at the 100 billion parameter scale. Yet, whether they are still effective for larger models remains to be investigated.

2.1.1.2 Position Embedding

Position embedding is essential for Transformers. Unlike RNNs, which inherently process sequences in order, the vanilla attention mechanism disregards the positional distances between words, and the Transformer has to rely on position embeddings for position modeling. Initially, the Transformer [4] utilized sinusoidal embeddings, a non-trainable form of position embedding that is added directly to word embeddings. Later, [8] introduced trainable position embeddings, which is initialized randomly and are updated through gradient descent during training. Subsequently, [9] and [10] proposed relative positioning, where the attention mechanism incorporates biases based on the relative positions of words to better model varying distances. Recently, [11] introduced the concept of rope position embedding, based on the principle that the dot product of vectors correlates with their magnitudes and the angles between them. By rotating vectors in space proportionally to their positions, this method naturally integrates positional information into attention scores. [12] has found that this approach outperforms trainable position embeddings. Yet, these approaches may not work well when extrapolated to long sequences, and more effective methods need to be explored.

2.1.1.3 Attention

The attention mechanism models interactions between words and is a significant component of the Transformer architecture. Enhancements to the attention module have predominantly focused on two aspects: modeling long texts and optimizing the Key-Value (KV) cache. (1) Modeling Long Texts: The vanilla attention mechanism has a complexity of $O(n^2)$, which significantly increases computational costs for long texts. To address this, the Sparse Transformer [13] employs sparse attention, utilizing pre-designed attention patterns to avoid the computation of attention over long sequences. Another approach, Reformer [14], uses Locality-Sensitive Hashing (LSH) to reduce computational complexity. Additionally, [15] compressed context information to shorten sequences, thereby reducing overhead. Others have explored retrieval-based methods [16, 17]. This area of research continues to hold considerable potential for future advancements. (2) Optimizing KV Cache: classic

Transformers use multi-head attention (MHA), which requires storing extensive key-value caches during inference, slowing down model generation. To mitigate this, [18] proposed Multi-Query Attention (MQA), which employs multiple key heads but only a single value head, substantially reducing the key-value cache and enhancing computational speed. However, [19] found that this could degrade model performance, leading to the development of grouped query attention. This method allows multiple key heads to share a single value head, effectively serving as a hybrid between MQA and MHA, balancing computational complexity and performance more effectively. Recently, [20] introduced multi-head latent attention, which compresses keys and values into a single latent space, thereby reducing the key-value cache while maintaining robust representational capacity.

2.1.1.4 Layer Normalization

Layer normalization (LayerNorm) is important for stabilizing the distribution of hidden states, a key to train large language models. In the classical Transformer architecture, LayerNorm is positioned between residual blocks, hence termed Post-LN. Researchers [21] observed that this configuration could lead to high gradients near the output layers and very small gradients near the input layers, resulting in unstable gradients and challenging training dynamics. To address this issue, the Pre-LN configuration was proposed [21], placing LayerNorm on the residual pathways before attention or Feed-Forward Network (FFN) module. Experiments have shown that this adjustment leads to more uniform gradient distribution. Building upon Pre-LN, other researchers introduced Sandwich-LN [22], which adds an additional LayerNorm at the output of the residual pathways, further enhancing the training stability. Beyond merely adjusting the position of LayerNorm, researchers have developed DeepNorm [23], which combines a tailored parameter initialization strategy with modified residual connections to stabilize training. This approach enables the training of Transformers with depths reaching up to 1000 layers. Nevertheless, there still lacks a theoretical understanding about how layer normalization affects the training stability, and more work needs to be done for scaling the model even further.

2.1.2 Scaling

Across different information generation scenarios, scaling has been a significant factor to the performance improvement. It is largely attributed to the discovery of scaling laws [24]. Scaling laws describe how loss decreases in a log-linear manner as model size or training data volume increases. It can be formulated as follows:

$$L(x) = L_{\infty} + k \cdot x^{-\alpha}, \tag{2.1}$$

where L is the loss, x is model size or data size, and k and α are coefficients. This scaling formula has become a crucial theoretical guide in the era of large models, suggesting that performance can be enhanced at a log-linear rate simply by scaling up the model size or training data. Based on these scaling laws, researchers also derived optimal model sizes given fixed computational resources [25]. Their findings indicate that as computational capacity expands, it is beneficial not only to increase the training step but also the model size. This insight has further facilitated the pursuit of large models. The correctness of scaling laws was first proposed in language modeling field and then validated in many other areas, including data mixture scaling laws [26], multimodal scaling laws [27], and scaling laws specific to Information Retrieval (IR) [28].

Despite wide recognition of scaling laws, there remains disagreement among researchers about whether scaling is the correct path to the future. This stems from two main concerns: the uncertain relationship between loss and practical metrics and the inference costs associated with large models.

- **Loss vs. metric improvement:** The first arguing point is whether a linear reduction in loss can translate into super-linear improvements in actual metrics. If metrics could improve super-linearly with linear increases in computational effort, scaling up models would be highly advantageous. However, if the decrease in loss only results in linear or sublinear metric improvements, the diminishing improvements make scaling an inefficient option. The relationship between loss and metric performance remains an open question. Some researchers [29] believe that metrics can improve super-linearly, which is termed emergent abilities. This is further supported by Du et al. [30], who observed a jump in metrics when loss reaches a certain threshold. Additionally, [31] introduced the concept of "grokking" to explain emergence, showing that models might suddenly exhibit strong generalization capabilities when provided with sufficient computational resources. Nevertheless, some researchers [25] argued that such phenomena do not exist, showing that a well-trained smaller model can outperform a larger, undertrained one. Schaeffer et al. [32] demonstrated that emergent abilities are artifacts of discrete metric functions and found that continuous metric functions do not exhibit such behaviors. McKenzie et al. [33] even found that scaling results in worse metric scores. The existence of specific emergent abilities remains unresolved and needs to be investigated in future work.
- **Inference cost considerations:** Early studies on scaling laws did not account for the higher inference costs associated with larger models. Thus, the arguments that larger models are better [25] do not apply when the inference costs are considered. Instead, small models demonstrate potential to lower the inference costs. As shown by Fang et al. [28], the optimal model sizes become significantly smaller when accounting for inference costs. Besides, [34] show that smaller models can utilize more sampling steps during inference and thus perform better. Consequently, many recent studies focus on extensively training small models. For example, Llama [3] and MiniCPM [35] are trained with data

and steps that far exceed the guidance suggested by scaling laws. In the future, the models may be used on a phone to build up intelligent interaction with users. Thus, it is important to develop high-performing small models.

2.1.3 Training

Generative models in different scenarios are similar in training. For example, they usually use autoregressive training objectives, pretraining to Supervised Fine-Tuning (SFT) to Reinforcement Learning from Human Feedback (RLHF) training stages, and prompt tuning procedure. In this section, we focus on the text generation scenario. We first discuss the training objectives and then show the three training stages. Finally, we discuss how to design the prompts after the model is trained.

2.1.3.1 Training Objectives

For generative language models, the training objective is usually next token prediction. However, this was not widely used when Transformers first appeared. Initially, masked language modeling was the prevalent training objective during the bidirectional encoder representations from transformers (BERT) era [8]. It masks 15% of the words in a text randomly, and the model is tasked with predicting these masked words. This approach allows the model to utilize bidirectional attention, enhancing its representational capabilities. Even today, BERT models perform better than autoregressive models on tasks requiring bidirectional attention. However, a significant drawback of this method is the gap between its training setup and downstream tasks, necessitating a fine-tuning phase for adaptation to various applications. Thus, its zero-shot generalization capabilities are very limited.

Next token prediction was developed to address the inability of masked language modeling to generalize zero-shot to downstream tasks. The authors of GPT-2 [36] proposed that all Natural Language Processing (NLP) tasks could be reformulated as next token prediction tasks. By training models on this task, models could be directly applied to any downstream task without the need for specific fine-tuning. In fact, research nowadays demonstrates the effectiveness of this idea. Mathematically, next token prediction can be represented with the following formula:

$$P\left(x_{t+1} \mid x_1, \ldots, x_t\right), \tag{2.2}$$

which is to predict the probability of the next token x_{t+1} given the sequence of previous tokens.

2.1.3.2 Training Stages

The training process of language models typically unfolds in three stages: pre-training, SFT, and RLHF. Each phase presents unique challenges and methodologies.

Pre-training is the most resource-intensive stage. It is training a randomly initialized model on a large dataset to develop a robust linguistic capability. Several challenges arise during this stage: (1) Large models are especially difficult to train from random initialization. During training, there are often spikes in training loss or difficulty in converging [6, 23, 37]. We discussed various architectural improvements in Sect. 2.1.1 to address these instabilities, yet a definitive solution remains an open issue. (2) The computational demand is substantial. Pre-training requires stable and efficient use of computational resources [1]. It often involves parallel processing across multiple machines, which can lead to low utilization rates of computing resources [38]. Zeng et al. [6] reported numerous hardware failures during pre-training. (3) The quality of pre-training data is crucial [39]. Given the vast amount of data needed, efficiently filtering out low-quality data is essential. The filtering methods usually employ neural scoring models and based on the credibility of the site [40, 41].

SFT is to train the model on instruction-response pairs [42]. The model can thus learn to follow instructions or engage in dialogue [3]. To enhance dataset diversity, researchers often leverage different types of NLP tasks. The quality of the dataset is significant and requires a skilled annotation team. Besides, it is also important to label safety-related data, which helps instruct the models to learn to reject inappropriate requests [3].

RLHF focuses on aligning the model with human preferences based on human feedback [43, 44]. The process starts by sampling real human prompts to which the model generates multiple responses. These responses are then compared by users or third-party annotators. A reward model is trained based on these human preferences. Subsequently, reinforcement learning techniques utilize the reward model to guide the model updates. This approach significantly enhances the quality of model outputs, especially in creative writing tasks. However, a major challenge is the generalizability of the reward model; as the model evolves, the reward model may no longer accurately assess the quality of outputs. Continuous iterations of this process are necessary to mitigate this issue [3]. Recently, there are also some offline reinforcement learning algorithms that do not necessitate training a reward model, such as direct preference optimization (DPO) [45]. Yet studies [46] show that such offline learning methods still underperform the online learning methods.

2.1.3.3 Prompt Optimization

Generative models are highly sensitive to input prompts; an effective prompt can significantly enhance the quality of the model's output [47]. Therefore, optimizing

prompts for a generative model is a crucial area of research. Here are three main directions:

- **Designing prompt templates:** Researchers often design prompts that mimic human thought processes to guide the model effectively. This includes using structured thought patterns like chain of thought [48], tree of thought [49], and self-consistency [50], which help the model organize and process information in a logical manner.
- **Iterative optimization of prompt templates:** As with reinforcement learning, this method continuously iterates and refines the prompt templates based on the generation feedback. Given that prompt templates are typically discrete, researchers usually employ large language models to conduct prompt updates [51, 52].
- **Training prompt rewriting models using user interaction logs:** This approach harnesses the rich feedback contained within user interaction logs to tap into user insights. By analyzing how users interact with the model, researchers can train an automated model to rewrite prompts more effectively. This method leverages real-world data to better align the prompts with user intentions and improve the model's responses [53, 54].

2.1.4 Multi-Modal Applications

The rapid advancement of language models has significantly helped progress in the multimodal domain. Language models facilitate the understanding of multimodal data and developments in multimodal generation. We will discuss these two aspects separately.

2.1.4.1 Multi-Modal Understanding

Multimodal understanding involves models processing inputs from multiple modalities to produce relevant textual responses. For example, GPT-4o can process textual, visual, and auditory input. Challenges in this area include designing model structures that can handle multimodal inputs and crafting appropriate training objectives. Here, we focus on how visual signals are integrated into large language models:

In terms of aligning multimodal inputs, there are mainly three approaches:

- **Object detection-based input:** This method involves detecting objects within an image, extracting their features and associated spatial information, and then feeding this data into the language model [55, 56]. While this approach is effective, it tends to be slow due to the processing time required for object detection.
- **Visual encoding:** Another method encodes images directly using a visual encoder, which converts images into a latent vector representation before

integration with the model [57–61]. This method can sometimes result in the loss of detail.

- **Patch-based input:** The most efficient approach involves dividing images into several patches, transforming them with a simple linear layer, and directly inputting them into the model without the need for a complex visual encoder [62].

In terms of training methods, there are mainly four types of training objectives:

- **Contrastive learning or image-text matching:** These tasks require the model to correctly categorize images and their corresponding textual descriptions, aligning the representations of text and images [61, 63, 64].
- **Image captioning:** The model generates captions based on images, which helps it learn to understand the visual content [58–61].
- **Fine-grained image understanding:** The model is tasked to describe specific areas of an image or locate particular objects within an image. This helps enhance the model's detailed comprehension of visual elements [58, 65].
- **Image generation:** This task is reconstructing the original pixels of an image that has been blurred or corrupted [58, 66].

These methodologies and training objectives are crucial for advancing models' capabilities to process and interpret complex multimodal information effectively. This facilitates a more natural interaction with users.

2.1.4.2 Multi-Modal Generation

Multi-modal generation models, such as text-to-image generation, have substantially revolutionized the field of art creation. Traditionally, Generative Adversarial Networks (GAN) [67] and autoregressive methods [68] are mainstream methods. However, they are computationally expensive and cannot produce high-quality results. Recently, diffusion [69, 70] emerges as a new state-of-the-art method in multimodal generation. It perturbs the data with noise and learns to reconstruct the original data.

Language models are increasingly applied in the multimodal generation domain, such as in image [71, 72] and video generation [73, 74]. Language models are primarily utilized for processing training data and reformulating prompts.

In terms of training data, the titles associated with real-world images or videos often contain significant noise. If generative models are trained directly on these noisy titles, it could lead to inaccurate semantic understanding. To address this, language models can be used to filter and regenerate text descriptions within the training data [75, 76]. For instance, a multimodal understanding model could first be trained and then used to relabel videos or images to obtain more precise and detailed text descriptions. Experimental results have shown that this method significantly improves the fidelity of model generations to prompts.

During inference, multimodal generation models are highly sensitive to the input prompts. Many users do not know how to craft effective prompts and thus get

unsatisfying responses [77]. As a result, it is common to train a language model to rewrite user-provided prompts to enhance the quality of the generated images [75]. One of the challenges here is the difficulty in annotating such rewriting training data, as even system developers may not always know the optimal prompts, let alone crowdsourced workers [78]. To overcome this, some researchers collect a large number of user-shared effective prompts as training data [79]. Others build prompt-rewriting models based on user log data, capturing preferences, and feedback for training [53].

2.2 Information Synthesis

Other than generating information directly, another important research and application direction is to use the power of generative AI models, particularly LLMs, to integrate existing information and generate grounded responses accordingly. For simplicity, we refer to this paradigm as *information synthesis*. The key difference between information generation and information synthesis is the source of information. Information generation relies on the internal knowledge and information gathered through the training of generative AI models to create the model outputs, while information synthesis requires external sources to provide information to the models, and the models serve more as a integrator than a creator. There are multiple reasons why information synthesis is considered more reliable than generation in several IA scenarios. Here we discuss two of the most significant ones, i.e., model hallucination and external knowledge.

Hallucinating, which refers to the behavior of generative AI models that create responses and outputs that are not grounded by facts or existing supporting materials, is rooted in the foundation of most existing generative AI systems. For instance, LLMs create responses based on the next token prediction task, which formulates the generation of language as a probabilistic process and generates the next token in the output based on a probabilistic distribution (over the vocabulary) predicted by neural networks [1, 3]. The probabilistic model of LLMs allows them to capture knowledge in large-scale data efficiently and effectively, but it also introduces inevitable variance in their generation process. In other words, it is well acknowledged that it is theoretically impossible to prevent LLMs from generating data that are not seen in their training process [80]. While the ability of hallucinating is the source of creativity for LLMs (and for humans as well), it is not always desirable in practice, particularly for tasks with high requirements on result precision, reliability, and explainability. Therefore, asking the generative AI models to integrate human-created or factually grounded materials instead of generating information on their own is often considered more effective and robust to hallucination-sensitive applications.

The need for external knowledge is another key reason why we may prefer information synthesis over information generation. Despite the fact that modern generative AI models are trained with an incredibly large amount of data gathered

from the Web, there are many cases where we still need to retrieve and find support from external knowledge collections to finish certain tasks. Examples include the use of private datasets, vertical domain applications that require special knowledge, tasks that involve time-sensitive data, etc. It is usually inefficient or prohibitive to update large-scale generative AI models such as LLMs with task-oriented external data through model pre-training or SFT [81–83]. Even if possible, such paradigm is not preferred because the internal knowledge structures of most generative AI models are still a mystery (at least of today), and there is no guarantee that the models could behave and use the external information as we expect. In contrast, using generative AI models as information synthesizer gives us not only more flexibility but also more transparency and control over system outputs.

In this section, we discuss how generative AI models, particularly LLMs, can serve as effective information synthesizers for IA. We start with introducing one of the most popular information synthesis paradigm, i.e., Retrieval-Augmented Generation (RAG), and then discuss several other directions that utilize LLMs for corpus modeling and understanding.

2.2.1 Retrieval-Augmented Generation

RAG refers to the process of augmenting LLMs with data retrieved from external collections or synthesizing multiple retrieval results with LLMs for downstream applications [84, 85]. While the popularity of RAG rose after the release of large-scale pre-trained language models such as GPT [1] and BART [86], relevant topics and techniques have already been studied for at least more than two decades in both the IR and NLP communities, e.g., extractive and abstractive summarization that generates summary based on retrieved sentences [87, 88] or answer extraction from top retrieved document [89]. A major reason why RAG-like techniques were not as attractive as they are today is the limited performance of generative models before the era of LLMs. After ChatGPT [1] demonstrated superior ability text generation at the end of 2022, there have been many studies and surveys on RAG and its applications in LLMs [84, 90, 91]. As the intent of this chapter is not to provide yet another survey on existing RAG papers, we focus the following discussions on several present and future directions for RAG and their relations underneath.

2.2.1.1 Naive RAG

Naive RAG refers to the paradigm that directly feeds documents or other types of information retrieved by a retrieval system to the input (e.g., prompts) of a generative AI model and hope that the model can generate better output with or without a specific target task [92]. It is also referred to as the "Retrieve-then-Read" framework that has been used in reading comprehension and text summarization before LLMs hit the world [93]. Given an input (could be a query or a specific

task instruction), we first retrieve relevant information (usually entities, passages, or documents) from an external corpus or previous inputs (e.g., the memory of an agent [94, 95]) with a retrieval system. Then, we craft a input prompt with the retrieval results and feed it to the LLM. The LLM will generate the final response based on the input request and the retrieved information. This paradigm has already been proven to be effective in multiple IA tasks such as question answering [85].

Since LLMs are purely used as black-box tools to process the retrieved documents and input request in naive RAG, existing studies in this direction mainly focus on the development of better retrieval systems and prompt design for RAG. The studies on retrieval systems, unsurprisingly, are highly similar to those in IR, which involve indexing, query processing, first-stage retrieval, re-ranking, etc. These topics and system components have already been studied in the IR community for more than five decades. Perhaps the most notable difference is that recent studies on naive RAG often prefer the use of neural retrieval models (e.g., dense retrieval models [96]) over traditional term-matching models (e.g., BM25 [97]). An important reason behind this is that neural retrieval models share similar theoretical background and model structures with LLMs. This makes joint optimization possible in modern RAG systems, which we discuss in Sect. 2.2.1.3.

The design of input prompts with retrieval results, on the other hand, is relatively more under-explored before the rise of LLMs. It has been well recognized that prompt formats, even when the contents are same, could significantly affect the performance of LLMs. How to feed retrieval results effectively into the prompts of LLMs for RAG has thus attracted a lot of attention recently [93, 98, 99]. Studies have found that LLMs exhibit significant position bias over the input result sequences [100, 101] and has different perspectives on relevance with human experts [102]. Since prompts are the main interaction interface between retrieval and generation, their design principles and downstream effects on naive RAG are of great value both in research and real-world applications. Particularly, how to craft effective RAG prompts automatically could be a fruitful direction to explore. Existing studies have shown that high-quality prompt writers can be automatically learned based on downstream task performance and user logs in image generation [53], and it is widely believed that similar techniques have also been used in popular LLM chatbots [103]. Yet how to do this for RAG remains to be a question to be answered.

2.2.1.2 Modular RAG

In contrast to naive RAG methods, modular RAG treats retrieval systems as functional modules to support LLMs [104]. While some works view this retrieval module as one type of many tools that can be learned and used by LLMs [105], it is widely acknowledged that retrieval systems possess an irreplaceable position in modern LLM applications due to its diverse nature and significant importance [84]. Broadly speaking, existing studies on using retrieval systems as functional modules

for LLM generation mainly focus on the three "W" questions, namely, *when to retrieve*, *what to retrieve*, and *where to retrieve*.

The question of *when to retrieve* refers to the timing of functional call for retrieval systems. In contrast to LLMs that directly create responses based on their internal parameter space without explicit evidence grounding, retrieval systems produce reliable and explainable information directly by searching external corpus. From this perspective, the best timing to call the retrieval system is when LLMs start to hallucinate or produce wrong results. Yet identifying such timing is difficult because we neither know the correct answers in advance or understand the internal mechanisms of LLMs (at least as of today) [106]. One naive yet effective method is to retrieve supporting evidence for LLM inference with a fixed time interval, such as every fixed number of generated tokens [107, 108] or every sentence [109]. More advanced paradigms involve the analysis of knowledge boundary [110] and the estimation of prediction uncertainty in LLMs [106, 111]. Theoretically speaking, since the study of *when to retrieve* shares similar motivations and foundations with the study of hallucination detection, existing studies on LLM hallucination [112, 113] could provide important inspiration for research on this topic. Promising directions include better fact-checking systems for LLMs [114] and more investigations on how to characterize the confidence and uncertainty of LLM predictions based on both external behavior and internal state analysis [111].

The question of *what to retrieve* focuses on analyzing the intents and information needs of LLMs in inference. LLMs often need the help of different tools and systems to finish different tasks [105]. However, in contrast to other tools widely studied in tool learning, retrieval itself is a complicated systems with dynamic and free-form inputs, data collections, and outputs. Therefore, understanding what exactly is needed by LLMs and how to formulate it in the language of retrieval systems is an important problem. Most existing studies on RAG naively use the whole or local context of LLM inference as the queries to retrieval systems and assume that these context contain enough information to guide retrieval [90]. A slightly better solution is to use the terms that LLMs have low confidence to formulate queries since uncertain tokens represent cases where LLMs have limited knowledge to generate responses and thus need more information [106]. As long studied in the IR community, the formulation of an effective query requires deep understanding of the user's intent, and many of the important context information behind a user intent is not explicitly expressed in the words they wrote [115]. Therefore, a more theoretically principled method to answer *what to retrieve* in RAG is to analyze the internal state of LLMs and infer their information needs directly. For example, Su et al. [111] directly formulate queries based on the internal attention distribution of LLMs (Fig. 2.2) and improve the performance of RAG for nearly 20% on several benchmark datasets without changing the retrieval system. This demonstrates the potential of future studies in this direction.

Where to retrieve refers to the question of how to identify the correct information sources for RAG. Studies in this direction are particularly related to the research on multi-source retrieval [116] and tool learning [105]. To answer different requests related to the use of information collected from different databases or data

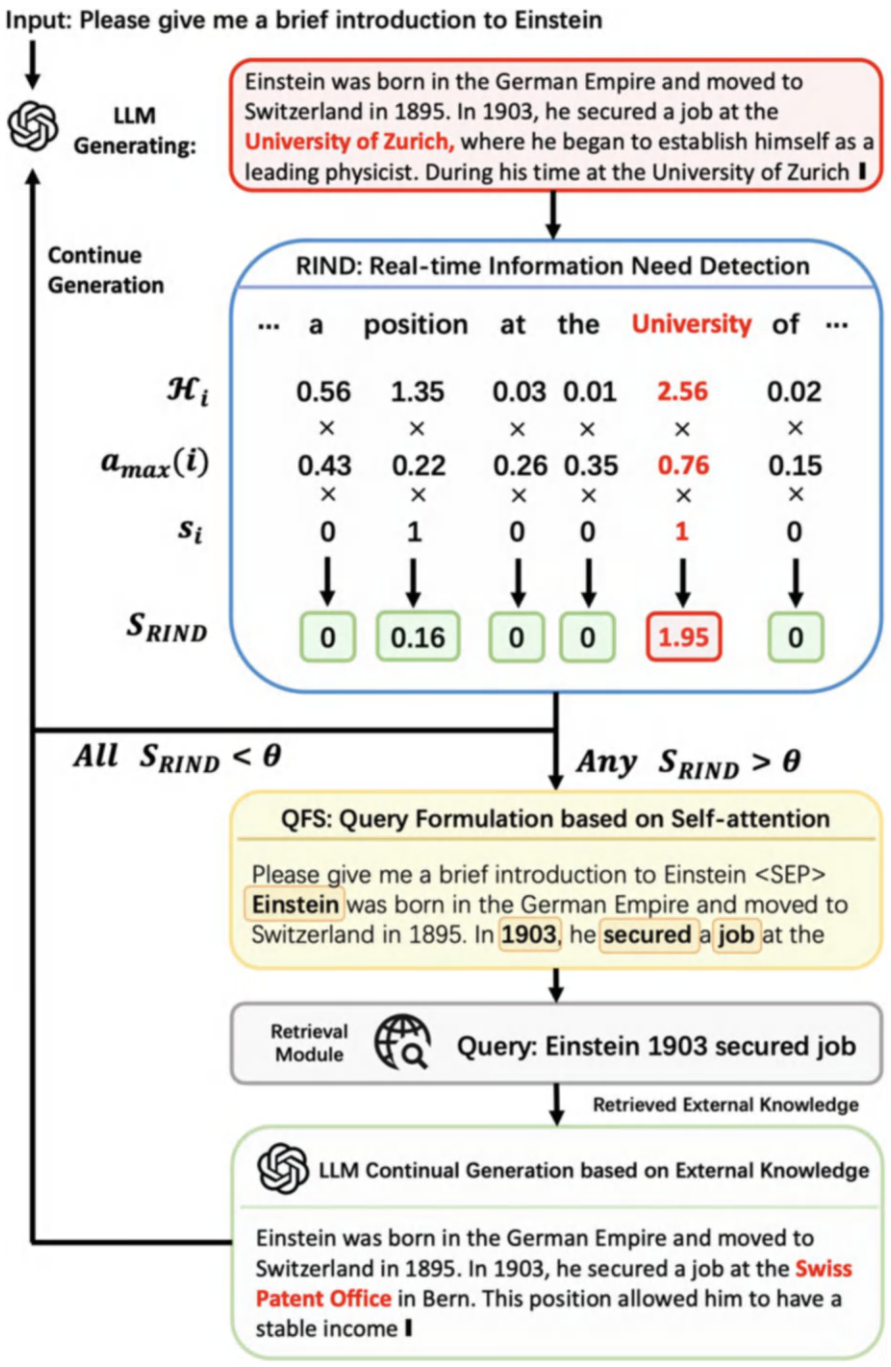

Fig. 2.2 Su et al. [111] generate queries for RAG based on the internal attention distribution of LLMs

collections, LLMs need to learn how to interact with each information sources effectively and efficiently. Studies of tool learning focus on teaching LLMs to use tools according to the context, and retrieval systems are usually considered as one type of tools to use. However, retrieval itself could be a complicated problem when we possess multiple data collections with different characteristics. In search engines, information sources are broadly categorized based on their modality, and we usually build separate systems for each of them (e.g., the "Images," "News," "Videos" tabs on Google). While commercial search engines may aggregate results from different sources into a single page, the ultimate Search Engine Result Page (SERP) shown to users are just a list of results, and it is up to the users to decide which they want to see and how to use these results for downstream applications. In contrast, when using LLMs, users often request LLMs to directly answer their question instead of listing a couple of candidates [117, 118], so it is the job of LLMs to decide where to retrieve the information given the current context. While studies of how to navigate user queries to search indexes built from different information sources have been widely studied in the IR community [119–122], how to do it for RAG with modern generative AI models is, to the best of our knowledge, still underexplored. Existing literature on RAG mostly works on a single retrieval collection (usually a text corpus), but it is obvious that no single collection can satisfy the needs of LLMs in different tasks. For instance, when writing a legal case document, the judge needs to collect and organize information from evidences, complaints, counterclaims, court records, as well as legal articles and previous cases. How to navigate the generation model to retrieve and integrate information from different sources jointly for downstream applications is a practical and potentially fruitful research question for RAG.

2.2.1.3 Optimization of Retrieval and Generation

As discussed in several RAG surveys [84, 90], the optimization of RAG systems usually involves the optimization of three components, i.e., the retriever, the generator, and the augmentation method. If we further step back and look at the high-level goals of RAG optimization, we could also categorize it based on how we evaluate the RAG system, namely, the evaluation from the perspectives of retrievers, generators, or the joint systems. The evaluation from the retriever perspectives is not particularly different from existing studies on ranking evaluation. The underlining assumption of this is that once the LLMs are fed with the passages or documents that contain the correct information, they should be able to produce the correct answers directly. Therefore, the evaluation and optimization of a RAG system could downgrade to the evaluation and optimization of a classic retrieval/ranking systems, to where most existing works on dense retrieval and Learning to Rank (LTR) could be applied [123, 124]. Yet there are still differences between RAG and traditional retrieval tasks as the queries are no longer issued by users. How to formulate queries efficiently and effectively from LLMs for the retriever is a worthy research question,

and studies in this direction have already shown potential in improving the overall quality of RAG systems [111].

From the perspective of generators, RAG evaluation and optimization focus more on improving the robustness and effectiveness of LLM generation based on a fixed set of retrieval results [108]. This often means extra training or fine-tuning on LLMs to improve their fundamental ability in information processing. For example, retrieved documents could be lengthy, and LLMs are usually not good at processing long input context [101]. Therefore, how to design efficient LLMs that can take long context inputs efficiently and effectively has been a popular research problem that have been widely studied by researchers from both academia and industry [100]. We have seen many companies show off their models based on how many input tokens they can process in one request. In addition, since retrieval results are fed as a part of the LLM inputs, whether the LLMs can generate the response based on the retrieved documents instead of their internal knowledge could be seen as a special type of instruction-following ability. Studies have been conducted to teach LLMs to utilize retrieval results faithfully and constantly in RAG systems [125] On the other hand, factors such as irrelevant results and ranking perturbations are well acknowledged to be harmful for the performance of generators in RAG, so there are also studies that try to improve the robustness of LLMs from the perspective of RAG. For example, [126] proposes to fine-tune LLMs with the presence of retrieval results (i.e., retrieval-augmented fine tuning) so that LLMs can learn the domain-specific knowledge introduced by the retriever and improve their robustness against potential distracting information from retrieval.

From the perspective of augmentation methods, existing research mostly focuses on the joint optimization of the RAG system as a whole. In other words, the loss functions of RAG optimization should be built from the performance metrics of downstream tasks directly. While this paradigm is appealing, it often has strict requirements on the design of RAG systems. Particularly, it is difficult to apply such joint optimization algorithms on a RAG system in which retrievers and generators are loosely connected through prompts constructed from discrete retrieval results. While reinforcement learning could solve the problem in theory, its empirical performance when being used as the solo optimization algorithms for ranking systems is still not satisfying at this point [127]. If you already have a good retriever and only conduct fine-tuning with a fixed LLM, then it may work [128], but this still does not look like a perfect solution because reinforcement learning is usually subject to large variance in practice. To the best of our knowledge, how to directly connect the training of retrievers with the auto-regressive loss of the generators in RAG is still an open question. Answering this question requires us to go deep into the structure of generative AI models and retrieval models and develop new model structures that can take advantages from studies on both sides.

2.2.1.4 Retrieval Planning and Composite Information Needs

As discussed above, the initial motivation behind the studies of RAG mostly focuses on using the power of retrieval systems to improve the quality of responses generated by LLMs in terms of reliability and informativeness. While it is widely acknowledged that problems such as hallucination and high computation cost in supervised fine-tuning will continue to be significant for generative AI models in a short period of time, there are also concerns, especially from the IR community, that retrieval could become less important with the rapid evolution of LLMs [129]. In fact, ChatGPT has already shown similar accuracy and better user satisfaction on factoid question answering than traditional Web search engines [1]. However, the rise of generative AI models also brings brand new opportunities for IR. One of them is the possibility of moving from SERPs that simply list result candidates to a real information agent that solve complicated tasks with composite information needs.

Today, most people treat IR systems as *unit information solvers*. Despite their actual task characteristics, users first decompose their goals into a couple of unit information needs (usually expressed with separate queries) and then issue them one by one to search engines or recommendation systems to find the corresponding answers. An important reason behind the popularity of this paradigm is that, at least of today, IR systems are not capable of doing complicated information tasks with composite needs and multi-step planning. For example, we can use a search engine to find a survey on RAG by searching "survey of RAG," but cannot write such a survey directly by retrieving and analyzing papers from publication collections. The job of information need decomposition and retrieval planning has always been human's.

Fortunately, with the help of generative AI models such as LLMs, it is now possible to push the boundary of IR systems and tackle such advanced information tasks for users. Composite retrieval is not a new concept in IR [130], but previous studies refer to the phrase as retrieval paradigms that cluster results from multiple sources and show them in groups for specific user queries [131]. While this represents one type of composite needs, it is relatively simple as the target user queries usually are mostly topic specific and keyword based. Complicated information tasks such as survey generation and professional document writing often involve multi-step planning and multi-round interactions between the retrieval results and response generation. To build powerful IR systems or agents that can solve such composite information tasks, we need to construct collaborative systems that deeply connect the retrieval, planning, and generation. For instance, we need to conduct generation-oriented retrieval optimization to build retrieval framework and model interfaces for downstream task planner and response generators; we also need to design retrieval-oriented generation models that can decompose information needs, navigate the retrieval process, and gather information from multiple sources to generate the final results. Research on these directions could be fruitful and significantly extend the scope of IR in the era of generative AI.

2.2.2 *Corpus Modeling and Understanding*

In contrast to using RAG, another line of studies try to use generative AI models to replace traditional retrieval systems. Directly answering a user's information need instead of showing ten blue links has long been an important goal for the development of intelligent IR systems [132]. With the rise of LLMs, such vision is now achievable in a significant extent. For example, LLM-based chatbots such as ChatGPT can answer multiple types of user queries with direct answers [118]. Metzler et al. [133] has discussed several paradigms in which pre-trained language models can help IR systems answer a user's information needs directly without listing references. The intuition is to use neural network-based language models to store the corpus knowledge in parameter space and pull relevant answers or information directly from it based on user's queries. Depending on how the problem is formulated, several research directions have emerged. Specifically, in this section, we discuss two of them, namely, Generative Retrieval (GR) and domain-specific modeling.

2.2.2.1 Generative Retrieval

The idea of *generative retrieval* comes from the idea of differentiable index proposed by Metzler et al. [133]. The original name used in the paper was *model-based IR*, but after the rise of generative AI models, some researchers start to refer to studies in this direction as generative retrieval (GR). The core idea of GR is twofold, i.e., the differentiable index and the generation of doc IDs.

Inspired by the superior performance of pre-trained language models, particularly BERT [8] and GPT [1], generative retrieval wants to explore the possibility of replacing traditional term-based index (e.g., inverted index) in retrieval systems with large-scale neural networks. In contrast to dense retrieval models that build neural encoders to project documents to latent semantic spaces and build explicit indexes based on document vectors, GR tries to build implicit indexes in the parameter space of neural networks. For instance, Differentiable Search Indexing (DSI) and its variations [134–137] have tried to train pretrained language models on the target corpus directly and then treat the model's parameter as an "index" of the corpus. Studies in this direction argue that by training the neural models to encode the whole corpus, documents and information would be implicitly stored in the parameters of the models, and these parameter-based indexes have better storage efficiency than traditional term-based or vector-based indexes [134]. They also argue that such paradigm can unify the multi-stage retrieval pipeline so that indexes can be trained directly for the final retrieval objectives. However, storing raw document content directly in limited parameter spaces often lead to significant information loss (which is reflected in the suboptimal retrieval performance of GR models [138]), and using model parameters as indexes makes the whole system uncontrollable by both system developers and users. While the former could be alleviated by using large-

scale models, the latter is still an unresolved problem for GR. For example, it is difficult, if not impossible, to remove or update a document indexed in the parameter space when we do not know what exactly each parameter do in the neural models. Considering that dense retrieval models built with product quantization and inverted file systems can achieve state-of-the-art retrieval performance with similar latency and less storage than term-based models with inverted indexes [139], whether the idea of differentiable indexes in GR is worth its price is still a controversial question.

Another important characteristic of GR models is to retrieve documents by generating sequences of doc IDs through autoregression. Since documents are stored implicitly in model parameters, to actually retrieve a real document, GR models use user's queries as prompts to generate document IDs, which usually consist of a couple of special tokens that exclusively identify each relevant document. Since the birth of GR, a variety of document IDs have been proposed, which can be broadly categorized as IDs with explicit tokens [134, 135, 137] and IDs with implicit tokens [136, 140, 141]. GR models with explicit ID tokens try to label each document with sequences of real terms that have semantic or numerical meanings. Examples include keyword-based doc IDs and tree-based doc IDs [134]. Compared to vectors in dense retrieval, these methods have less flexibility and capability in document modeling as they discretize document semantic meanings with a limited number of tokens, and their retrieval performance is usually poor [140]. However, they have better explainability than other neural retrieval models because their doc ID tokens are constructed from real words or document clusters. To avoid the theoretical limitation of explicit token IDs and grant GR models with the same modeling capacity of dense retrieval models, several studies have proposed to build implicit token IDs with latent vectors [136, 140, 141]. The idea is to represent each document with a sequence of latent vectors so that fine-grained semantic information would not be lost. These types of GR models are highly similar to existing dense retrieval models since both of them represent each document with latent vectors. The major difference is that the former uses a sequence of vectors from a learned codebook constructed in training, while the latter builds separate vectors for each document directly from their raw content. [142] have proved that GR models with implicit tokens are equal to a multi-vector dense retrieval models in theory. Also, the use of a learned codebook for implicit token vectors is theoretically the same with a dense retrieval system that uses cluster-based product quantization [139, 143]. Therefore, the performance upper bound of GR (with implicit tokens) and dense retrieval is the same in theory. While some believe that GR models could have lower latency as they don't need to search among millions of documents on the fly, this is a questionable argument because the inference of a large-scale neural model is usually much slower than a vector-based search on distributed systems. Also, the maintenance of data in a neural model is much more complex than it is in a vector-based database. Perhaps the future potential of GR does not lay in retrieval effectiveness or efficiency but some other perspectives such as explainability.

2.2.2.2 Domain-Specific Modeling

LLMs, particularly those with instruction tuning, can respond to user's queries directly. This exactly matches the initiative of a long-standing vision of IR systems to directly answer user's need without listing a couple of documents [133]. Therefore, ever since the rise of ChatGPT, there has been a serious discussion on whether LLMs are future search engines in practice [129]. Yet apart from the hallucination problem discussed in previous sections, there are other challenges that prevent generative AI models like LLMs to serve as a major information accessing tool for modern users. One of them is how to teach LLMs to understand and use knowledge from external corpus not included in their initial training process. If we treat each external corpus as a domain-specific dataset, then the studies in this direction are essentially the same with the construction of domain-specific LLMs. While RAG can help LLMs adapt to new domains quickly, their performance is limited when the understanding of input documents from the external corpus requires domain knowledge that the LLMs do not possess in advance [83].

To solve the above problem and build usable IA systems with LLMs on domain-specific data, one of the most popular method is to conduct continued pre-training or supervised fine-tuning of LLMs on the target domain corpus. The idea is to apply similar training strategies used in model pre-training on the new corpus so that LLMs can better capture knowledge in the new domain. Example studies in this direction include techniques on data selection [82] and tokenizers adaptation [144] that directly use the target corpus to train LLMs. Many domain-specific LLMs have been developed, including legal LLMs, financial LLMs, etc. [145–147] The continued pre-training of LLMs on external corpus has been shown to be effective on many domain-specific tasks such as domain QA and text generation. However, modeling external corpus through this method may not be preferred in practice when we do not have enough computation resources to train LLMs or cannot access the parameters of them. Also, till the end of the today, the internal knowledge structure and learning mechanism of LLMs are still unknown, and applying naive continued pre-training algorithms on external corpus could hurt the performance of LLMs in unexpected way. Therefore, researchers have designed several knowledge editing techniques on LLMs to explore the possibility of injecting knowledge with no or low cost on the general effectiveness of LLMs [148, 149]. Studies in this direction are still in an early stage as most existing methods only work on fixed and limited updating rules and knowledge entity triples [150], but it could be fruitful in the future since domain adaptation and external corpus modeling is a wide need of LLM applications in practice.

Besides continued pre-training, another paradigm to model external corpus and domain knowledge is to build separate language models for each corpus and combine them with the large general LLMs to form a collaborative system. The intuition behind this is relevant to the idea of LLM agents where each LLM could serve different roles in the system to accomplish tasks together. It is widely acknowledged that the emergence of abilities only present in large-scale models [29], but training models with such large scale (e.g., GPT-4 [1]) is

usually prohibitive, even with parameter efficient algorithms [151]. Inspired by the superior instruction following ability of LLMs, researchers have explored the possibility of building small models for external corpus modeling and use them to communicate domain-specific knowledge to large general LMs [83]. In other words, the small models can serve as domain knowledge "consultants," while large general models can serve as the decision-makers that finish domain-specific tasks based on the guidance of the small models. Experiments have shown that such a paradigm can improve black-box LLMs' performance on domain-specific tasks with low cost and high flexibility. While the overall idea of prompt general LLMs with domain-specific prompts is similar to the framework of RAG, building an actual LM for corpus modeling enables us to capture implicit domain knowledge (e.g., the fine-grained differences between law articles [152]) and potentially save tokens in prompts. There are concerns on whether this paradigm is still worthy when we have more powerful LLMs that include more domain-specific data in training. However, since many users prefer to keep their data private to themselves due to multiple safety and privacy concerns, this paradigm and RAG could continue to be appealing in practice.

2.3 Summary and Future Directions

In this chapter, we introduce the foundations and applications of generative AI models in information accessing. Instead of analyzing how generative AI models like LLMs could improve the existing modules of search engines and recommendation systems, we focus on how they could revolutionize information access with new methodologies and system design. Particularly, we discuss two new paradigms brought by generative AI models, namely, information generation and information synthesis.

Information generation refers to scenarios where users can use generative AI models to create information that directly satisfies their information needs. Here, we delved into the core components of generative models, including model architectures (with a focus on Transformers and their improvements), scaling laws, and training methodologies. We examined the debates surrounding continual model scaling, the importance of prompt optimization, and the extension of these models to multi-modal applications for information access.

Information synthesis refers to the paradigm that utilizes the superior instruction-following and logic-reasoning ability of LLMs to aggregate and synthesize existing information. We extensively discuss one of the most representative techniques, i.e., RAG, on this direction, and introduce various approaches from naive implementations to more sophisticated modular systems. We describe the challenges and opportunities in optimizing RAG systems, highlighting the need for joint retrieval-generation optimization and the potential of several relevant research directions such as composite retrieval with planning. Besides RAG, we also discuss some alternative paradigms that use generative AI models to model corpus knowledge directly,

such as generative retrieval, which aims to replace traditional indexing methods with neural network-based approaches, and domain-specific model training, which conducts continued pre-training or fine-tuning on LLMs with the target corpus. We discussed the potential and limitations of these approaches, including issues of system controllability and cost efficiency.

Overall, research on how generative AI models could reshape modern information access systems is still at an early stage today. As discussed above, existing studies on information generation and information synthesis either focus on simple information tasks (such as writing a poem, answering a factoid question, etc.) or reply on simple system design (e.g., feeding all documents to LLMs as prompts) that obviously cannot fully exploit the power of modern retrieval and generation models. Therefore, we believe that there are two major directions worth exploring in the next couple of years (at least). The first one is to move from simple and unit information retrieval tasks (e.g., factoid question answering) to more complicated information tasks that used to be "impossible" for modern IR systems. Examples include retrieval with composite needs (e.g., "help me plan a wedding in Amherst, MA") or tasks that require planning and multiple rounds of retrieval and generations (e.g., "write a survey on RAG"). These tasks are used to require human experts to decompose the needs and conduct retrieval, analysis, and result aggregations. With the help of generative AI, accomplishing them automatically with machines is now possible. The second direction is to explore better techniques to communicate, collaborate, or even unify retrieval and generation systems for information access. While studies of RAG have attracted considerable attention, existing works mostly use retrieval systems as plug-in tools for LLMs without digging into their internal connections and differences. Examples such as how to understand the information needs of LLMs, how to communicate the retrieved results to LLMs, and how to optimize generators for retrieval and retriever for generation are all important yet underexplored research topics. There are many questions related to each of these topics that are worthy of detailed investigation, including the design of new training paradigms, the development of agent-like system frameworks, potential problems and bias introduced by off-policy and on-policy training for the joint system, etc.

When ChatGPT first arrived, people from the IR community were worried that such generative AI models could overthrow all existing IR systems and crush everything in the field [129], as it has almost happened in NLP. Interestingly, in simulated social experiments on human-AI competitions, [153] found that if human producers do not extend their capacities with the help of generative AI, they will eventually be "replaced" by AI. From this perspective, the future of IR research in the era of generative AI lies in how to extend the scope of IR with generative AI models to finish more complicated information tasks and develop more general system architectures that do not just retrieve a list of documents but also perform more sophisticated information processing and planning.

References

1. OpenAI: GPT-4 technical report. CoRR abs/2303.08774 (2023). https://doi.org/10.48550/ARXIV.2303.087742303.08774
2. Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J., Wen, J.: A survey of large language models. CoRR abs/2303.18223 (2023). https://doi.org/10.48550/ARXIV.2303.182232303.18223
3. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
4. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.U., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc., New York (2017)
5. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Trans. Signal Process. **45**(11), 2673–2681 (1997)
6. Zeng, A., Liu, X., Du, Z., Wang, Z., Lai, H., Ding, M., Yang, Z., Xu, Y., Zheng, W., Xia, X., et al.: Glm-130b: An open bilingual pre-trained model. arXiv preprint arXiv:2210.02414 (2022)
7. Le Scao, T., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A.S., Yvon, F., Gallé, M., et al.: Bloom: A 176b-parameter open-access multilingual language model (2023)
8. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT (1), pp. 4171–4186. Association for Computational Linguistics, New York (2019)
9. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**, 140–114067 (2020)
10. Press, O., Smith, N.A., Lewis, M.: Train short, test long: Attention with linear biases enables input length extrapolation. arXiv preprint arXiv:2108.12409 (2021)
11. Su, J., Lu, Y., Pan, S., Wen, B., RoFormer, Y.L.: Enhanced transformer with rotary position embedding (2021). https://doi.org/10.1016/j. neucom (2023)
12. Black, S., Biderman, S., Hallahan, E., Anthony, Q., Gao, L., Golding, L., He, H., Leahy, C., McDonell, K., Phang, J., et al.: Gpt-neox-20b: an open-source autoregressive language model. arXiv preprint arXiv:2204.06745 (2022)
13. Child, R., Gray, S., Radford, A., Sutskever, I.: Generating long sequences with sparse transformers. arXiv preprint arXiv:1904.10509 (2019)
14. Kitaev, N., Kaiser, L., Levskaya, A.: Reformer: the efficient transformer. arXiv preprint arXiv:2001.04451 (2020)
15. Munkhdalai, T., Faruqui, M., Gopal, S.: Leave no context behind: Efficient infinite context transformers with infini-attention. arXiv preprint arXiv:2404.07143 (2024)
16. Grave, E., Joulin, A., Usunier, N.: Improving neural language models with a continuous cache. arXiv preprint arXiv:1612.04426 (2016)
17. Izacard, G., Grave, E.: Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. arXiv (2020). https://arxiv.org/abs/2007.0128
18. Shazeer, N.: Fast transformer decoding: One write-head is all you need. arXiv preprint arXiv:1911.02150 (2019)
19. Ainslie, J., Lee-Thorp, J., Jong, M., Zemlyanskiy, Y., Lebrón, F., Sanghai, S.: GQA: training generalized multi-query transformer models from multi-head checkpoints. arXiv preprint arXiv:2305.13245 (2023)
20. DeepSeek-AI: DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model (2024)

21. Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Zhang, H., Lan, Y., Wang, L., Liu, T.: On layer normalization in the transformer architecture. In: International Conference on Machine Learning, pp. 10524–10533. PMLR, New York (2020)
22. Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., Lin, J., Zou, X., Shao, Z., Yang, H., et al.: CogView: Mastering text-to-image generation via transformers. Adv. Neural Inf. Proces. Syst. **34**, 19822–19835 (2021)
23. Wang, H., Ma, S., Dong, L., Huang, S., Zhang, D., Wei, F.: Deepnet: Scaling transformers to 1,000 layers. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
24. Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling laws for neural language models. arXiv preprint arXiv:2001.08361 (2020)
25. Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D.d.L., Hendricks, L.A.,Welbl, J., Clark, A., et al.: Training compute-optimal large language models. arXiv preprint arXiv:2203.15556 (2022)
26. Ye, J., Liu, P., Sun, T., Zhou, Y., Zhan, J., Qiu, X.: Data mixing laws: optimizing data mixtures by predicting language modeling performance. arXiv preprint arXiv:2403.16952 (2024)
27. Henighan, T., Kaplan, J., Katz, M., Chen, M., Hesse, C., Jackson, J., Jun, H., Brown, T.B., Dhariwal, P., Gray, S., et al.: Scaling laws for autoregressive generative modeling. arXiv preprint arXiv:2010.14701 (2020)
28. Fang, Y., Zhan, J., Ai, Q., Mao, J., Su, W., Chen, J., Liu, Y.: Scaling laws for dense retrieval. arXiv preprint arXiv:2403.18684 (2024)
29. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al.: Emergent abilities of large language models. arXiv preprint arXiv:2206.07682 (2022)
30. Du, Z., Zeng, A., Dong, Y., Tang, J.: Understanding emergent abilities of language models from the loss perspective. arXiv preprint arXiv:2403.15796 (2024)
31. Power, A., Burda, Y., Edwards, H., Babuschkin, I., Misra, V.: Grokking: generalization beyond overfitting on small algorithmic datasets. arXiv preprint arXiv:2201.02177 (2022)
32. Schaeffer, R., Miranda, B., Koyejo, S.: Are emergent abilities of large language models a mirage? Adv. Neural Inf. Proces. Syst. **36**, 1–13 (2024)
33. McKenzie, I.R., Lyzhov, A., Pieler, M., Parrish, A., Mueller, A., Prabhu, A., McLean, E., Kirtland, A., Ross, A., Liu, A., et al.: Inverse scaling: when bigger isn't better. arXiv preprint arXiv:2306.09479 (2023)
34. Mei, K., Tu, Z., Delbracio, M., Talebi, H., Patel, V.M., Milanfar, P.: Bigger is not always better: Scaling properties of latent diffusion models. arXiv preprint arXiv:2404.01367 (2024)
35. Hu, S., Tu, Y., Han, X., He, C., Cui, G., Long, X., Zheng, Z., Fang, Y., Huang, Y., Zhao, W., et al.: MiniCPM: unveiling the potential of small language models with scalable training strategies. arXiv preprint arXiv:2404.06395 (2024)
36. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
37. Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., et al.: OPT: open pre-trained transformer language models. arXiv preprint arXiv:2205.01068 (2022)
38. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al.: Palm: scaling language modeling with pathways. J. Mach. Learn. Res. **24**(240), 1–113 (2023)
39. Gunasekar, S., Zhang, Y., Aneja, J., Mendes, C.C.T., Del Giorno, A., Gopi, S., Javaheripi, M., Kauffmann, P., Rosa, G., Saarikivi, O., et al.: Textbooks are all you need. arXiv preprint arXiv:2306.11644 (2023)
40. Yang, A., Xiao, B., Wang, B., Zhang, B., Bian, C., Yin, C., Lv, C., Pan, D., Wang, D., Yan, D., et al.: Baichuan 2: open large-scale language models. arXiv preprint arXiv:2309.10305 (2023)

41. Bi, X., Chen, D., Chen, G., Chen, S., Dai, D., Deng, C., Ding, H., Dong, K., Du, Q., Fu, Z., et al.: Deepseek LLM: scaling open-source language models with longtermism. arXiv preprint arXiv:2401.02954 (2024)
42. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S.S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Valter, D., Narang, S., Mishra, G., Yu, A.W., Zhao, V., Huang, Y., Dai, A.M., Yu, H., Petrov, S., Chi, E.H.-h., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q.V., Wei, J.: Scaling instruction-finetuned language models. ArXiv abs/2210.11416 (2022)
43. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Adv. Neural Inf. Proces. Syst. **35**, 27730–27744 (2022)
44. Schulman, J.,Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal Policy Optimization Algorithms (2017)
45. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: your language model is secretly a reward model. In: Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems, vol. 36, pp. 53728–53741. Curran Associates, Inc., New York (2023)
46. Xu, S., Fu, W., Gao, J., Ye, W., Liu, W., Mei, Z., Wang, G., Yu, C., Wu, Y.: Is DPO superior to PPO for LLM alignment? a comprehensive study. arXiv preprint arXiv:2404.10719 (2024)
47. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G.: Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Comput. Surv. **55**(9), 1–35 (2023)
48. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Adv. Neural Inf. Proces. Syst. **35**, 24824–24837 (2022)
49. Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., Narasimhan, K.: Tree of thoughts: Deliberate problem solving with large language models. Adv. Neural Inf. Proces. Syst. **36**, 1–14 (2024)
50. Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., Zhou, D.: Self-consistency improves chain of thought reasoning in language models. In: 11th International Conference on Learning Representations (ICLR 2023), pp. 1–15. arXiv preprint arXiv:2203.11171 (2023)
51. Zhou, Y., Muresanu, A.I., Han, Z., Paster, K., Pitis, S., Chan, H., Ba, J.: Large language models are human-level prompt engineers. arXiv preprint arXiv:2211.01910 (2022)
52. Yang, C., Wang, X., Lu, Y., Liu, H., Le, Q.V., Zhou, D., Chen, X.: Large language models as optimizers. arXiv preprint arXiv:2309.03409 (2023)
53. Zhan, J., Ai, Q., Liu, Y., Chen, J., Ma, S.: Capability-aware prompt reformulation learning for text-to-image generation. arXiv preprint arXiv:2403.19716 (2024)
54. Zhan, J., Ai, Q., Liu, Y., Pan, Y., Yao, T., Mao, J., Ma, S., Mei, T.: Prompt refinement with image pivot for text-to-image generation. In: ACL (2024)
55. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems, pp. 13–23. (2019)
56. Chen, Y.-C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: universal image-text representation learning. In: European Conference on Computer Vision, pp. 104–120. Springer, Berlin (2020)
57. Huang, Z., Zeng, Z., Liu, B., Fu, D., Fu, J.: Pixel-BERT: aligning image pixels with text by deep multi-modal transformers. arXiv preprint arXiv:2004.00849 (2020)
58. Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., Yang, H.: OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In: International Conference on Machine Learning, pp. 23318–23340. PMLR, New York (2022)

59. Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Adv. Neural Inf. Proces. Syst. **35**, 23716–23736 (2022)
60. Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., et al.: CogVLM: visual expert for pretrained language models. arXiv preprint arXiv:2311.03079 (2023)
61. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. In: International Conference on Machine Learning, pp. 19730–19742. PMLR, New York (2023)
62. Kim, W., Son, B., Kim, I.: Vilt: vision-and-language transformer without convolution or region supervision. In: International Conference on Machine Learning, pp. 5583–5594. PMLR, New York (2021)
63. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: vision and language representation learning with momentum distillation. Adv. Neural Inf. Proces. Syst. **34**, 9694–9705 (2021)
64. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR, New York (2021)
65. Yu, T., Yao, Y., Zhang, H., He, T., Han, Y., Cui, G., Hu, J., Liu, Z., Zheng, H.-T., Sun, M., et al.: RLHF-V: towards trustworthy MLLMs via behavior alignment from fine-grained correctional human feedback. arXiv preprint arXiv:2312.00849 (2023)
66. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254 (2021)
67. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. In: International Conference on Machine Learning, pp. 1060–1069. PMLR, New York (2016)
68. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International Conference on Machine Learning, pp. 8821–8831. PMLR, New York (2021)
69. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
70. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684–10695 (2022)
71. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Adv. Neural Inf. Proces. Syst. **33**, 6840–6851 (2020)
72. Zhang, C., Zhang, C., Zhang, M., Kweon, I.S.: Text-to-image diffusion model in generative ai: A survey. arXiv preprint arXiv:2303.07909 (2023)
73. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4195–4205 (2023)
74. Singh, A.: A survey of ai text-to-image and ai text-to-video generators. In: 2023 4th International Conference on Artificial Intelligence, Robotics and Control (AIRC), pp. 32–36. IEEE, New York (2023)
75. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al.: Improving image generation with better captions. Computer Science. **2**(3), 8 (2023). https://cdn.openai.com/papers/dall-e-3.pdf
76. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R., Ramesh, A.: Video generation models as world simulators (2024)
77. Oppenlaender, J.: A taxonomy of prompt modifiers for text-to-image generation. In: *Behaviour & Information Technology*, pp. 1–14

78. Liu, V., Chilton, L.B.: Design guidelines for prompt engineering text-to-image generative models. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, pp. 1–23 (2022)
79. Hao, Y., Chi, Z., Dong, L., Wei, F.: Optimizing prompts for text-to-image generation. Adv. Neural Inf. Proces. Syst. **36**, 1–17 (2024)
80. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A., Fung, P.: Survey of hallucination in natural language generation. ACM Comput. Surv. **55**(12), 1–38 (2023)
81. Arefeen, M.A., Debnath, B., Chakradhar, S.: Leancontext: Cost-efficient domain-specific question answering using LLMs. Nat. Lang. Process. J. **7**, 100065 (2024)
82. Aharoni, R., Goldberg, Y.: Unsupervised domain clusters in pretrained language models. arXiv preprint arXiv:2004.02105 (2020)
83. Li, H., Ai, Q., Chen, J., Dong, Q., Wu, Z., Liu, Y., Chen, C., Tian, Q.: Blade: Enhancing black-box large language models with small domain-specific models. arXiv preprint arXiv:2403.18365 (2024)
84. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, H.: Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997 (2023)
85. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks. Adv. Neural Inf. Proces. Syst. **33**, 9459–9474 (2020)
86. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: Bart: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7871–7880 (2020)
87. Moratanch, N., Chitrakala, S.: A survey on extractive text summarization. In: 2017 International Conference on Computer, Communication and Signal Processing (ICCCSP), pp. 1–6. IEEE, New York (2017)
88. Lin, H., Ng, V.: Abstractive summarization: a survey of the state of the art. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 9815–9822 (2019)
89. Bajaj, P., Campos, D., Craswell, N., Deng, L., Gao, J., Liu, X., Majumder, R., McNamara, A., Mitra, B., Nguyen, T., Rosenberg, M., Song, X., Stoica, A., Tiwary, S., Wang, T.: MS MARCO: A Human Generated MAchine Reading COmprehension Dataset (2018)
90. Zhao, P., Zhang, H., Yu, Q., Wang, Z., Geng, Y., Fu, F., Yang, L., Zhang, W., Cui, B.: Retrieval-augmented generation for ai-generated content: A survey. arXiv preprint arXiv:2402.19473 (2024)
91. Asai, A., Min, S., Zhong, Z., Chen, D.: Retrieval-based language models and applications. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts), pp. 41–46 (2023)
92. Guu, K., Lee, K., Tung, Z., Pasupat, P., Chang, M.: Retrieval augmented language model pre-training. In: International Conference on Machine Learning, pp. 3929–3938. PMLR, New York (2020)
93. Ma, X., Gong, Y., He, P., Zhao, H., Duan, N.: Query rewriting for retrieval augmented large language models. arXiv preprint arXiv:2305.14283 (2023)
94. Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., et al.: A survey on large language model based autonomous agents. Front. Comp. Sci. **18**(6), 186345 (2024)
95. Zhang, Z., Bo, X., Ma, C., Li, R., Chen, X., Dai, Q., Zhu, J., Dong, Z., Wen, J.-R.: A Survey on the Memory Mechanism of Large Language Model based Agents (2024)
96. Zhan, J., Mao, J., Liu, Y., Guo, J., Zhang, M., Ma, S.: Optimizing dense retrieval model training with hard negatives. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1503–1512 (2021)
97. Robertson, S., Zaragoza, H., et al.: The probabilistic relevance framework: Bm25 and beyond. Found. Trends Inf. Retr. **3**(4), 333–389 (2009)

98. Mao, S., Jiang, Y., Chen, B., Li, X., Wang, P., Wang, X., Xie, P., Huang, F., Chen, H., Zhang, N.: Rafe: ranking feedback improves query rewriting for rag. arXiv preprint arXiv:2405.14431 (2024)
99. Chan, C.-M., Xu, C., Yuan, R., Luo, H., Xue, W., Guo, Y., Fu, J.: RQ-RAG: learning to refine queries for retrieval augmented generation. arXiv preprint arXiv:2404.00610 (2024)
100. Li, T., Zhang, G., Do, Q.D., Yue, X., Chen, W.: Long-context LLMs struggle with long in-context learning. arXiv preprint arXiv:2404.02060 (2024)
101. Liu, N.F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., Liang, P.: Lost in the middle: how language models use long contexts. Trans. Assoc. Comput. Linguistics **12**, 157–173 (2024)
102. Faggioli, G., Dietz, L., Clarke, C.L., Demartini, G., Hagen, M., Hauff, C., Kando, N., Kanoulas, E., Potthast, M., Stein, B., et al.: Perspectives on large language models for relevance judgment. In: Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval, pp. 39–50 (2023)
103. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G.: Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Comput. Surv. **55**(9) (2023). https://doi.org/10.1145/3560815
104. Wang, X., Yang, Q., Qiu, Y., Liang, J., He, Q., Gu, Z., Xiao, Y., Wang, W.: KnowledGPT: Enhancing large language models with retrieval and storage access on knowledge bases. arXiv preprint arXiv:2308.11761 (2023)
105. Qin, Y., Hu, S., Lin, Y., Chen, W., Ding, N., Cui, G., Zeng, Z., Huang, Y., Xiao, C., Han, C., Fung, Y.R., Su, Y., Wang, H., Qian, C., Tian, R., Zhu, K., Liang, S., Shen, X., Xu, B., Zhang, Z., Ye, Y., Li, B., Tang, Z., Yi, J., Zhu, Y., Dai, Z., Yan, L., Cong, X., Lu, Y., Zhao, W., Huang, Y., Yan, J., Han, X., Sun, X., Li, D., Phang, J., Yang, C., Wu, T., Ji, H., Liu, Z., Sun, M.: Tool Learning with Foundation Models (2023)
106. Jiang, Z., Xu, F., Gao, L., Sun, Z., Liu, Q., Dwivedi-Yu, J., Yang, Y., Callan, J., Neubig, G.: Active retrieval augmented generation. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 7969–7992. Association for Computational Linguistics, Singapore (2023). https://doi.org/10.18653/v1/2023.emnlp-main.495. https://aclanthology.org/2023.emnlp-main.495
107. Ram, O., Levine, Y., Dalmedigos, I., Muhlgay, D., Shashua, A., Leyton-Brown, K., Shoham, Y.: In-context retrieval-augmented language models. Trans. Assoc. Comput. Linguistics **11**, 1316–1331 (2023)
108. Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., Van Den Driessche, G.B., Lespiau, J.-B., Damoc, B., Clark, A., et al.: Improving language models by retrieving from trillions of tokens. In: International Conference on Machine Learning, pp. 2206–2240. PMLR, New York (2022)
109. Trivedi, H., Balasubramanian, N., Khot, T., Sabharwal, A.: Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 10014–10037. Association for Computational Linguistics, Toronto, Canada (2023). https://doi.org/10.18653/v1/2023.acl-long.557. https://aclanthology.org/2023.acl-long.557
110. Ni, S., Bi, K., Guo, J., Cheng, X.: When do LLMs need retrieval augmentation? mitigating LLMs' overconfidence helps retrieval augmentation. arXiv preprint arXiv:2402.11457 (2024)
111. Su, W., Tang, Y., Ai, Q., Wu, Z., Liu, Y.: DRAGIN: Dynamic Retrieval Augmented Generation based on the Information Needs of Large Language Models (2024)
112. Su, W., Wang, C., Ai, Q., HU, Y., Wu, Z., Zhou, Y., Liu, Y.: Unsupervised Real-Time Hallucination Detection based on the Internal States of Large Language Models (2024)
113. Liu, T., Zhang, Y., Brockett, C., Mao, Y., Sui, Z., Chen, W., Dolan, B.: A token-level reference-free hallucination detection benchmark for free-form text generation. arXiv preprint arXiv:2104.08704 (2021)

114. Fadeeva, E., Rubashevskii, A., Shelmanov, A., Petrakov, S., Li, H., Mubarak, H., Tsymbalov, E., Kuzmin, G., Panchenko, A., Baldwin, T., et al.: Fact-checking the output of large language models via token-level uncertainty quantification. arXiv preprint arXiv:2403.04696 (2024)
115. Cronen-Townsend, S., Croft, W.B., et al.: Quantifying query ambiguity. In: Proceedings of HLT, vol. 2, pp. 94–98 (2002)
116. Arens, Y., Chee, C.Y., Hsu, C.-N., Knoblock, C.A.: Retrieving and integrating data from multiple information sources. Int. J. Cooperative Inf. Syst. **02**(02), 127–158 (1993). https://doi.org/10.1142/S0218215793000071
117. Wang, J., Mo, F., Ma, W., Sun, P., Zhang, M., Nie, J.-Y.: A User-Centric Benchmark for Evaluating Large Language Models (2024)
118. Wang, J., Ma, W., Sun, P., Zhang, M., Nie, J.-Y.: Understanding User Experience in Large Language Model Interactions (2024)
119. Beitzel, S.M., Jensen, E.C., Chowdhury, A., Grossman, D., Frieder, O., Goharian, N.: Fusion of effective retrieval strategies in the same information retrieval system. J. Am. Soc. Inf. Sci. Technol. **55**(10), 859–868 (2004)
120. Wu, S., McClean, S.: Performance prediction of data fusion for information retrieval. Inf. Process. Manag. **42**(4), 899–915 (2006)
121. Cormack, G.V., Clarke, C.L., Buettcher, S.: Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 758–759 (2009)
122. Lee, C.-J., Ai, Q., Croft, W.B., Sheldon, D.: An optimization framework for merging multiple result lists. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, pp. 303–312 (2015)
123. Liu, T.-Y., et al.: Learning to rank for information retrieval. Found. Trends Inf. Retr. **3**(3), 225–331 (2009)
124. Zhan, J., Mao, J., Liu, Y., Zhang, M., Ma, S.: Learning to retrieve: How to train a dense retrieval model effectively and efficiently. arXiv preprint arXiv:2010.10469 (2020)
125. Arora, D., Kini, A., Chowdhury, S.R., Natarajan, N., Sinha, G., Sharma, A.: Gar-meets-rag paradigm for zero-shot information retrieval. arXiv preprint arXiv:2310.20158 (2023)
126. Zhang, T., Patil, S.G., Jain, N., Shen, S., Zaharia, M., Stoica, I., Gonzalez, J.E.: RAFT: Adapting Language Model to Domain Specific RAG (2024)
127. Xu, Z., Tran, A., Yang, T., Ai, Q.: Reinforcement learning to rank with coarse-grained labels. arXiv preprint arXiv:2208.07563 (2022)
128. Shi, W., Min, S., Yasunaga, M., Seo, M., James, R., Lewis, M., Zettlemoyer, L., Yih, W.-t.: REPLUG: retrieval-augmented black-box language models. arXiv preprint arXiv:2301.12652 (2023)
129. Ai, Q., Bai, T., Cao, Z., Chang, Y., Chen, J., Chen, Z., Cheng, Z., Dong, S., Dou, Z., Feng, F., et al.: Information retrieval meets large language models: a strategic report from Chinese IR community. AI Open **4**, 80–90 (2023)
130. Bota, H., Zhou, K., Jose, J.M., Lalmas, M.: Composite retrieval of heterogeneous web search. In: Proceedings of the 23rd International Conference on World Wide Web, pp. 119–130 (2014)
131. Amer-Yahia, S., Bonchi, F., Castillo, C., Feuerstein, E., Mendez-Diaz, I., Zabala, P.: Composite retrieval of diverse and complementary bundles. IEEE Trans. Knowl. Data Eng. **26**(11), 2662–2675 (2014)
132. Kolomiyets, O., Moens, M.-F.: A survey on question answering technology from an information retrieval perspective. Inf. Sci. **181**(24), 5412–5434 (2011)
133. Metzler, D., Tay, Y., Bahri, D., Najork, M.: Rethinking search: making domain experts out of dilettantes. SIGIR Forum **55**(1) (2021) https://doi.org/10.1145/3476415.3476428
134. Tay, Y., Tran, V., Dehghani, M., Ni, J., Bahri, D., Mehta, H., Qin, Z., Hui, K., Zhao, Z., Gupta, J., et al.: Transformer memory as a differentiable search index. Adv. Neural Inf. Proces. Syst. **35**, 21831–21843 (2022)
135. Tang, Y., Zhang, R., Guo, J., Chen, J., Zhu, Z., Wang, S., Yin, D., Cheng, X.: Semantic-enhanced differentiable search index inspired by learning strategies. In: Proceedings of the

29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 4904–4913 (2023)
136. Sun, W., Yan, L., Chen, Z., Wang, S., Zhu, H., Ren, P., Chen, Z., Yin, D., Rijke, M., Ren, Z.: Learning to tokenize for generative retrieval. Adv. Neural Inf. Proces. Syst. **36**, 1–17 (2024)
137. Zhuang, S., Ren, H., Shou, L., Pei, J., Gong, M., Zuccon, G., Jiang, D.: Bridging the gap between indexing and retrieval for differentiable search index with query generation. arXiv preprint arXiv:2206.10128 (2023)
138. Nguyen, T., Yates, A.: Generative retrieval as dense retrieval. arXiv preprint arXiv:2306.11397 (2023)
139. Zhan, J., Mao, J., Liu, Y., Guo, J., Zhang, M., Ma, S.: Learning discrete representations via constrained clustering for effective and efficient dense retrieval. In: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining. WSDM '22, pp. 1328–1336. Association for Computing Machinery, New York (2022). https://doi.org/10.1145/3488560.3498443
140. Zeng, H., Luo, C., Jin, B., Sarwar, S.M., Wei, T., Zamani, H.: Scalable and effective generative information retrieval. In: Proceedings of the ACM on Web Conference 2024. WWW'24, pp. 1441–1452. Association for Computing Machinery, New York (2024). https://doi.org/10.1145/3589334.3645477
141. Zeng, H., Luo, C., Zamani, H.: Planning Ahead in Generative Retrieval: Guiding Autoregressive Generation through Simultaneous Decoding. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 469–480 (2024)
142. Wu, S., Wei, W., Zhang, M., Chen, Z., Ma, J., Ren, Z., de Rijke, M., Ren, P.: Generative retrieval as multi-vector dense retrieval. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1828–1838 (2024)
143. Zhan, J., Mao, J., Liu, Y., Guo, J., Zhang, M., Ma, S.: Jointly optimizing query encoder and product quantization to improve retrieval performance. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management. CIKM '21, pp. 2487–2496. Association for Computing Machinery, New York (2021). https://doi.org/10.1145/3459637.3482358
144. Sachidananda, V., Kessler, J.S., Lai, Y.-A.: Efficient domain adaptation of language models via adaptive tokenization. arXiv preprint arXiv:2109.07460 (2021)
145. Huang, Q., Tao, M., Zhang, C., An, Z., Jiang, C., Chen, Z., Wu, Z., Feng, Y.: Lawyer llama technical report. arXiv preprint arXiv:2305.15062 (2023)
146. Cui, J., Li, Z., Yan, Y., Chen, B., Yuan, L.: ChatLaw: open-source legal large language model with integrated external knowledge bases. arXiv preprint arXiv:2306.16092 (2023)
147. Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., Mann, G.: BloombergGPT: a large language model for finance. arXiv preprint arXiv:2303.17564 (2023)
148. Dai, D., Dong, L., Hao, Y., Sui, Z., Chang, B., Wei, F.: Knowledge neurons in pretrained transformers. arXiv preprint arXiv:2104.08696 (2021)
149. Meng, K., Bau, D., Andonian, A., Belinkov, Y.: Locating and editing factual associations in GPT. Adv. Neural Inf. Proces. Syst. **35**, 17359–17372 (2022)
150. Liu, J., Yu, P., Zhang, Y., Li, S., Zhang, Z., Ji, H.: EVEDIT: event-based knowledge editing with deductive editing boundaries. arXiv preprint arXiv:2402.11324 (2024)
151. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
152. Li, H., Ai, Q., Chen, J., Dong, Q., Wu, Y., Liu, Y., Chen, C., Tian, Q.: Sailer: structure-aware pre-trained language model for legal case retrieval. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1035–1044 (2023)
153. Yao, F., Li, C., Nekipelov, D., Wang, H., Xu, H.: Human vs. Generative AI in Content Creation Competition: Symbiosis or Conflict? (2024)

Chapter 3
Interactions with Generative Information Retrieval Systems

Mohammad Aliannejadi, Jacek Gwizdka, and Hamed Zamani

Abstract Recent advancements in generative artificial intelligence have provided unique opportunities for seamless information access and discovery, particularly through natural language interactions. These technologies enable users to easily describe their needs and provide interactive feedback. This chapter provides an overview of the opportunities and challenges in interacting with information access systems powered by generative artificial intelligence technologies. We focus on user interfaces in these systems and various interactions for describing and clarifying users' needs, refining the result list produced by the system, providing proactive feedback to the system, the system proactively initiating conversations, explaining the result list, and enabling multi-modal interactions for information access.

3.1 Introduction

At its core, information access and seeking is an interactive process. In existing search engines, interactions are limited to a few pre-defined actions, such as "requery," "click on a document," "scrolling up/down," "going to the next result page," "leaving the search engine," etc. A major benefit of moving toward generative Information Retrieval (IR) systems is enabling users with a richer expression of information need and feedback and free-form interactions in natural language and beyond. In other words, the actions users take are no longer limited by the clickable links and buttons available on the search engine result page, and users

M. Aliannejadi
University of Amsterdam, Amsterdam, The Netherlands
e-mail: m.aliannejadi@uva.nl

J. Gwizdka
University of Texas at Austin, Austin, TX, USA
e-mail: jacekg@utexas.edu

H. Zamani (✉)
University of Massachusetts Amherst, Amherst, MA, USA
e-mail: zamani@cs.umass.edu

R. W. White, C. Shah (eds.), *Information Access in the Era of Generative AI*,
The Information Retrieval Series 51, https://doi.org/10.1007/978-3-031-73147-1_3

can express themselves freely through natural language. This can go even beyond natural language, through images, videos, gestures, and sensors using multi-modal generative IR systems. This chapter briefly discusses the role of *interaction* in generative IR systems. We will first discuss different ways users can express their information needs by interacting with generative IR systems (Sect. 3.2). We then explain how users can provide explicit or implicit feedback to generative IR systems and how they can consume such feedback (Sect. 3.3). Next, we will cover how users interactively can refine retrieval results (Sect. 3.4). We will expand upon mixed-initiative interactions and discuss clarification and preference elicitation in more detail (Sect. 3.5). We then discuss proactive generative IR systems, including context-aware recommendation, following up past conversations, contributing to multi-party conversations, and feedback requests (Sect. 3.6). Providing explanations is another interaction type that we briefly discuss in this chapter (Sect. 3.7). We will also briefly describe multi-modal interactions in generative information retrieval (Sect. 3.8). Finally, we describe emerging frameworks and solutions for user interfaces with generative AI systems (Sect. 3.9). We conclude with a question: Will the myriad interaction possibilities afforded by generative AI systems be embraced by a broad user base, or will they remain merely a research curiosity?

3.2 Expressing Information Needs

An information need is what prompts users to seek information through various means, such as asking others, consulting printed resources, other media, or searching online. It arises from the awareness of a gap in a user's knowledge or understanding, necessitating the acquisition of information to bridge that gap [12, 25]. Bridging the gap helps fulfill a specific purpose or goal, which is typically driven by a work task [13].

Prompt-based interactions with Large Language Models (LLMs), and, more broadly, multi-modal interactions with LLMs-based systems, provide an opportunity to fundamentally rethink the processes of searching for, finding, and using information and how to support these activities. This fresh perspective has the potential to significantly transform the user experience by enhancing how users express their information needs and achieve their goals.

We will frame our considerations using the information need model proposed by Robert Taylor in the 1960s [69, 70]. Taylor identified four levels of information need, each helping us understand how users formulate questions in their minds, how they articulate them, and how they interact with information systems. The four levels of information need are (1) **visceral need**, an inexpressible, unformulated need, felt as a vague sense of dissatisfaction; (2) **conscious need**, where the user is aware of the need but cannot fully articulate it; (3) **formalized need**, which can be clearly expressed and defined; and (4) **compromised need**, which is the articulated need, as presented to an information system, often simplified or altered to fit the system's capabilities.

Traditional search systems typically support levels 3 and 4, but not 1 and 2. We believe that LLMs-based information access systems have the potential to support all four levels. Therefore, we use these four levels to structure our speculative list of ways users could be assisted in their interactions with generative AI. We will draw, in part, on well-known information-seeking models [40, 45].

Support for **visceral need**: (1) *Exploratory interactions* provide users with broad, exploratory dialogue that might help users *clarify* their thoughts and suggest related topics to help users better understand and articulate their needs. This is an example of Clarification, which we describe in Sect. 3.5. (2) *Prompt suggestions* offer prompt suggestions or follow-up questions to guide users toward more specific questions. This is an example of Proactive Interactions, which we describe in Sect. 3.6.

Support for **conscious need**: (1) *Partial expression of needs*: accept partially formed questions or statements of need. (2) *Proactive support for refinement*: generate relevant information that helps users *refine* their understanding of what they're looking for. (3) *Guided conversations*: engage in a dialogue to help users articulate their needs more precisely. We describe such approaches in more detail in Result Refinement (Sect. 3.4) and Proactive Interactions (Sect. 3.6).

Support for **formalized need**: (1) *Direct queries*: respond directly to well-formulated questions with relevant information. (2) *Structured responses*: provide detailed, structured responses that address specific aspects of the user's need. (3) *Advanced features*: offer options (e.g., filters) for further exploration or *clarification* based on the formalized need.

Finally, support for **compromised need**: (1) *Flexibility of syntax*: offer flexibility to allow for iterative refinement of queries without strict syntax requirements. (2) *Flexibility of language*: interpret and respond to a wide range of query formats, reducing the need for users to adapt their language significantly. (3) *Feedback loop*: offer feedback on questions and suggesting modifications or alternative phrasings to better match the user's needs and the system's capabilities. We describe such approaches in more detail in Proactive Feedback (Sect. 3.3).

Overall, the key advantages of LLMs in assisting users at all four levels of information need are:

- **Natural language processing**: LLMs can understand and respond to queries expressed in natural language, making them accessible even at the visceral and conscious need levels.
- **Contextual understanding**: Advanced LLMs can maintain context over multiple interactions, allowing for a more nuanced exploration of information needs.
- **Broad knowledge base**: LLMs draw upon a vast range of information, potentially addressing needs across various domains and levels of specificity.
- **Adaptive responses**: LLMs can tailor their responses based on the perceived level of the user's information need, understanding and responding to both simple and complex questions and providing more or less detail as appropriate.
- **Iterative refinement**: The conversational nature of LLMs interactions allows users to refine their queries progressively, moving from visceral to formalized needs through dialogue.

– **Enhanced expressiveness**: Prompt-based interactions allow users to express their needs in more nuanced and detailed ways. Users can specify the format, tone, and depth of the information they seek, which can lead to more tailored and useful outputs. For instance, users can request summaries, detailed explanations, comparisons, or creative content, depending on their needs.

However, it is important to note that while LLMs offer powerful capabilities in addressing information needs across Taylor's levels, they also have limitations. They may sometimes provide plausible-sounding but incorrect information, lack true understanding of context beyond the immediate conversation, and cannot replace the critical thinking and expertise of human information professionals in complex scenarios.

While LLMs can offer enhanced capabilities for expressing information needs, they also introduce new challenges. Such as *capability gap*: users may struggle to formulate their intentions clearly and effectively, leading to a gap between what they want and what the LLMs provides. *Instruction gap*: users need to learn how to craft effective prompts, which can involve understanding the LLMs's capabilities and limitations. *Evaluation of outputs*: users must critically evaluate the LLMs's responses for accuracy and relevance, as LLMs can sometimes generate incorrect or misleading information. A recent paper introduced these three gaps and termed them collectively the *Gulf of Envisioning* [64].

In the following sections, we address selected aspects of user-LLMs-based-system-interactions, Sects. 3.3, 3.4, 3.5, 3.6, 3.7, and 3.8. In Sect. 3.9, User Interfaces, we discuss recent user interface frameworks and solutions.

3.3 Proactive Feedback

Recent developments in LLMs have paved the path toward complex interactions between the user and the system. Generative IR models are able to satisfy users' information needs in multiple interaction turns. Among many possibilities, this enables users to provide feedback to the system. Feedback can be provided when it is explicitly requested by the system, for example, in the form of clarifying questions or preference elicitation [2, 49, 52, 82]. Section 3.5 discusses these aspects in more detail. Feedback can be also requested for assessing the quality of the system at the end or in the middle of a conversation. For instance, Amazon's Alexa Prize Challenge [51] has sought explicit rating feedback from users upon the completion of the conversation. Zamani et al. [85] introduce the possibility of improving this simple approach by asking context-aware questions for feedback and making natural language interactions within the conversation.

Feedback can be provided proactively by the user, which is the focus of this section. Perhaps the simplest type of feedback that users provide can be in the form of *repeating or reformulating the user's need in the same search session*. If detected, this often means that the user's need has not been addressed yet. Besides

such simple scenarios, users may provide *explicit positive or negative feedback*. Explicit positive feedback are often easier to identify and interpret. They are often in the form of appreciation and hold a positive sentiment. Explicit negative feedback, on the other hand, is more challenging, more diverse, and perhaps more important for system designers as they help the system improve and identify its limitations. Pointing out what parts of the system's response is inaccurate and why it does not satisfy the user's needs and expressing frustration and disappointment are examples of explicit negative feedback. Current state-of-the-art technologies often cannot successfully take advantage of explicit negative feedback and often limit themselves to acknowledging the system's limitations and apologizing to users. There is huge potential in successfully comprehending negative feedback from users.

In generative IR systems, grounding as relevance feedback is also relevant to the concept of explicit feedback. Trippas et al. [71] define grounding as discourse for the creation of mutual knowledge and beliefs. Examples include providing indirect feedback by reciting their interpretation of the results. This process can potentially enable Conversational Information Seeking (CIS) systems to better understand a user's awareness of the results, background knowledge, or information need.

We would like to highlight the potential in providing implicit feedback as well. Progress in commercial (Web) search engines is in debt to large-scale implicit feedback collected from user interactions, such as clicks, skipped results, dwell time, and cursor (mouse) movement. Implicit feedback in generative IR systems is more challenging, because it is more likely to deal with abandonment in each session. This means that users may leave the system as they receive the answer they want without providing any positive feedback. Alternatively, they may leave the system as they lose hope in getting the right answer from the system. Besides abandonment, changing topics and asking follow-up questions can be interpreted as an implicit feedback signal in generative IR. Interpreting these user behaviors is essential in improving generative IR systems.

Research in understanding and modeling implicit (negative) feedback is relatively sparse, and future technologies can greatly benefit from further research in this space.

3.4 Result Refinement

3.4.1 An Overview of Result Refinement

Result refinement is relatively understudied, compared to other modes of interaction in generative IR. Search result refinement has a long history of research in IR, especially in areas such as information filtering (e.g., recommender systems) where users access semi-structured information [16]. Figure 3.1 shows an example search result page from Amazon.com, where users are able to select certain attributes of the items (e.g., size) in the catalog to narrow down the results being presented

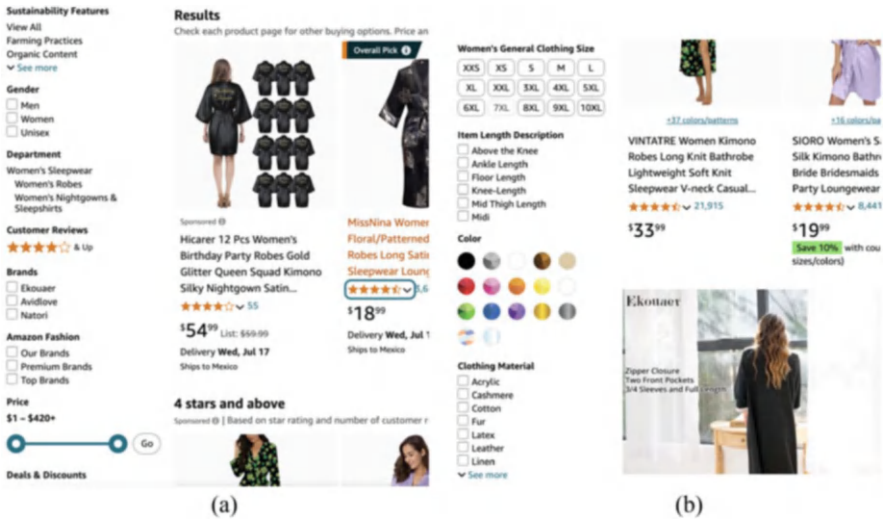

Fig. 3.1 Examples of search result refinement from Amazon.com. The refinement panes on the left help users browse through the search results

to them. Search result refinement for semi-structured data is a relatively trivial task, as the refinement pane usually concerns the most important attributes of the items, given the query and the top item list. In the preference-based search literature, example-critiquing approaches have been explored [73], where the model suggests examples to the user, and with the user's feedback, it then models the user's preference. In conversational recommender systems, a similar approach is taken as part of the preference elicitation process [37]. In this process, the conversational system starts the conversation by asking the user's opinion about movies, aiming to optimize the decision space. A similar approach is taken in conversational product recommendation [95, 96]. In these works, the high-level idea is to extract important attributes from user reviews of products and model a probabilistic decision space. Then the conversational system takes a greedy approach in which, at every step, it aims to ask about an item attribute that minimizes the uncertainty of the decision space. Search result refinement is more challenging in Web search, where the system deals with unstructured data. One of the earliest, simplest, and yet most effective ways is using vertical in the search result page [8]. Search result verticals divide the search results based on very high-level categories, such as images, videos, and news. Even though very high level, it still can be considered as a naïve approach to refinement, as it approaches the user information from the result type. In most cases, the same user query can be satisfied with different modalities, which turns out to be one of the most important aspects of search, hence major commercial search engines still employ this approach. Finally, some early approaches tried to diversify but also refine search results based on automatically extracted information facets. Faceted search [72] provides a means of navigation through topic facets for users, enabling

them to narrow down their information needs, as well as the search space. These early systems mainly relied on automatic facet extractors [36].

3.4.2 Technical Challenges

In the generative era, result refinement faces both algorithmic and interactive challenges.

Algorithmic Challenges As the items or documents are being represented using model parameters, refining the results based on a single attribute of the item is less trivial. To address this challenge, several works study controllable recommendation via disentanglement [17], where the goal is to represent items as separated attribute vectors instead of a single latent vector. Some of these attributes would be mapped to actual attributes in the catalog (e.g., color, style) or some latent attributes. LLMs have shown to be capable of extracting query facets, relying solely on their intrinsic knowledge [41]. However, as shown in the literature, LLMs are not yet capable of effectively grounding [63], which leads to suboptimal planning of LLMs utilizing their intrinsic knowledge to take the best next action. For example, in conversations where most humans would ask for refinement, LLMs fail to take the same action.

Interactive Challenges As mentioned above, there has been research on various modes of refinement, i.e., search verticals, item attributes, faceted search, and example critique. While each of these modes has been utilized for a specific interaction medium (e.g., Web search vs. conversational search), generative systems could potentially mix them, for example, prompting the user about their preferred search result modality, rather than making an assumption. Moreover, Chen et al. [14] review the interactive challenges of LLMs in the light of personalization, highlighting the importance of user–system interactions in result presentation, specifically refinement. Among other challenges, they refer to laborious data collection for training LLMs to be effective interactive systems, which can hinder the learning process.

3.5 Clarification

In a generative retrieval setting where the system aims to provide a comprehensive response to the user, whether in a conversational or Web search setting, it is of utmost importance to ensure that the user's intent is predicted with high confidence. This is particularly critical, as in traditional Web search scenarios, the system would diversify the list of results to ensure that various facets or interpretations of the query are covered in the top results [57]. However, in a generative scenario, usually, a single answer is provided to the user, limiting the information that can be exchanged between the user and the system.

3.5.1 An Overview of Search Clarification

Clarifying questions have been studied extensively [34] in the context of conversational question-answering [52], information-seeking conversations [2], and Web search [82].

Another line of research studies the role of mixed-initiative interactions for user preference elicitation [37, 49]. The goal here is to understand the user preference when multiple documents (items) can be deemed relevant to their information need. Radlinski et al. [49] study this problem for movie recommendation, where the user information need is typically generic (e.g., "romantic movies") with multiple potentially relevant items. The dialogue system's goal in this setting is to engage in a conversation to elicit user preference in a more fine-grained way.

There has been a body of research studying the effect of mixed-initiative interventions such as clarifying questions on user experience [35, 84, 97, 98]. Kiesel et al. [35] study the effect of voice query clarification on user experience and find, even in cases where the system performance is not improved, users have better experience. In Web search, Zamani et al. [84] study the effect of incorporating a clarification pane on the search result page, implemented in Bing.com. Analyzing the click logs, they find that the clarification pane improves user experience. More specifically, among the seven templates they use to generate the clarifying questions, they find clear preference towards certain question templates in terms of user engagement. Zou et al. [97] study the effect of the clarification pane in the same setting in a controlled experimental setup where they introduce three quality levels and measure user satisfaction and performance. They find that asking a low-quality question in a search session risks lower user engagement with questions of higher quality in the same session. This finding was confirmed in follow-up work [98].

User engagement (i.e., click-through rate) can be considered as a user-oriented quality measure of clarifying questions. Sekulic et al. [58, 60] extract various search result pages (SERPs)- and document-based features to predict user engagement while interacting with clarifying questions in a Web-based interface [83]. Rahmani et al. [50] study the effect of various query- and question-based features to predict user satisfaction in the MIMICS dataset [83] where they find, among others, a positive sentiment in the clarifying question leads to higher user satisfaction. Sekulic et al. [61] instead predict the usefulness of clarifying questions in the retrieval pipeline. Following an early study on the effect of different types of clarifying questions on retrieval performance [38], they train a classifier to predict the usefulness of a clarifying question and its answer in the retrieval pipeline and incorporate it in the retrieval pipeline if only it is predicted to be useful.

3.5.2 Technical Challenges

Planning While the early works in this area focused mainly on ranking clarifying questions from a pre-collected question bank [3, 5], more recent studies aim toward leveraging the generation power of LLMs to generate clarifying questions [91]. However, generative systems based entirely on LLMs are not effective in proactive interactions, especially in generating clarifying questions when necessary [24, 63]. Initial experiments reveal the power of LLMs in understanding the context of a query or a search session [1] and generate potential questions based on the context when prompted [21]; however, they fail at planning when to ask and which question to ask [21, 63]. Shaikh et al. [63] conduct a study where they compare human–human conversations with system–human conversations and find that LLMs fail at effectively planning when to ask clarifying questions in a conversation, even though they can generate high-quality questions if they are explicitly prompted to do so. Deng et al. [22] propose a proactive chain-of-thought approach to enhance the planning capability of LLMs such as ChatGPT and show that it has a considerable effect on their interaction capabilities.

Evaluation Evaluating generative systems comes with various challenges. On top of that, evaluating interactive generative systems involves even more challenges as the user response to a system output is required. A line of research looks at simulating and modeling the user–system interactions in a mixed-initiative setting [4, 10, 11, 48, 55, 59, 88]. User simulation can be beneficial to generative IR models in two ways: (i) they provide a means for evaluating generated content, and (ii) they can be used for training. Zhang and Balog [88] propose a user simulator for conversational recommendation to evaluate the system performance. This is followed by the work done by Sekulic et al. [59] and Owoicho et al. [48] in using GPT-based models to simulate users in a mixed-initiative information-seeking conversational system where the main goal of the simulator is to provide an answer to a generated clarifying question. They show that such simulators can lead to reliable evaluation of conversational systems.

There are various considerations to take into account in simulating and evaluating interactive generative systems:

- User effort: In interacting with the system, users bear different levels of cognitive load, which can lead to user fatigue as the number of interactions increases.
- User information gain: To model the true value of a clarifying question in a conversation, we need to model both the gain and effort a clarifying question brings to the conversation [4, 10].
- Information nuggets: Information gain can be modeled by breaking the user's information need into information nuggets and measuring how much asking a certain clarifying question would help us provide further information nuggets to the user.

– User model: As proposed by Balog [11], an effective user simulator should have various components, including a user mental model. Realistically, a single user simulator does not cover the needs and behavior of the wide range of users interacting with the system.

3.6 Proactive Interactions

Typically, users initiate the interaction with a generative retrieval system, for example, by submitting a chitchat utterance, asking a question, or submitting an action request. In mixed-initiative conversational systems, the agent is also able to initiate the conversation. This is also called a *proactive*, system-initiative, or agent-initiative conversation. Existing generative AI systems are relatively underdeveloped when it comes to proactive interactions [43]. A major reason is that initiating a conversation by the system is not only challenging but can also be risky; frequent and non-relevant proactive interactions may annoy users and hurt user satisfaction and trust [85]. Therefore, whether and when to initiate a proactive interaction are the key decisions a proactive CIS system should make.

3.6.1 An Overview of Proactive Generative Retrieval Systems

Wadhwa and Zamani [74] explored proactive conversational information access systems, discussing their challenges and opportunities. The authors introduced a taxonomy of proactive interactions, delineating three dimensions: (1) initiation moment (*when* to initiate a conversation), (2) initiation purpose (*why* to initiate a conversation), and (3) initiation means (*how* to initiate a conversation). They identified five purposes for initiating interactions: (1) filtering streaming information, (2) context-aware recommendation, (3) following up a past user–system conversation, (4) contributing to a multi-party human conversation, and (5) requesting feedback from users. A generic pipeline for these systems is depicted in Fig. 3.2. In this pipeline, several algorithms constantly monitor the user's context and information streams to produce conversation initiation instances, which are stored in a database. A conversation initiator component then selects an appropriate instance based on the situation, initiating a fluent and accurate utterance. Figure 3.2 is sufficiently generic for illustrating proactive interactions in generative retrieval models, and we use it to describe research and open questions in proactive retrieval in more detail.

Initiating a conversation through recommendation stands as one of the most common scenarios for proactive interaction. For instance, a conversational information access system might suggest an item based on the user's situational context, such as their location, time, and preferences. It is worth noting the distinction from traditional conversational recommendation setups, where users typically initiate the conversation by requesting specific items [66, 89]. Recent efforts in joint modeling

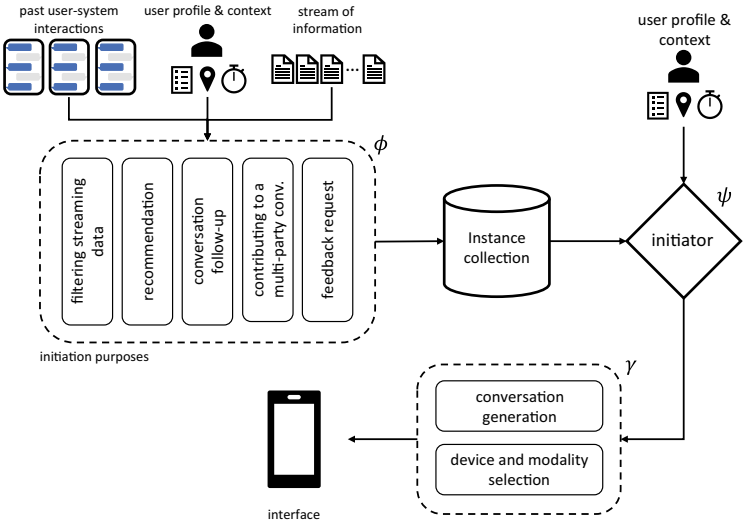

Fig. 3.2 A generic pipeline for conversation initiation in CIS systems by Wadhwa and Zamani [74]

of search and recommendation and developing unified information access systems [80, 81, 87] represent a step toward developing proactive, and thus mixed-initiative, systems in search and recommendation. However, proactive conversations extend beyond mere recommendations.

For example, Avula and Arguello [9] devised a system for conducting wizard-of-Oz experiments, investigating proactive interactions during conversational collaborative search. This system could seamlessly integrate into collaborative platforms like Slack,[1] where during a collaborative search task, an external user (acting as a wizard) provides information. Though advancements in this area are nascent, there exists considerable potential for systems to initiate context-based conversations, engaging users and eliciting feedback.

Consider a scenario where a user employs a mapping application to navigate to a restaurant. Leveraging contextual cues, a proactive generative retrieval system could subsequently initiate a conversation upon the user's return journey, inquiring about their dining experience. Such interactions not only enhance user engagement but also facilitate feedback collection, aiding in profile refinement. Similarly, in situations where a user encounters difficulty in task completion, a conversational system could autonomously engage in conversation, offering assistance [85].

[1] https://slack.com/

3.6.2 User Responses to Proactive Interactions

In generative retrieval systems, users have the freedom to provide a natural language response in any form, and they can be categorized as follows [74]:

- **Null action:** Users provide no response to the initiated conversation. It is important to note that null action should not necessarily be construed as negative feedback, as users may find the initiation useful but may not desire further engagement.
- **Interruption or negation:** Users respond in a manner consistent with terminating any further engagement by the generative retrieval system. It is perhaps safe to interpret such responses as negative feedback.
- **Relevant response:** Users provide a pertinent response to the initiated interaction, typically occurring when the interaction involves a question or solicits feedback.
- **Postpone:** Users respond to the initiated conversation and request the system to remind them at a later time.
- **Critique or clarification-seeking response:** Users engage further with the generative retrieval system, either seeking more information or critiquing existing engagement.
- **Follow-up:** Users provide a follow-up response to obtain additional information or perform actions related to the initiated conversation.
- **Topic drift:** Users respond but shift the topic of the initiated conversation.

3.6.3 Technical Challenges

Here, we outline key technical hurdles in implementing the pipeline shown in Fig. 3.2.

Producing System-Initiative Instances The initial step in the system-initiation pipeline involves identifying reasons for initiating a conversation and generating a proactive instance. Proactive instances encapsulate all relevant information about a conversation, including its purpose, content, and context. This process entails addressing each initiation purpose component outlined in Fig. 3.2. While some purposes, such as filtering streaming information and recommendation, have received attention in the literature, others like following up a past conversation or contributing to a multi-party conversation remain relatively unexplored. Thus, a major technical challenge lies in developing models capable of identifying the reasons for conversation initiation across various goals, including filtering information, recommendation, conversation follow-up, contributing to multi-party conversations, or requesting feedback.

Developing an Initiator Model The subsequent step involves selecting a proactive instance from the instance collection using an initiator component. The primary

challenge in this component stems from our limited understanding of the optimal moment to initiate a conversation. Consequently, future research should emphasize conducting user studies to explore the ideal timing for conversation initiation. Weak signals gleaned from user interactions with existing conversational systems, even those lacking proactive capabilities, could provide valuable insights. For instance, instances, where users initiate trivial conversations (e.g., out of boredom), could serve as noisy but potentially useful signals for predicting optimal conversation initiation moments. Machine learning models trained on situational context and user profiles could leverage such signals. Furthermore, interactive systems that log user interactions offer the opportunity to iteratively refine prediction accuracy based on user feedback.

Generating System-Initiative Utterances The final step entails generating a (natural language) interaction based on a proactive instance and presenting it to the user. Techniques from dialogue systems and text generation research can be leveraged for this purpose. Since users typically do not anticipate proactive utterances, a notable technical challenge lies in providing context within the generated utterance to ensure user comprehension. This context could reference previous interactions with the system, user experiences, or explanations regarding the rationale behind initiating the conversation. Given that each instance is a structured data object, neural models designed for unstructured text generation from structured data, such as tables, could be potentially useful.

3.6.4 Evaluation of Proactive Systems

Assessing proactive generative IR systems poses significant challenges. While IR research has traditionally focused on creating collections for specific information-seeking tasks, these collections are typically based on predefined needs (e.g., Text REtrieval Conference (TREC)[2] tracks) or observations (e.g., clickthrough data). However, these evaluation methods do not readily apply to scenarios involving proactive interactions. Although evaluating proactive generative IR systems remains largely unexplored in the literature, we can envision two classes of evaluation methodologies: (1) modular evaluation and (2) end-to-end evaluation.

In modular evaluation, the quality of each component in Fig. 3.2 is evaluated in isolation. For example, how accurate is the initiator component in identifying opportune moments for proactive interactions? This methodology simplifies evaluation in proactive systems, but does not provide a complete picture of the overall performance of the system from the user's perspective and does not reflect real-world complexities.

[2] https://trec.nist.gov/

In end-to-end evaluation, one can explore both offline and online evaluation strategies. For offline evaluation, each instance would encompass all necessary information for the system at a given timestamp, including past user–system interactions, user profiles, situational contexts, and streams of new information. The model's performance would then be assessed based on the generated proactive interactions, if applicable. Crafting a single evaluation metric capable of capturing all facets of conversation initiation evaluation presents a challenge, necessitating further investigation. Recently, Samarinas and Zamani [56] introduced a large-scale benchmark for proactive interactions to ongoing multi-party human conversations and proposed normalized proactive discounted cumulative gain (npDCG) for end-to-end evaluation of such systems. In a separate investigation, Sen et al. [62] suggested evaluating proactive recommendation within search sessions by aggregating a correlation measure over the session. This measure assesses the relationship between the expected outcome—comprising the list of documents retrieved with a true user query—and the predicted outcome, representing the list of documents recommended by a proactive search system.

In the realm of online evaluation, conventional A/B tests can serve as a valuable tool for assessing the system's efficacy. Additionally, interpreting user feedback—both positive and negative—can provide valuable insights into system performance.

3.7 Explanation

Explanation can be seen as a critical tool in search result presentation in generative systems, as users are interested in comprehensive justification and explanation of the presented results [14, 27]. Also, it can lead to more user trust in the results, potentially aiding the user to distinguish between a low-quality and a high-quality response.

3.7.1 An Overview of Explanation in Information Retrieval

Zhao et al. [94] provide a survey on the explainability of LLMs where they provide a taxonomy of explanations, together with methods for explaining Transformer-based LLMs. Also, they discuss various methods for evaluating explanations for both local and global explanations. Krishna et al. [39] show that not only are explanations useful in user–system interactions, but they also improve the performance of LLMs. They study automatic rationale generation in a Chain of Thought (CoT) manner. Deng et al. [23] show that rephrasing the user input leads to a better understanding of the user request, which in turn results in better performance of the LLM, which is complementary to CoT reasoning. In their tutorial, Anand et al. [6, 7] review Transformer-based explanation generation. Zhang et al. [90] address search explainability via the lens of query understanding, where the system's task is to

predict the user intent considering their query as input. LiEGe [79] explains all the documents in the ranking jointly using a listwise explanation generator.

Evaluating explanations is challenging. For free-text generations, human evaluation is employed. In other cases, because of a lack of explanation, proxy explanations such as clicks, query descriptions, query aspect annotation, and topic annotation can be used. For feature-based models, explanations are evaluated based on the effectiveness of predicted features. As for counterfactual explanations, model-based evaluation is employed.

3.7.2 Modes of Explanation in Generative Information Retrieval

The main mode of explanation used in generative models is free-form text, where the model would further elaborate why the provided answer is relevant to the user's input. The explanation often consists of two major parts: (i) a further description of user information need and (ii) an explanation of the reasons why the generated response is relevant to the user's input. The system has a limited information bandwidth and cannot present users with multiple intents of their query. Therefore, describing what the system "thinks" the user wants helps the user understand whether the system understands their intent or not [90]. This type of explanation aims to ensure the user that their information need is properly understood by the system and can lead to increased trust in the system. Also, in case of misunderstanding the user's information need, it provides the opportunity for the user to realize what is missing in their input. This can be seen as similar to scanning the Search Engine Result Page (SERP) by the user, through which the user would have an idea if the system understands their information need correctly.

Another form of explanation is to provide citations. This has been studied more extensively in the NLP community where the generated text is attributed via source citation [29]. The URL citations are supposed to provide evidence of the source of information from the Web. However, there are concerns regarding the quality of the citations, as there is no clear way of controlling the large language models (LLM) to ground its responses on the cited page [93]. Citing source documents while being useful as a form of explanation still does not provide a comprehensive idea of the relevance of the source. Comparing it to a typical SERP where the users are exposed to the URLs of the results, users already have a quality perception by scanning through the page title, summary, and URL. Even though the LLM-based search interfaces aim to mimic this experience, it is not yet clear which parts of the generated response are extracted from the cited document. Moreover, it is not clear how much the system depends on its intrinsic knowledge (i.e., model parameters) vs. the retrieved document. Therefore, more research in this area is required to understand how much different techniques and modes of explanation affect the users' perception of quality and trust. One potential alternative is to treat the system

as an information-gathering tool [53], rather than an information system. In such cases, the responses would look like "After searching the web, I found numerous sources of information about your query. Two of more trustworthy sources mention that . . . ." With such a response, not only does the user learn about the search space of the given query, but also they learn about the most important information extracted from the topic documents.

3.8 Multi-modal Interactions

Research has demonstrated the advantageous role of multimodal signals in both keyword-based and recommendation-driven searches, spanning from contextual item recommendations [33, 76] to visual and multimedia recommendations [46]. These signals also address challenges like cold-start issues [15, 18, 47] and aid in explaining and visualizing recommendation outcomes [68]. A recent survey by Deldjoo et al. [19] offers insights into the role of multimedia content in recommendation systems, delineating how such content—comprising audio, visual, and textual elements—enriches real-world recommendation challenges.

A significant challenge in multimedia information systems lies in fusing multiple modalities to derive meaningful representations. Recent advancements in multi-modal large language models employ joint representation techniques to establish a latent space where multiple modality information can be compared. However, aligning content data like text and images is relatively straightforward compared to aligning content with user preferences such as ratings or social media data.

Deldjoo et al. [20] explored multi-modal conversational information seeking tasks from multiple perspectives. They investigated (1) *why* multi-modal interactions should be used, (2) *which* tasks to support in multi-modal conversational systems, (3) *when* to integrate multiple modalities in conversations, and (4) *how* to research multiple modalities and conversations to enable multi-modal conversational information seeking. Deldjoo et al. [20] highlight the importance of each of these perspectives through a real-world example:

> Imagine a person is cycling along the road on their way to work. She is planning her day, including tasks from presenting a budget, hosting a new client, picking up their children after school, and making dinner. The cyclist passes a flower on the sideroad, which caught her eye and wanted to know what this plant is. Since she is cycling on a busy road, she quickly stops, takes a photo, and keeps riding. Meanwhile, she asks her earbuds to tell her which plant that was by a spoken query such as "what was that plant and is it edible?"

The authors argue that generative IR systems with multi-modal interactions and multi-modal sensors can accomplish the user's need in this and even more complex scenarios. Dealing with multi-modal interactions is a multidisciplinary topic, spanning across research areas from IR, recommender systems, multi-media, human–computer interactions, computer vision, and even psychological and cognitive sciences. The intersection of the research areas that enable people to search for information through multi-modal conversations has not received the attention it

deserves, and it might partially be due to the complexity of the topic in terms of both modeling and evaluation. Prior work are mostly limited to two modalities (image and text), e.g., [67, 78], and further development in multi-modal foundation models [26, 42] and multi-modal retrieval-augmented generation models [54] is expected to speed up progress in this area.

3.9 User Interfaces

While in Sect. 3.2 we focused on general interaction methods to assist users in expressing their information needs when interacting with generative AI, in this section, we review recent work on interaction techniques and user interfaces for information access with LLMs. The design space is huge, and it is still under-researched and poorly understood. For example, out of approximately 750 pre-prints related to LLMs published on arXiv in the field of IR between 2020 and 2024, only 22 mentioned "user interface" in their abstracts.

New human–LLM interaction frameworks are only starting to emerge. For example, recent work [28] reviewed 73 papers published in HCI conferences since 2021 to investigate the dynamics of human–LLM interaction. Authors identified four key phases in the interaction flow and developed a taxonomy of four primary interaction modes. The four phases are *planning*, before an interaction; *facilitating*, during an interaction; *iterating*, refining an interaction; and *testing*, testing an interaction. The interaction modes include *standard prompting, user interface, context-based, and agent facilitator*. The *user interface* mode is of most interest to us as it enhances user interactions with LLMs beyond the conversational interface by improving input, output, iteration, and reasoning processes. This mode contains five approaches, which could be used separately or in combination. (1) *Structured prompt* approaches assist users in creating multi-component prompts, which could range from zero-shot to few-shot, and support specification of constraints. Tools like PromptMaker [31] combine prefixes, settings, and examples in prompt creation. (2) *Varying output* approaches allow users to specify output formats. Early examples like GenLine and GenForm [30] facilitate generation of user specified mixed outputs, such as HTML, JavaScript, and CSS code. User's control over output format allows for high level of personalization and, potentially, enhances consumption of information. (3) *Iteration of interaction* approaches include features such as debugging, error labeling, regenerating, and self-repairing, enabling users to refine their original prompts and workflows. BotDesigner [86], for instance, helps users identify and label errors within conversations and offers a "retry" button to regenerate outputs. (4) *Testing of interaction* facilitates the testing of various prompt variations, useful for quick testing of complex solutions. Tools like VISAR [92] use visual programming to enable rapid prototyping and testing of writing organization. (5) *UI to support reasoning* incorporates direct manipulation in the chain-of-thought process, allowing users to actively participate in and reorganize reasoning sequences. Other approaches in this area offer visual programming

techniques, such as chain designs and mind maps, and enable a more interactive and user-defined reasoning framework [32, 65, 92]. For example, Graphologue [32] introduced (1) graphical diagrams, which convert text-based responses from LLMs into diagrams; (2) graphical dialogues, which enable graphical, non-linear dialogues between humans and LLMs; and (3) interactive diagrams, which allow users to adjust graphical presentation and its complexity and submit context-specific prompts.

MacNeil et al. [44] explore three methods for integrating LLMs into user interfaces through a framework called Prompt Middleware. The three methods are (1) *Static Prompts*, which are predefined prompts generated by experts through prompt engineering. They can be invoked by using UI elements (e.g., buttons), allowing users to send high-quality prompts with minimal effort. This method leverages best practices but limits user control over prompt generation. (2) *Template-Based Prompts* involve generating prompts by filling in a template with options selected from the UI. The template integrates expertise and best practices, giving users more control through UI options. This method is exemplified by the FeedbackBuffet prototype, a writing assistant that uses template-based prompts to generate feedback on writing samples [44]. (3) *Free-Form Prompts* grants users full control over the prompting process. Although challenging, it is beneficial in scenarios where complete control is desired.

Wang et al. [75] present a proactive interface design that addresses challenges users face in initializing and refining prompts, providing feedback to the system, and managing cognitive load. They describe three interaction techniques (*Perception Articulation, Prompt Suggestions, Conversation Explanation*) and how they can be supported by user interface elements. Perception articulation is supported by a pre-task questionnaire and main prompt template—the first supports information need at the visceral level while the latter at the formalized level. Prompt suggestions are provided through supportive function tabs, which support conscious need. Conversation explanations are also delivered through supportive function tabs, with a feedback mechanism allowing users to rate the usefulness of these explanations. This feature supports compromised needs. Evaluation with participants demonstrated the effectiveness of these supportive functions in reducing cognitive load, guiding prompt refinement, and increasing user engagement. In interviews, participants appreciated the perception articulation functions for setting expectations and the conversation explanations for balancing expectations and satisfaction.

On one hand, the design space of user interfaces for LLMs offers a myriad of new interaction possibilities. On the other, taking advantage of the new possibilities can lead to complexity, which can make interfaces harder to comprehend and can overwhelm users. From the history of search interface evolution, we know that more complex search interfaces have not been widely accepted. For example, faceted search UIs led to a sharp learning curve and increased cognitive load [77]. History likes to repeat itself. Will it be the case with user interfaces for LLMs? Will the more complex interfaces for LLMs become only niche products?

3.10 Conclusions

As mentioned multiple times throughout this chapter, handling complex interaction types and modalities has been relatively under-explored, and the authors find it a rich area of investment for the further development of generative IR systems. This chapter pointed out prior work on various interaction types, from expressing information need to result refinement and mixed-initiative interactions, including clarification, feedback, and proactive interactions. Recent developments in (multimodal) foundation models, including LLMs, have paved the path toward better understanding complex user interactions, but we are still far from ideal generative IR systems that can satisfy user needs efficiently, effectively, fairly, and robustly.

Acknowledgments This work was supported in part by the Center for Intelligent Information Retrieval, in part by the Office of Naval Research contract number N000142212688, and in part by NSF grant number 2143434. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

References

1. Abbasiantaeb, Z., Yuan, Y., Kanoulas, E., Aliannejadi, M.: Let the LLMs talk: Simulating human-to-human conversational QA via zero-shot LLM-to-LLM interactions. In: Proceedings of the 17th ACM International Conference on Web Search and Data Mining, pp. 8–17. ACM, New York (2024)
2. Aliannejadi, M., Zamani, H., Crestani, F., Croft, W.B.: Asking clarifying questions in open-domain information-seeking conversations. In: SIGIR, pp. 475–484. ACM, New York (2019)
3. Aliannejadi, M., Kiseleva, J., Chuklin, A., Dalton, J., Burtsev, M.: Convai3: generating clarifying questions for open-domain dialogue systems (ClariQ). CoRR **abs/2009.11352** (2020)
4. Aliannejadi, M., Azzopardi, L., Zamani, H., Kanoulas, E., Thomas, P., Craswell, N.: Analysing mixed initiatives and search strategies during conversational search. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, pp. 16–26. ACM, New York (2021)
5. Aliannejadi, M., Kiseleva, J., Chuklin, A., Dalton, J., Burtsev, M.: Building and evaluating open-domain dialogue corpora with clarifying questions. In: EMNLP (1), pp. 4473–4484. Association for Computational Linguistics, Stroudsburg (2021)
6. Anand, A., Lyu, L., Idahl, M., Wang, Y., Wallat, J., Zhang, Z.: Explainable information retrieval: a survey. CoRR **abs/2211.02405** (2022)
7. Anand, A., Sen, P., Saha, S., Verma, M., Mitra, M.: Explainable information retrieval. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 3448–3451. ACM, New York (2023)
8. Arguello, J., Diaz, F., Callan, J.: Learning to aggregate vertical results into web search results. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, pp. 201–210. ACM, New York (2011)
9. Avula, S., Arguello, J.: Wizard of oz interface to study system initiative for conversational search. In: Proceedings of the 2020 Conference on Human Information Interaction and Retrieval. p. 447–451. CHIIR '20, Association for Computing Machinery, New York (2020). https://doi.org/10.1145/3343413.3377941. https://doi-org.ezp.lib.unimelb.edu.au/10.1145/3343413.3377941

10. Azzopardi, L., Aliannejadi, M., Kanoulas, E.: Towards building economic models of conversational search. In: European Conference on Information Retrieval (2). Lecture Notes in Computer Science, vol. 13186, pp. 31–38. Springer, Berlin (2022)
11. Balog, K.: Conversational AI from an information retrieval perspective: Remaining challenges and a case for user simulation. In: DESIRES. CEUR Workshop Proceedings, vol. 2950, pp. 80–90. CEUR-WS.org (2021)
12. Belkin, N.J.: Anomalous states of knowledge as a basis for information retrieval. Can. J. Inf. Sci. **5**, 133–143 (1980)
13. Byström, K., Hansen, P.: Conceptual framework for tasks in information studies. J. Am. Soc. Inf. Sci. Technol. **56**(10), 1050–1061 (2005). https://doi.org/10.1002/asi.20197
14. Chen, J., Liu, Z., Huang, X., Wu, C., Liu, Q., Jiang, G., Pu, Y., Lei, Y., Chen, X., Wang, X., Lian, D., Chen, E.: When large language models meet personalization: Perspectives of challenges and opportunities. CoRR **abs/2307.16376** (2023)
15. Cui, P., Wang, Z., Su, Z.: What videos are similar with you?: Learning a common attributed representation for video recommendation. In: Hua, K.A., Rui, Y., Steinmetz, R., Hanjalic, A., Natsev, A., Zhu, W. (eds.) Proceedings of the ACM International Conference on Multimedia, MM '14, pp. 597–606. ACM, New York (2014)
16. Cui, Z., Yu, F., Wu, S., Liu, Q., Wang, L.: Disentangled item representation for recommender systems. ACM Trans. Intell. Syst. Technol. **12**(2), 20:1–20:20 (2021)
17. Cui, Z., Yu, F., Wu, S., Liu, Q., Wang, L.: Disentangled item representation for recommender systems. ACM Trans. Intell. Syst. Technol. **12**(2), 20:1–20:20 (2021)
18. Deldjoo, Y., Dacrema, M.F., Constantin, M.G., Eghbal-zadeh, H., Cereda, S., Schedl, M., Ionescu, B., Cremonesi, P.: Movie genome: alleviating new item cold start in movie recommendation. User Model. User Adapt. Interact. **29**(2), 291–343 (2019)
19. Deldjoo, Y., Schedl, M., Cremonesi, P., Pasi, G.: Recommender systems leveraging multimedia content. ACM Comput. Surv. **53**(5), 106:1–106:38 (2020)
20. Deldjoo, Y., Trippas, J.R., Zamani, H.: Towards multi-modal conversational information seeking. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1577–1587. SIGIR '21, Association for Computing Machinery, New York (2021). https://doi.org/10.1145/3404835.3462806
21. Deng, Y., Lei, W., Lam, W., Chua, T.: A survey on proactive dialogue systems: Problems, methods, and prospects. In: IJCAI. pp. 6583–6591. ijcai.org (2023)
22. Deng, Y., Liao, L., Chen, L., Wang, H., Lei, W., Chua, T.: Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. In: EMNLP (Findings), pp. 10602–10621. Association for Computational Linguistics, Stroudsburg (2023)
23. Deng, Y., Zhang, W., Chen, Z., Gu, Q.: Rephrase and respond: Let large language models ask better questions for themselves. CoRR **abs/2311.04205** (2023)
24. Deng, Y., Zhang, A., Lin, Y., Chen, X., Wen, J., Chua, T.: Large language model powered agents in the web. In: WWW (Companion Volume), pp. 1242–1245. ACM, New York (2024)
25. Dervin, B., Nilan, M.: Information needs and uses. Ann. Rev. Inf. Sci. Technol. **21**, 3–33 (1986)
26. Fei, N., Lu, Z., Gao, Y., Yang, G., Huo, Y., Wen, J., Lu, H., Song, R., Gao, X., Xiang, T., Sun, H., Wen, J.: WenLan 2.0: Make AI imagine via a multimodal foundation model. Nat. Commun. **13**(1) (2022)
27. Gao, J., Wang, X., Wang, Y., Xie, X.: Explainable recommendation through attentive multi-view learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 3622–3629. AAAI Press, Washington (2019)
28. Gao, J., Gebreegziabher, S.A., Choo, K.T.W., Li, T.J.J., Perrault, S.T., Malone, T.W.: A Taxonomy for Human-LLM Interaction Modes: An initial exploration. In: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, pp. 1–11 (2024). https://doi.org/10.1145/3613905.3650786
29. Huang, J., Chang, K.C.: Citation: a key to building responsible and accountable large language models. CoRR **abs/2307.02185** (2023)

30. Jiang, E., Toh, E., Molina, A., Donsbach, A., Cai, C.J., Terry, M.: GenLine and GenForm: Two tools for interacting with generative language models in a code editor. In: Adjunct Proceedings of the 34th Annual ACM Symposium on User Interface Software and Technology, pp. 145–147. ACM, New York (2021). https://doi.org/10.1145/3474349.3480209
31. Jiang, E., Olson, K., Toh, E., Molina, A., Donsbach, A., Terry, M., Cai, C.J.: PromptMaker: Prompt-based prototyping with large language models. In: Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems, pp. 1–8. CHI EA '22, ACM, New York (2022). https://doi.org/10.1145/3491101.3503564
32. Jiang, P., Rayan, J., Dow, S.P., Xia, H.: Graphologue: Exploring large language model responses with interactive diagrams. In: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, pp. 1–20. ACM, New York (2023). https://doi.org/10.1145/3586183.3606737
33. Kaminskas, M., Ricci, F., Schedl, M.: Location-aware music recommendation using auto-tagging and hybrid matching. In: Seventh ACM Conference on Recommender Systems, RecSys '13, Hong Kong, October 12–16, 2013, pp. 17–24. ACM, New York (2013)
34. Keyvan, K., Huang, J.X.: How to approach ambiguous queries in conversational search: a survey of techniques, approaches, tools, and challenges. ACM Comput. Surv. **55**(6), 129:1–129:40 (2023)
35. Kiesel, J., Bahrami, A., Stein, B., Anand, A., Hagen, M.: Toward voice query clarification. In: The 41st International ACM Sigir Conference on Research & Development in Information Retrieval , pp. 1257–1260. ACM, New York (2018)
36. Kong, W., Allan, J.: Extracting query facets from search results. In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 93–102. ACM, New York (2013)
37. Kostric, I., Balog, K., Radlinski, F.: Generating usage-related questions for preference elicitation in conversational recommender systems. Trans. Recomm. Syst. **2**(2), 12:1–12:24 (2024)
38. Krasakis, A.M., Aliannejadi, M., Voskarides, N., Kanoulas, E.: Analysing the effect of clarifying questions on document ranking in conversational search. In: Proceedings of the 2020 ACM Sigir on International Conference on Theory of Information Retrieval, pp. 129–132. ACM, New York (2020)
39. Krishna, S., Ma, J., Slack, D., Ghandeharioun, A., Singh, S., Lakkaraju, H.: Post hoc explanations of language models can improve language models. In: Advances in Neural Information Processing Systems 36 (2023)
40. Kuhlthau, C.C.: Inside the search process: information seeking from the user's perspective. J. Am. Soc. Inf. Sci. **42**(5), 361–371 (1991–06)
41. Lee, J., Kim, J.: Enhanced facet generation with LLM editing. In: LREC/COLING, pp. 5856–5865. ELRA and ICCL (2024)
42. Li, C., Gan, Z., Yang, Z., Yang, J., Li, L., Wang, L., Gao, J.: Multimodal foundation models: From specialists to general-purpose assistants. Foundat. Trends® Comput. Graph. Vis. **16**(1–2), 1–214 (2024). https://doi.org/10.1561/0600000110
43. Liao, L., Yang, G.H., Shah, C.: Proactive conversational agents in the post-ChatGPT world. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 3452–3455. SIGIR '23, Association for Computing Machinery, New York (2023). https://doi.org/10.1145/3539618.3594250
44. MacNeil, S., Tran, A., Kim, J., Huang, Z., Bernstein, S., Mogil, D.: Prompt Middleware: Mapping Prompts for Large Language Models to UI Affordances (2023). https://doi.org/10.48550/arXiv.2307.01142
45. Marchionini, G.: Information Seeking in Electronic Environments. Cambridge University Press, Cambridge (1995)

46. McAuley, J.J., Targett, C., Shi, Q., van den Hengel, A.: Image-based recommendations on styles and substitutes. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, August 9–13, 2015, pp. 43–52. ACM, New York (2015)
47. Oramas, S., Nieto, O., Sordo, M., Serra, X.: A deep multimodal approach for cold-start music recommendation. In: Proceedings of the 2nd Workshop on Deep Learning for Recommender Systems, DLRS@RecSys 2017, Como, August 27, 2017, pp. 32–37. ACM, New York (2017)
48. Owoicho, P., Sekulic, I., Aliannejadi, M., Dalton, J., Crestani, F.: Exploiting simulated user feedback for conversational search: Ranking, rewriting, and beyond. In: SIGIR, pp. 632–642. ACM, New York (2023)
49. Radlinski, F., Balog, K., Byrne, B., Krishnamoorthi, K.: Coached conversational preference elicitation: A case study in understanding movie preferences. In: SIGdial, pp. 353–360. Association for Computational Linguistics, Stroudsburg (2019)
50. Rahmani, H.A., Wang, X., Aliannejadi, M., Naghiaei, M., Yilmaz, E.: Clarifying the path to user satisfaction: An investigation into clarification usefulness. In: EACL (Findings), pp. 1266–1277. Association for Computational Linguistics, Stroudsburg (2024)
51. Ram, A., Prasad, R., Khatri, C., Venkatesh, A., Gabriel, R., Liu, Q., Nunn, J., Hedayatnia, B., Cheng, M., Nagar, A., King, E., Bland, K., Wartick, A., Pan, Y., Song, H., Jayadevan, S., Hwang, G., Pettigrue, A.: Conversational AI: the science behind the Alexa prize. arXiv preprint 1801.03604 (2018)
52. Rao, S., III, H.D.: Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In: ACL (1), pp. 2737–2746. Association for Computational Linguistics, Stroudsburg (2018)
53. Ren, P., Liu, Z., Song, X., Tian, H., Chen, Z., Ren, Z., de Rijke, M.: Wizard of search engine: Access to information through conversations with search engines. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 533–543. ACM, Stroudsburg (2021)
54. Salemi, A., Altmayer Pizzorno, J., Zamani, H.: A symmetric dual encoding dense retrieval framework for knowledge-intensive visual question answering. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 110–120. SIGIR '23, Association for Computing Machinery, New York (2023). https://doi.org/10.1145/3539618.3591629
55. Salle, A., Malmasi, S., Rokhlenko, O., Agichtein, E.: Cosearcher: studying the effectiveness of conversational search refinement and clarification through user simulation. Inf. Retr. J. **25**(2), 209–238 (2022)
56. Samarinas, C., Zamani, H.: ProCIS: A benchmark for proactive retrieval in conversations. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '24, Association for Computing Machinery, New York (2024)
57. Santos, R.L.T., MacDonald, C., Ounis, I.: Search result diversification. Found. Trends Inf. Retr. **9**(1), 1–90 (2015)
58. Sekulic, I., Aliannejadi, M., Crestani, F.: User engagement prediction for clarification in search. In: Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part I 43. Lecture Notes in Computer Science, vol. 12656, pp. 619–633. Springer, Berlin (2021)
59. Sekulic, I., Aliannejadi, M., Crestani, F.: Evaluating mixed-initiative conversational search systems via user simulation. In: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, pp. 888–896. ACM, New York (2022)
60. Sekulic, I., Aliannejadi, M., Crestani, F.: Exploiting document-based features for clarification in conversational search. In: European Conference on Information Retrieval (1). Lecture Notes in Computer Science, vol. 13185, pp. 413–427. Springer, Berlin (2022)
61. Sekulic, I., Lajewska, W., Balog, K., Crestani, F.: Estimating the usefulness of clarifying questions and answers for conversational search. In: European Conference on Information Retrieval (3). Lecture Notes in Computer Science, vol. 14610, pp. 384–392. Springer, Berlin (2024)

62. Sen, P., Ganguly, D., Jones, G.: Procrastination is the thief of time: Evaluating the effectiveness of proactive search systems. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 1157–1160. SIGIR '18. Association for Computing Machinery, New York (2018). https://doi.org/10.1145/3209978.3210114
63. Shaikh, O., Gligoric, K., Khetan, A., Gerstgrasser, M., Yang, D., Jurafsky, D.: Grounding gaps in language model generations. In: Duh, K., Gomez, H., Bethard, S. (eds.) Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 6279–6296. Association for Computational Linguistics, Mexico City (2024). https://aclanthology.org/2024.naacl-long.348
64. Subramonyam, H., Pea, R., Pondoc, C., Agrawala, M., Seifert, C.: Bridging the gulf of envisioning: Cognitive challenges in prompt based interactions with LLMs. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, pp. 1–19. ACM, New York (2024). https://doi.org/10.1145/3613904.3642754
65. Suh, S., Min, B., Palani, S., Xia, H.: Sensecape: Enabling multilevel exploration and sensemaking with large language models. In: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, pp. 1–18. ACM, New York (2023). https://doi.org/10.1145/3586183.3606756
66. Sun, Y., Zhang, Y.: Conversational recommender system. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. p. 235–244. SIGIR '18, Association for Computing Machinery, New York (2018). https://doi.org/10.1145/3209978.3210002
67. Sundar, A., Heck, L.: Multimodal conversational AI: A survey of datasets and approaches. In: Liu, B., Papangelis, A., Ultes, S., Rastogi, A., Chen, Y.N., Spithourakis, G., Nouri, E., Shi, W. (eds.) Proceedings of the 4th Workshop on NLP for Conversational AI, pp. 131–147. Association for Computational Linguistics, Dublin (2022). https://doi.org/10.18653/v1/2022.nlp4convai-1.12. https://aclanthology.org/2022.nlp4convai-1.12
68. Tangseng, P., Okatani, T.: Toward explainable fashion recommendation. In: IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, March 1–5, 2020, pp. 2142–2151. IEEE, Piscataway (2020)
69. Taylor, R.S.: The process of asking questions. Am. Document. **13**(4), 391–396 (1962). https://doi.org/10.1002/asi.5090130405
70. Taylor, R.S.: Question-negotiation and information seeking in libraries. College Res. Libr. **29**(3), 178–194 (1968). https://doi.org/10.5860/crl_29_03_178
71. Trippas, J.R., Spina, D., Thomas, P., Sanderson, M., Joho, H., Cavedon, L.: Towards a model for spoken conversational search. Inf. Process. Manage. **57**(2), 102162 (2020). https://doi.org/10.1016/j.ipm.2019.102162
72. Tunkelang, D.: Faceted Search. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, San Rafael (2009)
73. Viappiani, P., Faltings, B., Pu, P.: Preference-based search using example-critiquing with suggestions. J. Artif. Intell. Res. **27**, 465–503 (2006)
74. Wadhwa, S., Zamani, H.: Towards system-initiative conversational information seeking. In: Proceedings of the Second International Conference on Design of Experimental Search and Information Retrieval Systems. pp. 102–116. DESIRES '21, CSUR (2021)
75. Wang, B., Liu, J., Karimnazarov, J., Thompson, N.: Task supportive and personalized human-large language model interaction: A user study. In: Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval. pp. 370–375 (2024). https://doi.org/10.1145/3627508.3638344
76. Wang, S., Wang, Y., Tang, J., Shu, K., Ranganath, S., Liu, H.: What your images reveal: Exploiting visual contents for point-of-interest recommendation. In: Proceedings of the 26th International Conference on World Wide Web, WWW 2017, pp. 391–400. ACM, New York (2017)

77. Wilson, M.: Evaluating the cognitive impact of search user interface design decisions. In: Proceedings of the 1st European Workshop on Human-Computer Interaction and Information Retrieval, pp. 27–30 (2011), http://ceur-ws.org/Vol-763/
78. Wu, Y., Macdonald, C., Ounis, I.: Multimodal conversational fashion recommendation with positive and negative natural-language feedback. In: Proceedings of the 4th Conference on Conversational User Interfaces. CUI '22, Association for Computing Machinery, New York (2022). https://doi.org/10.1145/3543829.3543837
79. Yu, P., Rahimi, R., Allan, J.: Towards explainable search results: A listwise explanation generator. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 669–680. ACM, New York (2022)
80. Zamani, H., Croft, W.B.: Joint modeling and optimization of search and recommendation. In: Proceedings of the First International Conference on Design of Experimental Search and Information Retrieval Systems, pp. 36–41. DESIRES '18, CSUR (2020)
81. Zamani, H., Croft, W.B.: Learning a joint search and recommendation model from user-item interactions. In: Proceedings of the 13th International Conference on Web Search and Data Mining, pp. 717–725. WSDM '20, Association for Computing Machinery, New York, NY (2020). https://doi.org/10.1145/3336191.3371818
82. Zamani, H., Dumais, S.T., Craswell, N., Bennett, P.N., Lueck, G.: Generating clarifying questions for information retrieval. In: Proceedings of the web conference 2020 WWW, pp. 418–428. ACM/IW3C2 (2020)
83. Zamani, H., Lueck, G., Chen, E., Quispe, R., Luu, F., Craswell, N.: MIMICS: A large-scale data collection for search clarification. In: CIKM, pp. 3189–3196. ACM, New York (2020)
84. Zamani, H., Mitra, B., Chen, E., Lueck, G., Diaz, F., Bennett, P.N., Craswell, N., Dumais, S.T.: Analyzing and learning from user interactions for search clarification. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 1181–1190. SIGIR '20, Association for Computing Machinery, New York (2020). https://doi.org/10.1145/3397271.3401160
85. Zamani, H., Trippas, J.R., Dalton, J., Radlinski, F.: Conversational information seeking. Foundat. Trends® Inf. Retri. **17**(3–4), 244–456 (2023). https://doi.org/10.1561/1500000081
86. Zamfirescu-Pereira, J., Wong, R.Y., Hartmann, B., Yang, Q.: Why Johnny can't prompt: How non-AI experts try (and fail) to design LLM prompts. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, pp. 1–21. CHI '23, Association for Computing Machinery, New York (2023). https://doi.org/10.1145/3544548.3581388
87. Zeng, H., Kallumadi, S., Alibadi, Z., Nogueira, R., Zamani, H.: A personalized dense retrieval framework for unified information access. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 121–130. SIGIR '23, Association for Computing Machinery, New York (2023). https://doi.org/10.1145/3539618.3591626
88. Zhang, S., Balog, K.: Evaluating conversational recommender systems via user simulation. In: Proceedings of the 26th ACM Sigkdd International Conference on Knowledge Discovery & Data Mining, pp. 1512–1520. ACM, New York (2020)
89. Zhang, Y., Chen, X., Ai, Q., Yang, L., Croft, W.B.: Towards conversational search and recommendation: System ask, user respond. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, pp. 177–186. CIKM '18, ACM, New York (2018). https://doi.org/10.1145/3269206.3271776. http://doi.acm.org/10.1145/3269206.3271776
90. Zhang, R., Guo, J., Fan, Y., Lan, Y., Cheng, X.: Query understanding via intent description generation. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pp. 1823–1832. ACM, New York (2020)
91. Zhang, Q., Naradowsky, J., Miyao, Y.: Ask an expert: Leveraging language models to improve strategic reasoning in goal-oriented dialogue models. In: ACL (Findings), pp. 6665–6694. Association for Computational Linguistics, Stroudsburg (2023)
92. Zhang, Z., Gao, J., Dhaliwal, R.S., Li, T.J.J.: VISAR: A human-AI argumentative writing assistant with visual programming and rapid draft prototyping. In: Proceedings of the 36th

Annual ACM Symposium on User Interface Software and Technology, pp. 1–30 (2023). https://doi.org/10.1145/3586183.3606800
93. Zhang, W., Aliannejadi, M., Yuan, Y., Pei, J., Huang, J.H., Kanoulas, E.: Towards fine-grained citation evaluation in generated text: A comparative analysis of faithfulness metrics (2024). https://arxiv.org/abs/2406.15264
94. Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., Du, M.: Explainability for large language models: A survey. CoRR **abs/2309.01029** (2023)
95. Zou, J., Kanoulas, E.: Learning to ask: Question-based sequential Bayesian product search. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, pp. 369–378. ACM, New York (2019)
96. Zou, J., Chen, Y., Kanoulas, E.: Towards question-based recommender systems. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 881–890. ACM, New York (2020)
97. Zou, J., Aliannejadi, M., Kanoulas, E., Pera, M.S., Liu, Y.: Users meet clarifying questions: toward a better understanding of user interactions for search clarification. ACM Trans. Inf. Syst. **41**(1), 16:1–16:25 (2023)
98. Zou, J., Sun, A., Long, C., Aliannejadi, M., Kanoulas, E.: Asking clarifying questions: To benefit or to disturb users in web search? Inf. Process. Manag. **60**(2), 103176 (2023)

Chapter 4
Adapting Generative Information Retrieval Systems to Users, Tasks, and Scenarios

Johanne R. Trippas, Damiano Spina, and Falk Scholer

Abstract Generative Information Retrieval (GenIR) signifies an advancement in Information Retrieval (IR). GenIR encourages more sophisticated, conversational responses to search queries by integrating generative models and chat-like interfaces. However, this approach retains core principles of traditional IR and conversational information seeking, illustrating its capacity to augment current IR frameworks.

In this chapter, we propose that introducing GenIR enhances traditional information retrieval tasks and expands their scope. This allows systems to manage more complex queries, including generative, critiquing, and extractive tasks. These advancements surpass traditional systems, handling queries with greater depth and flexibility. This sometimes-speculative chapter suggests Generative Information Access (GenIA), a term that more accurately encapsulates the widened scope and enhanced functionalities of GenIR, particularly in how this relates to tasks. By investigating the impact of GenIR, this discussion aims to reiterate that generative research should not abandon traditional interactive information retrieval research but rather incorporate it into future research and development efforts.

4.1 Introduction

In an era where the volume of digital information expands more rapidly than ever, the ability of IR systems to sift through data to understand and anticipate users' information needs becomes even more essential. Additionally, in IR, the emergence of GenIR systems represents a paradigm shift in how we search for—and use found—information. The next-generation information access systems not only retrieve documents that may be relevant to a user's query but ideally can combine, synthesize, or abstract information, making the information directly applicable.

J. R. Trippas (✉) · D. Spina · F. Scholer
RMIT University, Melbourne, VIC, Australia
e-mail: j.trippas@rmit.edu.au; damiano.spina@rmit.edu.au; falk.scholer@rmit.edu.au

R. W. White, C. Shah (eds.), *Information Access in the Era of Generative AI*,
The Information Retrieval Series 51, https://doi.org/10.1007/978-3-031-73147-1_4

This chapter explores what GenIR systems may mean for users. We argue that even though Generative Artificial Intelligence (GenAI) could help us toward genuine interactive IR, building on existing research is crucial. We, therefore, contextualize the broader GenIA and stress the importance of keeping the user central. We aim to bridge the gap between established IR principles and new generative technologies, ensuring that user needs, tasks, and contexts drive innovation in IR. We explore the dynamic interaction between advanced GenIR technologies and the user's information needs, tasks, and queries. We examine users' tasks in IR systems, from simple fact-finding to complex, exploratory searches and transfer these tasks to generative information-seeking.

We consider how the context (e.g., academic research, learning and teaching, or personalized personal information management) affects the requirements and expectations of a GenIR system. We discuss the integration of context-aware technologies that adapt the retrieval process to fit the user's current environment, device, or application, thereby enhancing the relevance and utility of retrieved information.

Through theoretical insights and practical examples, this chapter aims to provide an overview of current strategies and emerging trends in GenIR. This chapter emphasizes the need for an approach considering the dynamic interaction between users, tasks, and scenarios. Such an approach helps researchers and practitioners develop more efficient, user-friendly information access systems. The significance of this chapter lies in demonstrating how GenIR systems can enhance the IR process by providing more flexible, advanced, and user-centric approaches. The integration of GenIR within the broader context of GenIA offers the potential for dynamic personalization and improved task understanding. Additionally, the chapter highlights the human's role in ensuring the relevance and reliability of GenIR outputs and the importance of ethical considerations and user privacy in evaluating these systems.

4.1.1 Chapter Overview

In Sect. 4.2, we conceptualize that GenIR affects IR information needs, tasks, and queries. We suggest that the generative systems' flexibility enables more advanced tasks than traditional IR. We argue that even though the flexibility of GenIR systems introduces new capabilities to the search process, the core structure of traditional IR and Interactive Information Retrieval (IIR) remains. We introduce the parallel to conversational information seeking and suggest that we leverage prior research and apply it to the context of GenIR.

In Sect. 4.3, we reemphasize the importance of the user's centrality for GenAI. We highlight the potential of GenIR with more sophisticated user adaptation techniques, enabling dynamic personalization. We emphasize that even though GenIR is driven by advanced algorithms, humans-in-the-loop is indispensable for curating and refining the system's output to ensure relevance and reliability.

In Sect. 4.4, we extend the importance of tasks within IR. We then discuss how tasks are expanding in Artificial Intelligence (AI) while pulling this through to GenIR. Next, we distinguish between using GenIR to enhance system tasks versus user tasks. We then map commonly used information-seeking process stages from IR to GenIR. We conceptualize task complexity for GenIR systems and provide tasks that are suitable and less suitable for GenIR.

Next, Sect. 4.5 discusses how different scenarios and applications can use generative technology, including work, knowledge base access via customized conversational agents, learning and teaching, research, and personal information management.

Given the human-centered nature of this chapter, we discuss user evaluation in relation to GenIR in Sect. 4.6. We briefly overview commonly used user-based evaluations in IR, such as user studies, online evaluation, and implicit measures. We then propose challenges and considerations for evaluating GenIR systems, including its ethical considerations and user privacy.

Lastly, we conclude with an overview of the chapter in Sect. 4.7 and discuss the future proactivity of generative systems.

4.1.2 Chapter Approach and Definitions

Our approach is the following. We study past work on IR and IIR and suggest how future information access systems can leverage prior research and what may be different when GenIR is further developed.

GenIR and GenIIR represent emerging concepts within IR, which traditionally focuses on retrieving relevant information from a large corpus of documents based on a user's query. The new approaches incorporate generative models, especially those based on deep learning, to enhance the search process. We begin by defining the key concepts used in this chapter.

Generative Information Retrieval (GenIR). GenIR is a subset of IR technologies that leverage GenAI to enhance the search process. Unlike traditional IR systems, which focus primarily on matching keywords and returning pre-existing documents, GenIR systems can synthesize, critique, or create new content for user queries. GenIR systems aim to move beyond the limitations of keyword-based searches and static document retrieval, offering users more nuanced, conversational, and interactive search experiences. This approach opens new possibilities for automated content creation, question-answering systems, and personalized information delivery.

While traditional IR systems focus on efficiently finding and presenting existing information, GenIR systems extend this by, for example, creating or synthesizing new information in response to user queries. This fundamental difference in output (i.e., retrieving existing documents versus generating new content) represents a shift in how these systems address user information needs.

Generative Interactive Information Retrieval (GenIIR). GenIIR extends the GenIR concept by emphasizing the interactive nature of—and user centrality in—the search process. Similar to IIR, GenIIR keeps the user central. *Interactive* implies the involvement of humans, in contrast to GenIR, which is system oriented. GenIIR is a system that incorporates the capabilities of GenIR within an interactive framework that prioritizes user engagement and feedback throughout the search process. Unlike GenIR, which primarily focuses on the system's ability to generate and retrieve information, GenIIR emphasizes a collaborative search process where the user's inputs, queries, and feedback directly influence the generation and refinement of information. This approach leverages generative models to synthesize, adapt, and present information in response to user–system interactions (also seen as conversations).

The key to GenIIR is its dynamic, user-driven approach to IR, where the system understands and generates content based on initial queries and evolves its responses through continuous interaction. These interactions ensure the generated information aligns with the user's changing information needs and contexts. GenIIR fundamentally transforms the nature of retrieval by dynamically generating information tailored to the user's evolving needs during the interactions.

Generative Information Access (GenIA). GenIA represents a holistic approach to how users discover, interact with, and utilize information across multiple platforms and formats. The emphasis of GenIA is on the breadth of access and the innovative generation of information, rather than on the depth of the user–system interaction as is the case for GenIIR. Interaction is one component of GenIA, but not its defining feature. It leverages GenAI to retrieve existing information and create, synthesize, or enhance the content in real time. This includes transforming raw data into understandable narratives, generating visualizations from complex datasets, or creating new textual content that fills the gaps in existing information.

The relationships between GenIA, GenIR, and GenIIR are illustrated in Fig. 4.1, and for convenient reference, the definitions are summarized in Table 4.1.

4.1.3 Information Needs, Tasks, and Queries

Information needs, tasks, and queries are foundational to understanding how users interact with IR systems. Thus, we define these concepts. IR and GenIR information needs, tasks, and queries share fundamental similarities, as they all revolve around the user's need to find information. We will explore the differences between traditional IR and GenIR systems in the subsequent section, reflecting the evolving capabilities of GenAI technologies.

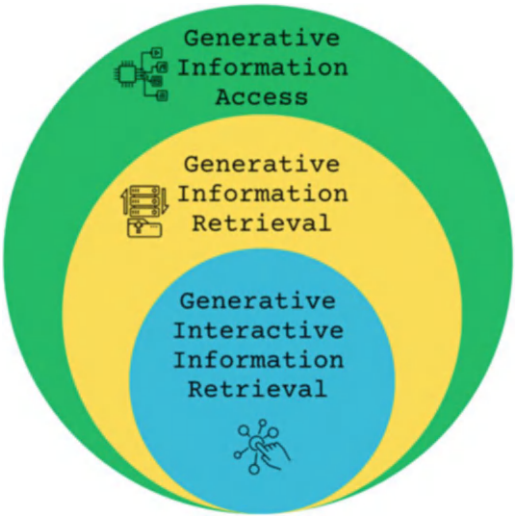

Fig. 4.1 Relationships of GenIA, GenIR, and GenIIR

Table 4.1 Relationship and short description of GenIA, GenIR, and GenIIR

Technology	Description
Generative Information Access (GenIA)	Broadest concept, encompassing generative AI techniques for information access
Generative Information Retrieval (GenIR)	Focuses on retrieving and generating content, enhancing search beyond keyword matching
Generative Interactive Information Retrieval (GenIIR)	Adds human interactivity (i.e., conversational elements) to GenIR, allowing dynamic user engagement

- **Information need:** The genesis of the IR process. An information need arises when users recognize a gap in their knowledge or require information on a particular topic or question [13]. It is the intent or requirement for information the user seeks to fulfill [74]. Information needs are often complex and may not be fully formed or explicitly understood by the user initially [21].
- **Tasks:** The notion of *task* has been widely studied in both IR and Information Science (IS) fields, with two broad perspectives. First, task may refer to a "goal," incorporating a specific scenario providing context for the need [68, 82]. This context elucidates the breadth and depth of the user's information requirements and can influence the search approach. Tasks, such as planning a trip, making a meal, or fixing a car, directly influence the search execution and the type of information deemed necessary, be it detailed explanations, quick facts, or comprehensive overviews [42]. The concept of *search task* focuses more specifically on actions and activities carried out by a user to resolve their information need, such as when interacting with an IR system. For example, Broder's taxonomy specifies search tasks as informational, navigational, or transactional [15].

– **Queries:** Queries are explicit expressions of information needs, formulated by users to interact with an IR system. A query translates the user's information need into a system-processable format, such as a set of keywords or a question. This translation is influenced by the user's understanding of their information need and their perception of the system's capability to meet that need. Crafting an effective query requires users to distil their information needs and task context into a concise and precise information request.[1]

The search for information, therefore, follows a logical flow, beginning with an underlying *information need*—a gap in our knowledge we aim to fill in a specific scenario. This need propels us to define a *task*—how we will acquire that information. Finally, we *translate this task into a query*—employing specific words or phrases to search for information in a system like a search engine. The success of this search hinges on how accurately the query represents our initial information requirement. For instance, the need to prepare an evening meal leads to a requirement of information for a recipe; the task involves using an appropriate Web search engine to seek a suitable recipe, and translating translates into the task of finding a nutritious meal and may manifest through queries such as "healthy dinner recipes" or "easy recipes with vegetables."

4.2 Does Generative Information Retrieval (GenIR) Change Information Retrieval (IR) Information Needs, Tasks, and Queries?

Searching for information in GenIR can be more dynamic and interactive than traditional IR. The process still begins with an *information need*, but GenIR allows for more interactions, such as conversational engagement, to better refine and understand the user's question (also referred to as *prompt*). Additionally, GenIR can generate new, synthesized information relevant to the user's scenario rather than simply returning existing documents.

Tasks in GenIR extend beyond traditional search and retrieval, incorporating direct question-answering, content summarization, and content creation based on the user's needs. This aspect of GenIR can adapt and respond to the nuances of the user's requirements in real time. Users might still need to craft keyword-based queries carefully, but they can also express their needs naturally. This ability to interpret and respond to conversational input improves the feedback loop between user input and information output, making responses more immediate and relevant to the user's context.

[1] In this text, we use the term "queries" broadly to encompass any system-oriented specification of an information need. It therefore includes things such as "keyword queries," "questions," "prompts," and "Boolean queries."

Table 4.2 Comparison of components in traditional IR vs. GenIR

Component	Traditional IR	GenIR
Information needs	Defined by user's desire to find specific information within existing documents	Interpreted flexibly, generating new content that fulfils the user's need
Tasks	Involves searching, browsing, filtering, and sorting through existing information	Extends to content creation, summarization, and question answering through content generation
Queries	Typically keyword-based queries, relying on precise user articulation	Can be more natural or conversational, with the system interpreting the query's intent

For example, consider a user interested in starting an urban garden and seeking information on sustainable practices. In a traditional IR scenario, the user might input several keyword-based queries such as "urban gardening tips," "sustainable urban gardening," or "how to start an urban garden," with the system providing relevant articles in response.

In contrast, with GenIR, the user could ask, "Can you guide me through starting a sustainable urban garden?" The GenIR system could then generate a step-by-step guide from multiple documents, including multimedia, Web pages, or personal documents. This guide could include selecting the right location, choosing plants based on the local climate, and implementing sustainable water drainage, all synthesized into a coherent, personalized response. This example highlights the transformative potential of GenIR in making the information retrieval process more aligned with natural human inquiry (similar to what is known from conversational information seeking) and potentially more efficient in addressing complex, multifaceted information needs (Table 4.2).

4.2.1 Fulfilling Information Needs with GenIR

While the GenIR search process introduces more dynamic interactions and content generation capabilities to broaden the search process, it retains the core structure of the traditional IR and IIR processes [25]. In essence, both approaches navigate from an information need, through a task, to formulating the information need (i.e., query or prompt), aiming to fulfil the user's search intent.

However, GenIR encapsulates a broader concept by integrating these foundational steps into a more fluid and conversation-like model. Rather than fundamentally altering the process, this development adds new interaction layers, understanding, and response generation to the established framework. GenIR's relationship with conversational information-seeking highlights this progression.

Conversational information seeking focuses on Natural Language Processing (NLP) and understanding to facilitate a dialogue-based interaction between the

user and the system [31, 64, 70, 79, 93]. This approach, for example, enables the system to ask clarifying questions [5], to refine search parameters based on user responses [26, 61], and to present information in a more conversational and accessible format [10, 78]. By building on the principles of conversational information seeking, GenIR should leverage prior research in the field, applying it within a generative context to produce synthesized information that directly addresses the user's needs. This connection to conversational information seeking enables researchers to draw upon existing studies and methodologies, to further develop and refine GenIR systems. The accumulated knowledge in understanding user intent, processing natural language queries, and generating relevant responses forms a solid foundation for advancing GenIR. This continuity ensures that innovations in GenIR are grounded in established IR and IIR research while at the same time expanding the boundaries of what information retrieval systems can achieve.

While GenIR introduces novel capabilities and a broader conceptual scope, its search process remains ingrained in the traditional IR framework, enriched by the advancements in conversational information seeking. This relationship validates the effectiveness of GenIR in meeting contemporary information needs and encourages a seamless integration of new technologies with existing IR research to enhance information seeking.

Figure 4.2 depicts the progression of tasks within information retrieval settings, highlighting the extension of capabilities by GenIR systems. The "Critique and evaluate" layer represents an advanced function where the system generates content and provides feedback, broadening the task's scope from mere creation to critical assessment. The diagram captures the concept of GenIR expanding the frontier of tasks beyond what was traditionally possible with search alone.

4.3 User-Centric Generative AI

People are at the heart of IR, as information-seekers and as "system component" as part of Human-in-the-Loop (HITL). Since information needs are inherently personal and unique to each individual, therefore *adapting* systems to users has been an important goal of much IR research. The adaptation of IR systems as the potential to support users in many ways includes:

- Providing more relevant search results by tailoring the search to account for individual preferences
- Reducing cognitive load by aligning information with the user's abilities and experience
- Providing context-sensitive adaptations based on a user's location, time, and device
- Continuously evolving to match the user's changing preferences

In addition, in trying to keep the user central, we also acknowledge people involved on the system side, through HITL approaches. This concept emphasizes

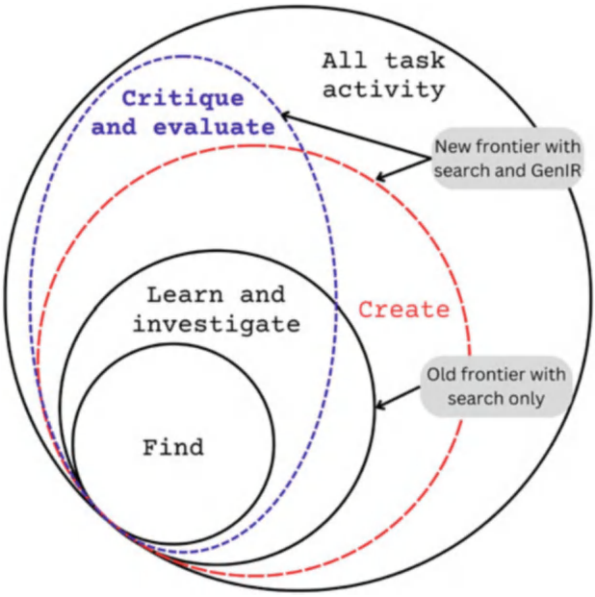

Fig. 4.2 The diagram is an adapted visualization of possible tasks [86], demonstrating different information retrieval and activity levels. The diagram highlights the progression from simple information finding to more complex tasks like learning, investigating, critiquing, evaluating, and creating. The new frontier with GenIR (marked by dashes) indicates that these systems can enable advanced tasks such as *critiquing* and *evaluating*, expanding beyond the traditional search frontier. This conceptualization shows how GenIR systems are pushing the boundaries of what can be achieved with IR, making it possible to engage in higher-order cognitive tasks (e.g., create and critique)

the necessity of integrating human insights within the system development and operation processes, HITL that the systems benefit from continuous human oversight and expertise. This approach enhances the system's adaptability, reliability, and overall effectiveness. By incorporating HITL methodologies, the aim is to create more robust systems integrating human judgment with advanced technological capabilities.

4.3.1 User Adaptation

User adaptation is crucial because there is a difference in effectiveness between a search engine designed for everyone and one personalized for an individual, as highlighted by Teevan et al. as the *potential for personalization* [76]. With the advent of GenIR, existing approaches for user adaptation are enhanced, and as the technology continues to develop, increasingly nuanced approaches are likely to become available. In addition, it is important to note that there is a wide variety

of literature on adaptivity [3, 45, 46, 58]. Therefore, the examples provided should be understood as illustrative rather than exhaustive.

4.3.1.1 User Characteristics and Individual Differences

User factors (personal characteristics and individual differences) can substantially impact how people interact with IR systems. For instance, *cognitive abilities*, such as working memory and processing speed, have been found to have a significant impact on how effectively users search and make use of an IR system [3]. Research has shown that users with more *search experience* tend to make better and more effective use of IR systems. This is because they employ more efficient search strategies based on their understanding of system features [45]. Additionally, users with greater domain knowledge or expertise exhibit different search behaviors, including the sites they visit, query length, and vocabulary breadth. These variations significantly impact overall search success [87].

4.3.1.2 User Adaptation Techniques

The individual difference factors of searchers have direct implications for the design of IR systems and the techniques that can be deployed to make them adaptable to different preferences, needs, and experiences. Adaptability aims to enhance the user experience by enabling the delivery of content that is more relevant, engaging, and accessible, furthering the mission to help the user resolve their information need.

User profiling and personalization involves collecting and analyzing data about individual users (such as their behavior, preferences, and interaction history) to tailor search results specifically to them, typically by re-ranking or filtering. Research in this area has explored various methods for creating dynamic user profiles, including machine learning algorithms that adapt to changes in user behavior over time. Depending on the data a system can collect, user profiles could be short or long term. For example, past browsing behaviors have been used to create user profiles, which are then applied to personalize search results by re-ranking items [58]. Other research has demonstrated that the Big Five personality traits can predict visual search performance [63]. Incorporating such individual differences into user profiles can allow IR systems to provide personalized recommendations and content that aligns with the characteristics of users.

Context-Aware Search considers the user's current context, such as location, device, time, and other situational factors, to provide more relevant search results [37]. For example, a user searching for "restaurants" on a mobile device would likely expect results tailored to their current location.

User Feedback Techniques aims to integrate information directly from the user to improve search results. Widely explored approaches are the use of relevance feedback, where a user may provide *explicit* information (e.g., by marking items in an initial search results list as being relevant or not) [46], or the system makes

use of *implicit* information (e.g., aggregated historical click behavior for the same query) [41] or even simply assumes the top items that were initially retrieved to be useful and uses this information to rewrite the query [20].

GenIR systems using techniques such as language models have the potential to further *personalize* responses, tailor language style, and adjust information complexity based on user profiles. These systems can leverage user profile data to generate content that matches the user's reading level, interests, and conversational preferences. Recent advances in language models enable them to perform tasks with little task-specific data [16], suggesting strong potential for personalization even based on limited user input. Moreover, transfer learning and fine-tuning on user-specific data will allow these models to adapt their output even further to suit individual users better.

4.3.2 The Role of Humans in the Loop for GenIR

In addition to GenIR users for accessing information, we acknowledge that people are integral to developing these systems. While GenIR models will continue to improve, humans still play a crucial role. They provide critical thinking to ensure that the information is useful, accurate, and ethical. HITL AI refers to a methodology where humans are actively involved in some or all stages of an AI system's training, testing, and deployment. This approach combines the efficiency of algorithms with the nuanced understanding and decision-making capabilities of humans. Mosqueira-Rey et al. [60] identifies three broad categories for HITL machine-learning approaches. These categories are differentiated by the degree of control machines have over the learning process. From highest to lowest, machine control degrees are Active Learning, Interactive Machine Learning, and Machine Teaching. In addition, HITL is now also used more broadly across various AI applications.

For GenIR, human–AI collaboration is crucial for curating and validating information. Generative models may be good at finding information but often struggle with understanding its nuance and accuracy. Humans act as fact-checkers, evaluating information for relevance, credibility, and potential bias. Generative models might misinterpret the true intent behind a search query. Humans refine searches by providing context or reformulating queries to meet users' needs better. Human domain expertise is invaluable for interpreting and evaluating information in specific fields. Expert knowledge helps distinguish relevant and irrelevant results, especially in complex or high-risk domains like healthcare or legal information retrieval.

Human input will continue to be key to address *ethical concerns*. Generative models can potentially surface harmful or offensive content. Therefore, humans need to be responsible for setting ethical guidelines and ensuring retrieved information is appropriate and unbiased. This aligns with the growing focus on Fairness, Accountability, Transparency, and Ethics (FATE) in AI systems, where

human oversight is crucial for mitigating bias and ensuring ethical outcomes. In addition, human feedback on the retrieved information is also crucial for improving generative models. This feedback can be used to train models to better understand user needs and return more accurate and relevant results in the future. This aspect aligns with the core principles of HITL Machine Learning (ML), where human feedback forms a continuous loop for improving the machine learning system [60].

4.4 Tasks and Information-Seeking Processes

4.4.1 Tasks in Information Retrieval

The concept of tasks (also referred to as work tasks, information-seeking tasks, search tasks, or IR tasks [71]) is central to the design and effectiveness of IR systems [11, 68, 88]. Tasks represent the goals or objectives that users aim to achieve, ranging from simple queries to complex information-seeking behaviors. Identifying and understanding these tasks are crucial for developing IR systems that align with user intentions and contextual needs. These IR systems leverage computational models to provide responses relevant to users' tasks. By tailoring the retrieval process to the characteristics of individual tasks, IR systems can provide more relevant, accurate, and useful results, thereby enhancing user satisfaction and improving the overall effectiveness of the search process. This task-centric approach to IR highlights the need for systems to understand beyond the content they index and the context and purpose behind user queries [54]. This enables a more nuanced and effective retrieval experience that aligns with the specific demands of different tasks.

Tasks are essential to users' search strategies, the type of information they seek, and how they engage with retrieval systems [55]. For instance, a well-defined task, such as looking up a specific fact (i.e., factoid information need), typically leads to direct and focused search behavior, with users employing precise queries and expecting quick, accurate answers. Conversely, more complex tasks (i.e., non-factoid information needs), such as conducting research for an academic paper, involve iterative search processes, refinement of information needs, and extensive interaction with the IR system to explore, contrast, and evaluate diverse information sources.

The influence of tasks extends to the design and functionality of IR systems themselves. Systems need to adapt to accommodate the varying requirements of different tasks, offering functionalities like query suggestion, personalized filtering, and context-aware retrieval to enhance user satisfaction and search efficiency [71, 91]. Understanding the task-driven nature of search behaviors and information needs is already essential for current systems; as discussed in the following sections, it may become even more crucial for developing generative IR systems that dynamically adapt to user contexts, anticipate information needs, and provide tailored responses.

4.4.2 Expanding IR Tasks into GenIR

Generative AI tasks include a range of "generative" tasks with varying levels of human-AI interaction, from fully automated content creation to collaborative co-creation where humans and AI work together to produce novel outcomes [24, 27]. These models are not only capable of generating text [53] but can be used for image creation [92], music composition [73], data augmentation [14], simulation [4], classification [69], or predictions [23]. As generative AI continues to evolve, the interaction between humans and AI systems is becoming more nuanced and sophisticated [7]. With AI evolving, applications and tasks are expanding too, from enhancing artistic output to scientific research. It has been suggested that the key in successful human–AI interaction lies in finding the right balance between leveraging the AI's capabilities and maintaining human oversight [24]. Ultimately, the goal is to harness the strengths of both humans and AI for improved outcomes.

Generative AI is also impacting the field of IR. By leveraging natural language processing and machine learning techniques, generative AI models can understand and interpret complex queries, providing more accurate and relevant search results [9]. In addition, these models can generate summaries, answer questions, and create content tailored to users' needs. These AI abilities enhance IR efficiency and open new possibilities for personalized and interactive search experiences.

4.4.3 Using GenIR to Enhance System Tasks

Tasks can be categorized as *system tasks*, given their execution by the IR system autonomously rather than by the end user. These tasks are integral to the system's enhanced capability to comprehend, process, and retrieve information relevant to the user's query. For example, a GenIR system can incorporate an automated *query expansion* technique. The automated query expansion process within a GenIR system can autonomously enhance a user's query to improve the relevance of results. This can be achieved with classic synonym additions or semantic enrichments but now completed by the underlying generative system. The outcome is search results that match the exact terms of the original query and include information linked to synonymous terms and related concepts.

An example of a system task query expansion is seen in Fig. 4.3.

These tasks encapsulate many functionalities within Natural Language Processing (NLP), Machine Learning (ML), and AI domains, aimed at improving the system's performance and effectiveness. By incorporating these advanced computational techniques, the IR system can better deal with the complexities of language and patterns within vast datasets. Consequently, these system-centric tasks are key in refining the system's responsiveness and reliability in delivering relevant search results, thereby contributing to the advancement of IR technologies.

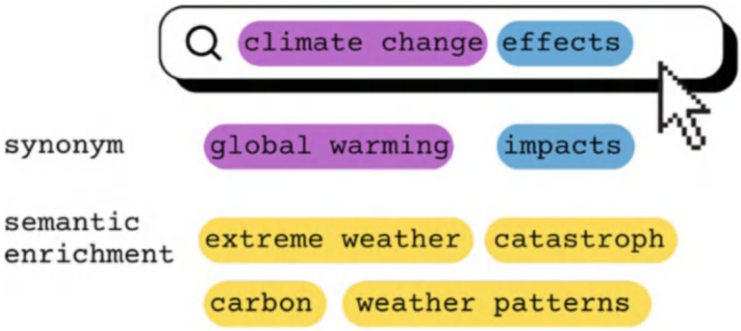

Fig. 4.3 Example of a typical GenIR system enhancement task for query expansion

4.4.4 Using GenIR to Support User Tasks: Mapping GenIR to Existing Information-Seeking Processes

We discuss possible GenIR actions and behaviors for three commonly used stages of the information-seeking process: *query formulation*, *search result exploration*, and *query reformulation* [67, 79], search stages equivalent to *express*, *examine*, and *reformulate* [57]. The information-seeking process model provides broad stages for possible actions and behaviors while providing a structure.

4.4.4.1 Information Need (Query/Prompt) Formulation

The initial stage of *query formulation*, or what can also be referred to as *information need (query/prompt) formulation*, is critical in the information-seeking process. It involves the user identifying and articulating their need for information into a query or prompt that the GenIR system can understand. This stage is critical because the entire search process's effectiveness is based on the user's ability to accurately express their information needs and the system's ability to interpret them correctly.

In traditional IR systems, query formulation often relies on the user's ability to distil their information need into a set of keywords or phrases. However, in GenIR, this process takes on a more dynamic and interactive character. GenIR systems, with their conversational capabilities, allow users to formulate their queries more naturally. This can include posing questions, making statements, or engaging in a dialogue with the system to refine the query.

Ideally, GenIR systems enhance the query formulation stage through interactive information needs refinement and prompt generation. The systems should engage with users to clarify ambiguities, request more context, or suggest different ways to phrase information needs, ensuring a deeper understanding of the user's intent. Additionally, for users unsure of how to express their information needs, GenIR can generate guiding information needs or prompts, aiding them in refining their

search objectives. This approach fosters a more intuitive and user-centric search experience.

In a recent interaction log study, it was indicated that generative AI prompts are often verbose and structured, encapsulating a broader range of information needs and *imperative* (e.g., directive) tasks distinct from traditional search queries [81]. The study showed that LLMs can support users in tasks beyond the three main types based on user intent: informational, navigational, and transactional [15]. Prompts also included instances where entire documents are copied and pasted into the "prompt box." These "document" prompts were often used to extract or summarize a user's personal data, indicative of pseudo-navigational tasks or personal information management. They identified unique tasks, such as text formatting and information extraction, that extend beyond traditional search queries and uncover a range of user intents, predominantly commands to the system.

An example session from this log analysis is shown in Table 4.3. The session starts with a generic question from the user wanting the system to explain the stages of a waterfall model in bullet points (Turn 1). Throughout the session, the user changes the way they are formulating their information need. For example, in Turn 2, the user specifies their initial need in more details.

4.4.4.2 Examine Generated Information

In the context of GenIR-enhanced information-seeking processes, the search result exploration phase expands to include the examination of traditional search results and newly generated information. This phase involves evaluating the relevance and usefulness of the initial search results and assessing the quality, novelty, and relevance of information generated by the AI system.

In the search process' "examine generated information" stage, personalization is crucial in tailoring the generated information to the individual's preferences and needs. This stage involves the AI system leveraging user profiles, search history, and contextual information to create personalized summaries, answers, or content directly relevant to the user's query. By doing so, the system must ensure that the generated information is grounded in accurate information and aligned with the user's interests and requirements, thereby enhancing the overall effectiveness of the information-seeking process. A recent paper investigated the *readability* of generative information systems' output and their accessibility barriers, especially for people with literacy difficulties [66]. This paper showed that responses from widespread large language models may not be accessible to people with cognitive and literacy challenges. The authors stress that generative systems have potential accessibility issues for users with low literacy or reading impairments. To address this, it is imperative for GenIA systems to incorporate mechanisms that can adapt the complexity of language and presentation of information according to users. This further highlights the important role of personalization for generative systems. Furthermore, the design of such systems should be informed by inclusive user

Table 4.3 Interaction session with user input to Google Bard [81]. The user input is verbatim. For example, turns 5, 8, 9, and 10 include bullet points directly copy-pasted by the user from one user looking for information about project management

Turn no.	User input
1	Explain the stages of the waterfall method in bullet points
2	Ok again but explain the distinct stages in the waterfall methodology
3	Give me a reference for this
4	Ok can you give me a Web page reference
5	•Evaluate the advantages and disadvantages of using Waterfall for the project, considering the specific project context
6	Ok project considerations for Agile and the PRINCE2
7	No like you did for Waterfall, project considerations for Agile and then PRINCE2
8	•Explain the significance of gathering user requirements in project management •Describe how user requirements were gathered in each of the evaluated methodologies: Agile, PRINCE2, and WAterfall
9	•Define the various job roles and responsibilities within project management •Compare the job roles and responsibilities in the three methodologies •Highlight any differences or similarities among them
10	•Describe the role of quality assurance, testing, and deployment in IT projects •Explain how each methodology handled quality assurance, testing, and deployment •Evaluate the effectiveness of these processes within the context of each methodology
11	Much more concise
12	Summarize the key findings of your investigation. Discuss the factors that could influence the choice of project management methodology in the new software development company
13	Give me a short conclusion:
14	No in paragraphs
15	Ok again, just a conclusion to this assignment
16	Rewrite in paragraphs
17	I need a conclusion for my assignment. Can you give me one?
18–19	[*Complete copy-paste document inputs*]
20	You missed out Agile
21	Please put Agile first
22	Ok rewrite that in paragraphs

studies, ensuring that language and information are accessible to everyone, thereby mitigating the risk of exacerbating existing information access inequalities.

When we investigate the example interaction session again in Table 4.3, we see that the user is trying to specify *how* the information is presented. In turn 11, the user specifies that they want the information presented in a much more concisely.

4.4.4.3 Information Need (Query/Prompt) Reformulation

As for all interactive search systems, the reformulation stage is critical in GenIR systems. This reformulation acts as the dynamic interface where users refine their queries in response to generated content and initial search outcomes. This iterative process is integral to GenIR, enabling users to adjust their information requests based on the presented information. By continuously refining their queries, users can further investigate their topic, leveraging the generative capabilities of the system to explore complex ideas and uncover connections. This feedback loop enhances the precision of search results and enriches the user's engagement with the information, demonstrating the unique interactivity and adaptability of GenIR systems.

Re-investigating the example interaction in Table 4.3, we can see many different reformulations. The search interaction excerpts highlight the iterative nature of the reformulation stage in the context of the user's search process. Each step, from initial, often imprecise requests ("Give me a short conclusion" in Turn 13) to more specific demands ("No in paragraphs" in Turn 14), illustrates how user queries evolve as they refine their need. This dynamic is crucial in both traditional and GenIR systems, where the capacity to adapt responses based on user feedback can enhance the relevance and utility of the information provided. For instance, requests like "You missed out Agile" and "Please put Agile first" (Turns 20–21) emphasize the importance of adaptability and specificity in search queries, including the need for systems that can flexibly accommodate changing user priorities and insights. In GenIR, this is particularly important, as the system must not only search but also generate content, demonstrating a sophisticated generative model for content creation. These reformulation interactions are practical examples of continuous feedback and essential for refining search outputs and accuracy. This capability to iterate and evolve search queries and responses is foundational in delivering a more personalized and effective search experience.

4.4.5 Conceptualizing Task Complexity for GenIR Systems

The field of IR has long recognized the diverse nature of information-seeking tasks and acknowledges that tasks vary in their complexity [18, 43, 88, 91]. Understanding this variation is crucial for developing information systems that effectively support users across a spectrum of needs. This section introduces a conceptual framework for categorizing information-seeking tasks by two critical dimensions: task complexity and *generative involvement*. For simplicity, we refer to task complexity as the number of steps, the intricacy of these steps, and the level of decision-making needed to complete a task. Figure 4.4 illustrates a continuum of task complexity and the level of generative intervention from AI as discussed below:

– **Basic information retrieval** (low task complexity, minimal generative involvement). It involves direct queries with precise answers, like looking up straight-

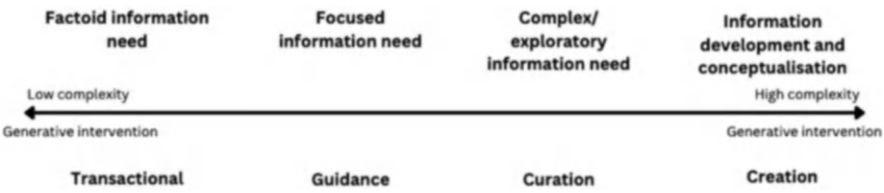

Fig. 4.4 Possible visualization of how generative systems interact depending on task complexity

forward facts. Interaction with GenIA systems is transactional, the user requests specific information, and the AI retrieves it with little to no additional generative contribution.

- **Guided topic expansion** (medium task complexity, moderate generative involvement). It entails broadening the scope of an inquiry to include related topics or concepts, requiring users to navigate through and select relevant information. The GenIA system aids this process by suggesting related areas and generating ancillary information that users can incorporate into their search.
- **In-depth analysis and synthesis** (high task complexity, substantial generative involvement). It requires comprehensive research and the integration of multiple information sources to construct detailed knowledge or insights. The GenIA system plays a significant role by generating complex outputs like summaries of extensive literature, which the user then critically evaluates and refines for their purposes.
- **Intelligent research design** (very high task complexity, interactive generative involvement). It involves the generation of new research frameworks, theoretical models, or innovative problem-solving approaches. The GenIA system and the user interact together, with the GenIA system proposing novel ideas and designs that the user iteratively refines, leading to sophisticated outcomes that may not have been achievable individually.

4.4.5.1 Tasks Less Suitable for GenIR

Based on the above conceptualization, we can see that not all search tasks are suited for a GenIR approach. The effectiveness of GenIR systems largely depends on the query's nature, the user's information need, and the context in which the information will be used. There are several scenarios where traditional IR systems might be more appropriate or where GenIR systems may need to be carefully designed to meet specific requirements:

- **Factoid information need.** Consider someone who wants to have an answer to a very concrete information need, "first person on South Pole." This query seeks a factual answer about a historical event related to world exploration. A traditional IR system would look for information from historical records,

exploration archives, or authoritative history Web sites to provide the name of the explorer who first reached the South Pole. It makes sense for a factoid task to embed the information in current authoritative information. In contrast, even though a generative system may contain the information, the user may not have to conduct extra fact-checking.

- **Legal or medical information.** In domains where the accuracy of information can have serious implications, such as legal and medical research, the conservative approach of traditional IR systems may be preferred. The potential for GenIR systems to synthesize information in ways that misinterpret complex legal statutes or medical guidelines necessitates a cautious application [48].
- **(Re)finding an original online document.** Imagine someone is working on a paper about tasks in information retrieval. They want to retrieve the topics of a previous Text Retrieval Conference (TREC) Track. Even though the user can ask for the topics in a GenIR system, users may prefer to access original documents directly from the sources, rather than receiving synthesized or generated content. In such cases, traditional IR systems that provide direct links to original sources would be more appropriate. In addition, many documents are not online and reside in physical archives or within proprietary databases. The researcher may need to consult these offline materials for academic rigor, necessitating a hybrid approach combining digital searches with traditional library methods.
- **Niche topics.** GenIR systems are typically trained on broad datasets, which may not cover highly specialized or niche topics sufficiently. For niche queries, traditional IR systems that index specialized databases or pay-walled articles might provide more comprehensive and relevant results.
- **Complex topics with critical and high-level reasoning**. While advances in AI and natural language processing have enabled GenIR systems to handle complex queries, there are still limitations in their ability to perform multi-step reasoning or to understand queries that require deep domain-specific knowledge. Complemented by human expertise, traditional IR systems may be better suited for these scenarios.

4.4.5.2 Tasks Suitable for GenIR

Next, we present example Information Access (IA) tasks that are suitable for GenIR:

- **Content creation.** GenIR systems are proficient at creating new, original content tailored to specific needs. This includes writing articles, generating reports, or producing creative pieces like short stories and poetry. The strength of GenIR in content creation lies in its ability to analyze vast amounts of data, understand context, and generate coherent, relevant text based on the user's input or prompts. GenIR can help streamline the content creation, offering efficiency and creativity while reducing the time and resources traditionally required for these tasks.

- **Content summarization.** With the overwhelming amount of online information, there is a growing need for concise summaries that capture the essence of longer texts. GenIR systems can automatically generate accurate, coherent summaries of articles, research papers, books, or reports, making information more accessible and digestible for readers.
- **Content extraction.** GenIR systems can help with content extraction, where specific information, data points, or insights need to be identified and extracted from large volumes of text or complex datasets. GenIR systems can parse through documents, identify relevant pieces of information based on the criteria set by the user, and generate summaries or reports highlighting the extracted content. This could save people time, and GenIR systems may identify patterns that the user may initially overlook.
- **Personalization.** Leveraging the strengths of GenIR, systems can craft personalized information. By analyzing a user's past interactions, search behaviors, or preferences, GenIR systems could curate content, increasing user satisfaction and engagement. The strategic deployment of GenIR for personalized suggestions enriches the user's experience by ensuring relevancy and is crucial in enhancing loyalty and improving conversion rates. GenIR may offer a more personalized, engaging, and user-centric service.

4.5 Scenarios and Applications

4.5.1 Work

Information access systems' role in working environments and work tasks has been studied for decades [51, 52, 56, 80]. There are a variety of tasks on which information access systems are used, including communication, documentation, planning, problem-solving, and admin and management, among others. In 2019, Trippas et al. [80] asked participants in an online survey about work tasks and how digital assistants can support them.

The survey was conducted from May 17 to July 2, 2018, with 410 respondents. One of the questions in the survey asked participants to describe the features or capabilities they would want to have in a hypothetical new piece of technology. Figure 4.5 shows an aggregation of the relevant responses—46.9% of the participants do not report a particular feature or report they do not need any. Features such as automatic "reminders," "scheduling," or "ubiquitous use" are easily recognizable in current applications such as e-mail clients or personal information managers. Yet other features such as "automatic e-mail," "profiling of other people," "note taking," or keeping oneself "up to date" were less obvious to foresee before the uptake of generative AI solutions such as ChatGPT.

A report by Microsoft [17] discusses the opportunities LLM-based technology can create to assist in work tasks. The report suggests that we are witnessing an

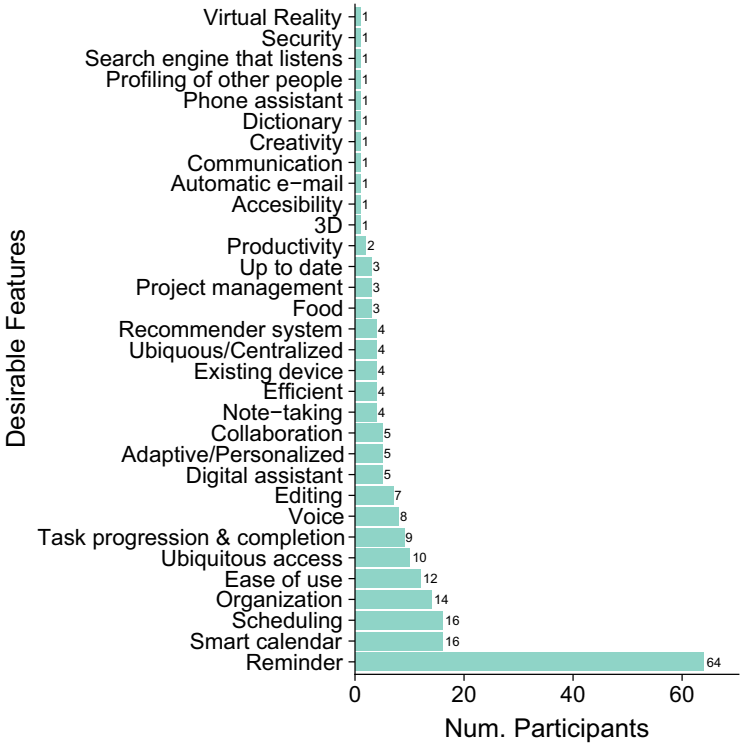

Fig. 4.5 Desirable features for an imaginary new piece of technology as reported by participants from [80]

expansion in the range of tasks automated by human–AI cooperation with assistants or copilots. For instance, the task of *finding* similar snippets of code implementing a particular functionality may shift to *critically analyze* a generated code that satisfies a given functional requirement. Synthesis tasks (e.g., summarizing a set of relevant documents or the discussion in a particular meeting) are also likely to be more automated in working settings, particularly for retrospective and real-time feedback in collaborative scenarios.

As generative IR enable automating more complex information-related tasks in work settings, it is important to consider the ethical implications that automation has in terms of the workforce. As every technological advancement reduces the need for manual labor, it is crucial to have measures that enable everybody to benefit from shared prosperity [44, 89].

4.5.2 Knowledge Base Access via Customized Conversational Agents

Recent developments in GenIR, including Retrieval-Augmented Generation (RAG), have sparked interest in both industry and academia in exploring the role of customized LLM-based conversational agents as a viable solution to provide access to information that is typically stored in private data and knowledge bases (e.g., information used by customer services, internal policies and procedures, project management data).

Pathiyan Cherumanal et al. [62] described *Walert*, an LLM-powered conversational agent customized to answer frequently asked questions about enrolment in computer science programs. Walert aims to demonstrate how common practices in conversational information seeking and IR can be adopted by practitioners when designing and evaluating customized LLM-based conversational agents and identify aspects in the evaluation process that may help practitioners better understand the suitability of GenIR in the context of customized information seeking systems, such as the importance of evaluating RAG at the retrieval phase—and the role of truncated rankings [8]—the inclusion of out of knowledge base (i.e., unanswerable) questions in the test collection, and the need of more informative evaluation measures to assess correctness of the generated responses at the end of the pipeline [62].

Another common problem in organizations and companies is that the vast amount of information generated and collected by different teams within the organization is not easily accessible. Information is typically stored in different data and knowledge bases and organized in different formats. Enterprise search [49] is a structured search system that enables stakeholders in an organization to find relevant information in decision-making processes with organizational or personal goals [85]. Reliable access to relevant information is arguably one of the most essential needs of an organization. And yet the nature of the problem—including data fragmentation, lack of redundancy in information, vocabulary gap, and secure access for different roles within the organization, among others—makes enterprise search one of the most challenging problems in IR [49]. Considering the capability of LLMs in characterizing language and semantic relationships, new GenIR techniques could considerably reduce the vocabulary gap present in enterprise search. In the past, enterprise search has attracted little attention in the academic community [25]. The increasing interest (both in industry and academia) of customized LLM-based conversational agents and RAG may push the state of the art in enterprise search.

4.5.3 Learning and Teaching

The arrival of ChatGPT has provoked a series of debates around the implications of generative AI in the education domain [1, 2, 22]. Recent studies bring empirical results to the discussion on ways students have been using GenAI tools [1]. Code

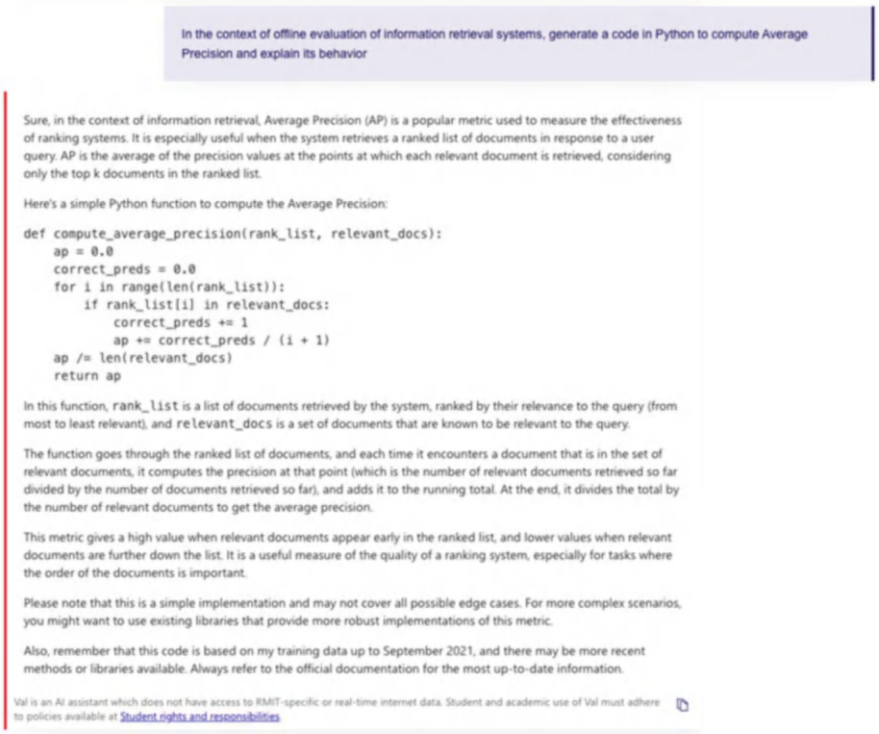

Fig. 4.6 Example of code generation for implementing Average Precision (AP) in Python by RMIT Val using GPT-4

and text generation for completing assignments, essays, and academic projects is one of the most common applications in learning and teaching environments. Although these applications may enhance productivity, they also raise some ethical concerns and implications to pedagogy and academic integrity [22].

Figure 4.6 shows the interaction with Royal Melbourne Institute of Technology (RMIT)'s Val learning assistant,[2] using GPT-4 to generate a snippet of code in Python to compute the Average Precision (AP) effectiveness measure.

Besides the code, the response includes an explanation of the code, as well as a disclaimer to warn the user that the code does not cover edge cases (e.g., topics without relevant documents in the qrels would give a divided-by-zero runtime error). An IR course student may prefer to interact with a conversational agent to learn about fundamental concepts such as AP instead of running a search in a commercial search engine—or reading the evaluation chapter of a classic IR book.

[2] https://www.rmit.edu.au/students/support-services/study-support/val [Accessed: 12 Apr 2024].

A promising direction is using GenAI for personalized and adaptive learning experiences [22, 34], in particular novel ways for providing personalized feedback or instruction tailored to individual student needs [22, 34].

The effective use of this type of technology in learning and teaching environments heavily relies on creating awareness of its limitations and critical engagement. There is also a demand for more established guidelines and policies to ensure the responsible use of GenAI in educational scenarios.

4.5.4 Research

Researchers perform a variety of tasks during the life cycle of a research project. Some common tasks across different study fields are planning, literature review, experimental design, data collection and analysis, writing and publishing, and collaboration. While many of these tasks overlap with the work tasks described in Sect. 4.5.1 (e.g., planning or collaboration), it is important to consider how GenIR can assist practitioners with research tasks.

The literature review phase is where most of the finding tasks occur. Although RAG systems [50]—i.e., systems that generate an answer from a set of passages retrieved from a knowledge base—specifically designed for scientific repositories may provide a complementary way to find relevant work, GenIR can still present unreliable information to researchers. Recent work explored the use of LLMs to make systematic reviews more cost-effective. [83] investigated the effectiveness of Boolean queries designed for systematic review literature search generated by transformed-based systems such as ChatGPT, showing a promising ground for research directions but also important caveats related to incorrect terms in the queries and non-determinism of prompts. Another work in the context of systematic reviews is reducing the number of retrieved documents that need to be manually screened by experts/researchers performing systematic reviews. Recent work in automatic document screening has explored the use of ChatGPT [6], fine-tuning [65], and zero-shot open-sourced LLMs [84]. Results indicate that techniques based on LLMs, particularly fine-tuned, can automatically be developed to screen documents for systematic reviews.

GenAI tools are instead becoming commonly available to assist researchers in refining their writing, e.g., by recommending alternative ways to formulate titles, abstracts, or sentences. Researchers may also benefit from using tools for synthesis or translation tasks [59].

Mittelstadt et al. [59] discuss how other tasks, such as data formatting and conversion, are likely to get more automated with the assistance of GenAI. However, the use of GenAI in other research tasks could compromise research integrity [28, 59], by increasing the risk of lack of reproducibility and transparency, especially if used without robust quality assurance protocols in the data collection and analysis phases.

As in other research fields, IR has also started to include GenAI approaches in research. In addition to automatic relevance assessment [33, 77] and simulation of user's interactions [12, 30] (see Chap. 6 for more details), recent work explore the use of GenAI to characterize tasks and information needs. Zendel et al. [94] explores the effectiveness of instruction-based LLMs to automatically classify the cognitive complexity of information needs described as backstories [94]. Alaofi et al. [4] explores the role of LLMs in generating new query variants for a given information need. Pathiyan Cherumanal et al. [62] use open-source LLMs as a data augmentation approach to generate training phrases to build the conversational model of a customized intent-based conversational agent.

4.5.5 Personalized Personal Information Management

Personal Information Management (PIM) is a set of practices to manage personal information ecosystems [19, 75]. This ecosystem includes various physical and digital information formats like emails, documents, Web content, and social media interactions. PIM enables users to control their information environment, enhancing their productivity, decision-making, and learning [39].

Incorporating GenAI and GenIR into PIM could potentially enhance personalized information access. GenAI can extrapolate new insights, link diverse data sources, and propose novel viewpoints, aiding knowledge integration from personal information sources. In addition, GenIR can improve how information is retrieved and presented to the user. GenIR can provide more relevant and digestible information by understanding the user's context and preferences. This can save time and effort in IR and make using information more efficient and enjoyable.

In this context, PIM is not just about managing information but also about effectively utilizing this information to achieve tasks and fulfill roles within individual contexts. This could include professional roles where specific information is needed to make decisions or personal roles where information could help plan activities or learn new skills, thus enhancing current PIM techniques, making it even more personalized. The combination of PIM, GenAI, and GenIR aims to create a more personalized, efficient, and insightful way of managing and utilizing information. This integrated approach can empower individuals to control their information environment and use information more effectively to achieve their goals. It represents a significant step forward in the evolution of PIM.

4.6 User Evaluation of Generative Information Retrieval

New approaches enabled through GenIR offer extensive opportunities to support users to resolve their information needs. To understand these new approaches and to support ongoing development and improvements, the ability to measure and evaluate system performance is a key requirement. The acir field has a strong history of evaluation. This includes approaches based on offline evaluation using test collections (often called the "Cranfield" methodology) [35] and instantiated through shared evaluation campaigns such as TREC, National Institute of Informatics Testbeds and Community for Information Access Research (NTCIR), Cross Language Evaluation Forum (CLEF), and Forum for Information Retrieval Evaluation (FIRE), online evaluation through techniques such as A/B testing [36], and user-based evaluation [40].

While the established approaches for evaluating IR systems provide a good foundation, they are not always directly usable in the context of new features that GenIR systems support. For example, Gienapp et al. [32] have recently proposed a framework for generative *ad hoc* retrieval—the task of ranking documents by their expected relevance in response to a single search query—that defines *utility*, *reading*, and *accumulation* components for an effectiveness metric. In line with traditional ad hoc retrieval evaluation based on test collections, this framework offers a promising direction to enable GenIR systems to be evaluated for ad hoc search, offering a clearly defined, repeatable, and cost-effective way to quantify effectiveness. However, similar to the use of test collections for traditional IR systems, this comes at the cost of simplification by essentially abstracting out the user and their interactions with the system.

As was highlighted in the previous sections of this chapter, GenIR in particular offers substantial new opportunities at the level of users, tasks, and scenarios: key opportunities arise in the interaction between users and systems, working to resolve an information need that is situated in the scope of a particular task—GenIIR. Evaluation here typically required user-focused approaches, rather than the use of test collections that typically abstract out the variability that users and interactions introduce. We therefore provide an overview of the key methodologies and associated considerations that arise in the context of user-based evaluation of these systems.

4.6.1 Current Information Retrieval Approaches to User Evaluation

4.6.1.1 User Studies

Evaluating the effectiveness of interactive systems can require careful study of the interactions between users and systems. It can be helpful to consider different approaches based on the goals of the research: exploratory, descriptive, and explanatory [40]. User studies can vary widely depending on the phenomenon being studied. Generally, they fall into exploratory, descriptive, and explanatory

categories, reflecting the level of researcher control. Exploratory studies involve minimal intervention, while explanatory studies often require extensive intervention for formal experimental inference.

- *Exploratory studies.* In situations where relatively little is known about the phenomenon, exploratory studies are useful to enable better understanding [72]. The aim is often to learn more about the phenomenon, which means that the research questions may be broad or open ended. As a result, exploratory studies are typically less structured. Exploratory studies often inform subsequent descriptive and explanatory studies.
- *Descriptive studies.* Descriptive studies aim to describe a phenomenon by careful observation and documentation [29]. Such studies can provide benchmarks of interactive systems and serve as taxonomies related to the phenomenon of interest.
- *Explanatory studies*. When variables of interest have been identified, explanatory studies offer a framework to determine relationships between them. This includes formal experiments to establish causality. Explanatory studies are sometimes termed "laboratory experiments," since they often take place in controlled conditions, with the aim of isolating the key variables of interest from possible confounding conditions [90].

4.6.1.2 Online Evaluation and Implicit Measures

Online evaluation aims to measure the effectiveness of IR systems by considering implicit indicators of user behavior as they interact with a live system. Indicators may be any measurable signals that reflect user activity and can range from low-level events such as the number of clicks on a hyperlink and the dwell time on particular Web page to higher-level events such as decisions to purchase items in an online store [36].

To establish the relative effectiveness of two systems, online evaluation typically makes use of A/B testing, a between-subjects experiment where users are randomly exposed to either system A or system B (the independent variable) to establish the presence or absence of an effect on the chosen implicit indicator (the dependent variable) [47].

4.6.2 Challenges and Considerations for Evaluating Generative Information Retrieval Systems

The evolution of IR to include conversational and generative aspects necessitates a deeper understanding of user needs and behaviors, especially since these systems may substantially change user expectations and interaction approaches. GenIR systems require rigorous user evaluation methodologies to ensure their effectiveness and relevance. User studies, incorporating both quantitative and qualitative

methodologies, will be essential in identifying the effectiveness of such systems in addressing complex user information needs.

A key open challenge for GenIR evaluation will be to establish realistic approaches for evaluating system output with users when we cannot control the system's output. For generative systems, a key consideration regarding user studies is the extent to which the system output needs to be controlled. Generative systems, by their nature, create "new" responses, and it may be difficult to ensure that such a system generates identical output even in response to the same input query.

The impact of this factor will vary depending on the type of study being conducted. In the context of *user studies*, this is, e.g., unlikely to be problematic for an open-ended exploratory study aiming to learn about interactions between users and chatbots but may present new complications into the design of an exploratory study in which the system output needs to be a controlled variable.

Since the indicators used in *online evaluation* rely on signals of user behavior, rather than the specific output of a GenIIR system, this evaluation approach can be used directly to evaluate systems that include new generative components. It is however important to bear in mind the usual limitations of online evaluation, namely, that the implicit indicators are very likely to only be a proxy for variables that are actually of interest, such as whether the system actually conveyed useful information to the user or whether the user was ultimately satisfied.

Other challenges in evaluating GenIR systems include accounting for the Natural Language Understanding (NLU) and Natural Language Generation (NLG) components, managing context and state across conversational turns, and ensuring the relevance and coherence of system responses. User evaluation methods will therefore need to be tailored to address these challenges, e.g., by incorporating scenario-based testing, user satisfaction surveys, and task completion rates as part of the evaluation criteria.

Beyond individual studies that focus on particular aspects of evaluation, the development and ongoing evaluation of GenIR systems will benefit from the use of *user-centered design principles*, involving users early and throughout the design process of such systems. This includes understanding user preferences for conversational interactions, personalization, and response generation. Design decisions should be informed by user feedback, ensuring that the system aligns with user expectations and information seeking behaviors.

4.6.2.1 Ethical Considerations and User Privacy

The ethical landscape of user evaluation in GenIR systems is complex, underscored by the importance of the ethical use of data and privacy considerations. In this context, the methodologies employed to gather, analyze, and store user data should be carefully designed to uphold the highest standards of privacy and ethics. In most countries, regulatory requirements around GenIR are developing; but perhaps even more importantly, ethical considerations and practices are a fundamental aspect of building trust and ensuring the integrity of the interaction between users and systems.

Transparency in the collection, usage, and storage of user data forms the cornerstone of ethical user evaluation. Users should be fully informed about what data is being collected, how it is being used, and where it is stored. This transparency is crucial not just for compliance with privacy laws and regulations, such as the General Data Protection Regulation (GDPR) in the European Union, but also for fostering a relationship of trust with users. When users understand how their data contributes to the improvement and effectiveness of GenIR systems, they are more likely to participate willingly in the evaluation process.

Informed consent is another critical element, ensuring that users are not just aware of how their data is used but have explicitly agreed to it. This consent should be obtained through clear, understandable language that avoids technical jargon, so that it is accessible to all users regardless of their background in technology. Moreover, informed consent should not be treated as a one-time process. Users should have the ability to withdraw their consent at any time, necessitating systems that can accommodate such requests without compromising the integrity of the data or the user experience.

A key technique in preserving user privacy is the *de-identification* or *anonymization* of user data. By removing or obfuscating identifiers that can link data back to an individual, researchers and developers can analyze patterns, behaviors, and feedback while minimizing risks around compromising user anonymity in situations such as data breaches or unauthorized access. Privacy-preserving methodologies extend beyond anonymization principles to include techniques such as differential privacy, which adds noise to the data to prevent the identification of individuals while still allowing for the aggregate data to be useful for analysis and system improvement. These methodologies ensure that the evaluation of GenIR systems can proceed without exposing sensitive user information or compromising the privacy of individual users.

4.7 Conclusion and the Future of Generative Systems

This chapter demonstrates that research on highly interactive information retrieval is not new. However, with new developments around generative technology, interactivity has become central again in information retrieval research. This *generative interactive information retrieval* resurgence may have the potential to make IA systems true assistants. However, questions such as *how much generation* is really needed for particular tasks and *what kind of interactivity best enhances user experience* remain open. The challenge lies in finding the right balance between generative capabilities and user control, ensuring that the systems are powerful but also intuitive and user-friendly. As we progress, we must continue exploring these questions and testing and refining generative interactive approaches to realize their potential in transforming IR into a more dynamic and collaborative process.

4.7.1 Proactivity in Generative Information Retrieval

Generative systems, despite their advancements, largely are still not proactive. The concept of "search engines that listen" aimed to introduce a more interactive dimension to IR systems [38, 79]. This vision sought to transform search engines from "responders" to active participants in the search process, capable of understanding and adapting to the user's context in real time.

With GenIR, we have progressed in expanding the task types these systems can handle, moving beyond traditional search queries to include content generation, summarization, critiquing, and even dialogue-like user interactions. However, this leap has yet to realize the proactive potential of GenIR systems fully. The envisioned "search engines that listen" imply initiative and anticipation, actively engaging with users, seeking clarification, and offering suggestions even before a query is fully articulated, perhaps even imply that a system is part of an agent that performs tasks on behalf of users.

To achieve genuinely proactive systems, enhancements in GenIR should focus on better interpreting user interactions, effectively using context, and applying predictive analytics to anticipate user needs. This shift toward proactive participation, making the system a co-navigator rather than just a responder, may improve the user experience, making IR more intuitive and aligning more with natural human information-seeking behaviors.

References

1. Abbas, M., Jam, F.A., Khan, T.I.: Is it harmful or helpful? Examining the causes and consequences of generative AI usage among university students. Int. J. Edu. Technol. Higher Edu. **21**(1), 10 (2024). ISSN: 2365-9440. https://doi.org/10.1186/s41239-024-00444-7. (Visited on 04/08/2024)
2. Adiguzel, T., Kaya, M. H., Cansu, F.K.: Revolutionizing education with AI: exploring the transformative potential of ChatGPT. Contemp. Educ. Technol. **15**(3), ep429 (2023). ISSN: 1309-517X. https://doi.org/10.30935/cedtech/13152. https://www.cedtech.net/article/revolutionizing-education-with-ai-exploring-the-transformative-potential-of-chatgpt-13152 (Visited on 03/05/2024)
3. Al-Maskari, A., Sanderson, M.: The effect of user characteristics on search effectiveness in information retrieval. Inf. Proc. Manag. **47**(5), 719–729 (2011). ISSN: 03064573. https://doi.org/10.1016/j.ipm.2011.03.002. https://linkinghub.elsevier.com/retrieve/pii/S030645731100029X (Visited on 03/15/2024)
4. Alaofi, M., Gallagher, L., Sanderson, M., Scholer, F., Thomas, P.: Can generative LLMs create query variants for test collections? An exploratory study. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '23, pp. 1869–1873. Association for Computing Machinery, New York, NY (2023). ISBN: 978-1-4503-9408-6. https://doi.org/10.1145/3539618.3591960. https://dl.acm.org/doi/10.1145/3539618.3591960 (Visited on 02/26/2024)

5. Aliannejadi, M., Zamani, H., Crestani, F., Croft, W.B.: Asking clarifying questions in open-domain information-seeking conversations. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '19: The 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 475–484. ACM, Paris (2019). ISBN: 978-1-4503-6172-9. https://doi.org/10.1145/3331184.3331265. https://dl.acm.org/doi/10.1145/3331184.3331265 (Visited on 03/06/2024)
6. Alshami, A., Elsayed, M., Ali, E., Eltoukhy, A.E., Zayed, T.: Harnessing the power of ChatGPT for automating systematic review process: methodology, case study, limitations, and future directions. Systems **11**(7), 351 (2023). ISSN: 2079-8954. https://doi.org/10.3390/systems11070351. (Visited on 03/25/2024)
7. Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P.N., Inkpen, K., Teevan, J., Kikin-Gil, R., Horvitz, E.: Guidelines for human-AI interaction. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. CHI '19: CHI Conference on Human Factors in Computing Systems, pp. 1–13. ACM, Glasgow Scotland (2019). ISBN: 978-1-4503-5970-2. https://doi.org/10.1145/3290605.3300233. https://dl.acm.org/doi/10.1145/3290605.3300233 (Visited on 02/27/2024)
8. Amigó, E., Mizzaro, S., Spina, D.: Ranking interruptus: When truncated rankings are better and how to measure that. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 588–598. ACM, Madrid (2022). ISBN: 978-1-4503-8732-3. https://doi.org/10.1145/3477495.3532051. (Visited on 03/28/2024)
9. Anand, A., Anand, A., Setty, V.: Query Understanding in the Age of Large Language Models. 2023. arXiv: 2306.16004. https://arxiv.org/abs/2306.16004 (Visited on 02/27/2024) Published in *Gen-IR@SIGIR 2023: The First Workshop on Generative Information Retrieval*
10. Avula, S., Chadwick, G., Arguello, J., Capra, R.: SearchBots: User engagement with ChatBots during collaborative search. In: Proceedings of the 2018 Conference on Human Information Interaction & Retrieval—CHIIR '18. The 2018 Conference, pp. 52–61. ACM Press, New Brunswick, NJ (2018). ISBN: 978-1-4503-4925-3. https://doi.org/10.1145/3176349.3176380. http://dl.acm.org/citation.cfm?doid=3176349.3176380 (Visited on 03/06/2024)
11. Bai, J., Nie, J.-Y.: Adapting information retrieval to query contexts. Inf. Proc. Manag. **44**(6), 1901–1922 (2008). https://www.sciencedirect.com/science/article/pii/S0306457308000824 (Visited on 02/29/2024)
12. Balog, K., Zhai, C.: User Simulation for Evaluating Information Access Systems (2023). arXiv: 2306.08550 [cs]. (Visited on 03/28/2024)
13. Belkin, N.J., Oddy, R.N., Brooks, H.M.: ASK for information retrieval: Part I. Background and theory. J. Document. **38**(2), 61–71 (1982). https://www.emerald.com/insight/content/doi/10.1108/eb026722/full/html (Visited on 03/05/2024)
14. Biswas, A., Md Abdullah Al, N., Imran, A., Sejuty, A. T., Fairooz, F., Puppala, S., Talukder, S. : Generative adversarial networks for data augmentation. In: Data Driven Approaches on Medical Imaging, Zheng, B., Andrei, S., Sarker, M.K., Gupta, K.D. (eds.), pp. 159–177. Springer Nature Switzerland, Cham (2023). ISBN: 978-3-031-47771-3 978-3-031-47772-0. https://doi.org/10.1007/978-3-031-47772-0_8. https://link.springer.com/10.1007/978-3-031-47772-0_8 (Visited on 02/27/2024)
15. Broder, A.: A taxonomy of web search. ACM SIGIR Forum **36**(2), 3–10 (2002). ISSN: 0163-5840. https://doi.org/10.1145/792550.792552. https://dl.acm.org/doi/10.1145/792550.792552 (Visited on 02/10/2024)
16. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems 33 (NeurIPS 2020) (2020)
17. Buttler, J., Jaffe, S., Baym, N., Czerwinski, M., Iqbal, S., Nowak, K., Rintel, S., Sellen, A., Vorvoreanu, M., Hecht, B., Teevan, J.: Microsoft New Future of Work Report 2023. Tech Report MSRTR-2023-34. Microsoft Research (2023). https://aka.ms/nfw2023

18. Byström, K., Järvelin, K.: Task complexity affects information seeking and use. Inf. Process. Manag. **31**(2), 191–213 v
19. Capra, R.: A survey of personal information management practices. In: Proceedings of Personal Information Management: PIM (2009)
20. Carpineto, C., Romano, G.: A survey of automatic query expansion in information retrieval. ACM Comput. Surveys **44**(1), 1–50 (2012). ISSN: 0360-0300, 1557-7341. https://doi.org/10.1145/2071389.2071390. https://dl.acm.org/doi/10.1145/2071389.2071390 (Visited on 03/16/2024)
21. Case, D.O., Given, L.M.: Looking for Information: A Survey of Research on Information Seeking, Needs, and Behavior. Emerald Group Publishing, Bingley (2016). https://books.google.com.au/books?hl=en&lr=&id=IAYvDAAAQBAJ&oi=fnd&pg=PP1&dq=scenarios+information+seeking&ots=Y-OTmzrOC3&sig=cAp_H91RVR5yXT65hvUM5tuUByA (Visited on 03/05/2024)
22. Chan, C.K.Y., Lee, K.K.W.: The AI generation gap: Are Gen Z students more interested in adopting generative AI such as ChatGPT in teaching and learning than their Gen X and millennial generation teachers? Smart Learn. Environ. **10**(1), 60–23 (2023). ISSN: 2196-7091. https://doi.org/10.1186/s40561-023-00269-3
23. Chen, B., Wu, Z., Zhao, R.: From fiction to fact: the growing role of generative AI in business and finance. J. Chinese Econ. Busin. Stud. **21**(4), 471–496 (2023). ISSN: 1476-5284, 1476-5292. https://doi.org/10.1080/14765284.2023.2245279. https://www.tandfonline.com/doi/full/10.1080/14765284.2023.2245279 (Visited on 02/27/2024)
24. Cheng, R., Smith-Renner, A., Zhang, K., Tetreault, J., Jaimes-Larrarte, A.: Mapping the design space of human-ai interaction in text summarization. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 431–455. Association for Computational Linguistics, Seattle (2022). https://doi.org/10.18653/v1/2022.naacl-main.33. https://aclanthology.org/2022.naaclmain.33 (Visited on 02/27/2024).
25. Croft, W.B., Metzler, D., Strohman, T.: Search Engines: Information Retrieval in Practice, vol. 520. Addison-Wesley Reading, Reading (2010). https://www.academia.edu/download/30740463/z2009_2465.pdf (Visited on 03/06/2024)
26. Dalton, J., Fischer, S., Owoicho, P., Radlinski, F., Rossetto, F., Trippas, J.R., Zamani, H.: Conversational information seeking: Theory and application. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '22, pp. 3455–3458. ACM, Madrid (2022). ISBN: 978-1-4503-8732-3. https://doi.org/10.1145/3477495.3532678. https://dl.acm.org/doi/10.1145/3477495.3532678 (Visited on 03/06/2024)
27. Ding, Z., Chan, J.: Mapping the Design Space of Interactions in Human-AI Text Co-creation Tasks (2023). https://doi.org/10.48550/arXiv.2303.06430 [cs]. http://arxiv.org/abs/2303.06430 (Visited on 02/27/2024) Published at NAAC'22: https://aclanthology.org/2022.naacl-main.33/
28. Duckham, M., Scholer, F., Barr, D., Blades, D., (Ben) Cheng, C.-T., Falzon, B., Forsyth, A., Given, L., McKay, D., Mention, A.-L., Mulet, X., Plebanski, M., Potts, J., Sanderson, M., Thangarajah, J., Thomas, J., Verspoor, K., Xue, C., Yarovsky, I., Yu, X., Zambetta, F.: Research Integrity and Generative AI. [object Object] (2023). https://doi.org/10.5281/ZENODO.10081201. https://zenodo.org/doi/10.5281/zenodo.10081201 (Visited on 03/05/2024)
29. Eickhoff, C., Teevan, J., White, R., Dumais, S.: Lessons from the journey: A query log analysis of within-session learning. In: Proceedings of the 7th ACM International Conference on Web Search and Data Mining, pp. 223–232. ACM, New York (2014). ISBN: 978-1-4503-2351-2. https://doi.org/10.1145/2556195.2556217. (Visited on 04/07/2024)

30. Engelmann, B., Breuer, T., Friese, J.I., Schaer, P., Fuhr, N.: Context-driven interactive query simulations based on generative large language models. In: Advances in Information Retrieval, Goharian, N., Tonellotto, N., He, Y., Lipani, A., McDonald, G., Macdonald, C., Ounis, I. (eds.), vol. 14609, pp. 173–188. Springer Nature Switzerland, Cham (2024). ISBN: 978-3-031-56059-0 978-3-031-56060-6. https://doi.org/10.1007/978-3-031-56060-6_12. (Visited on 03/28/2024)
31. Gao, J., Galley, M., Li, L.: Neural approaches to conversational AI. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. SIGIR '18, pp. 1371–1374. ACM, Ann Arbor, MI (2018). ISBN: 978-1-4503-5657-2. https://doi.org/10.1145/3209978.3210183. https://dl.acm.org/doi/10.1145/3209978.3210183 (Visited on 03/06/2024)
32. Gienapp, L., Scells, H., Deckers, N., Bevendorff, J., Wang, S., Kiesel, J., Syed, S., Fröbe, M., Zuccon, G., Stein, B., Hagen, M., Potthast, M.: Evaluating Generative Ad Hoc Information Retrieval. Nov. 2023. arXiv: 2311.04694 [cs]. (Visited on 03/28/2024) Published at SIGIR'24: https://dl.acm.org/doi/10.1145/3626772.3657849
33. Gilardi, F., Alizadeh, M., Kubli, M.: ChatGPT outperforms crowd workers for text-annotation tasks. Proc. Natl. Acad. Sci. **120**(30), e2305016120 (2023). ISSN: 0027-8424, 1091-6490. https://doi.org/10.1073/pnas.2305016120. (Visited on 03/28/2024)
34. Hai-Jew, S.: Generative AI in Teaching and Learning, 1st edn. IGI Global, Hershey (2023). ISBN: 9798369300749. https://doi.org/10.4018/979-8-3693-0074-9
35. Harman, D.K.: TREC: Experiment and evaluation in information retrieval, Voorhees, E.M., D.K. Harman (eds.). MIT Press, Cambridge (2005). https://mitpress.mit.edu/9780262220736/trec/
36. Hofmann, K., Li, L., Radlinski, F.: Online evaluation for information retrieval. Foundat. Trends®Inf. Retri. **10**(1), 1–117 (2016). ISSN: 1554-0669, 1554-0677. https://doi.org/10.1561/1500000051. (Visited on 04/07/2024)
37. Hosokawa, Y., Nakazawa, Y., Yamamoto, T.: Location-aware information retrieval for identifying local and distant landmark. In: Proceedings of the 29th Annual ACM Symposium on Applied Computing. SAC 2014: Symposium on Applied Computing, pp. 428–435. ACM, Gyeongju (2014). ISBN: 978-1-4503-2469-4. https://doi.org/10.1145/2554850.2554974. https://dl.acm.org/doi/10.1145/2554850.2554974 (Visited on 03/16/2024)
38. Joho, H.: Taciturn search, loquacious search. IEICE Tech. Rep. **115**(184), 19–19 (2015). https://ken.ieice.org/ken/paper/20150821Tb1s/eng/ (Visited on 03/06/2024)
39. Jones, W., Capra, R., Czerwinski, M., Dinneen, J.D., Gwizdka, J., Karadkar, U.: PIM 2024: The information we need, when we need it…: As we get ever closer, is this ideal still ideal? In: Proceedings of the 2024 Conference on Human Information Interaction and Retrieval. CHIIR '24, pp. 438–440. Association for Computing Machinery, Sheffield (2024). ISBN: 9798400704345. https://doi.org/10.1145/3627508.3638333
40. Kelly, D.: Methods for evaluating interactive information retrieval systems with users. Foundat. Trends®Inf. Retr. **3**(1–2), 1–224 (2009). https://www.nowpublishers.com/article/Details/INR-012 (Visited on 02/10/2024)
41. Kelly, D., Teevan, J.: Implicit feedback for inferring user preference: A bibliography. In: ACM SIGIR Forum **37**(2), 18–28 (2003). ISSN: 0163-5840. https://doi.org/10.1145/959258.959260. https://dl.acm.org/doi/10.1145/959258.959260 (Visited on 03/16/2024)
42. Kelly, D., Arguello, J., Edwards, A., Wu, W.C.: Development and evaluation of search tasks for IIR experiments using a cognitive complexity framework. In: Proceedings of the 2015 International Conference on The Theory of Information Retrieval, pp. 101–110. ACM, Northampton, MA (2015). ISBN: 978-1-4503-3833-2. https://doi.org/10.1145/2808194.2809465. https://dl.acm.org/doi/10.1145/2808194.2809465 (Visited on 03/28/2024)
43. Kelly, D., Arguello, J., Edwards, A., Wu, W.C.: Development and evaluation of search tasks for IIR experiments using a cognitive complexity framework. In: Proceedings of the 2015 International Conference on the Theory of Information Retrieval (2015), pp. 101–110. https://doi.org/10.1145/2808194.280946
44. Khogali, H.O., Mekid, S.: The blended future of automation and AI: Examining some long-term societal and ethical impact features. Technol. Soc. **73**, 102232 (2023).

ISSN: 0160791X. https://doi.org/10.1016/j.techsoc.2023.102232. https://linkinghub.elsevier.com/retrieve/pii/S0160791X23000374 (Visited on 03/06/2024)
45. Kim, K.-S., Allen, B.: Cognitive and task influences on web searching behavior. J. Am. Soc. Inf. Sci. Technol. **53**(2), 109–119 (2002). ISSN: 1532-2882, 1532-2890. https://doi.org/10.1002/asi.10014. https://onlinelibrary.wiley.com/doi/10.1002/asi.10014 (Visited on 03/15/2024)
46. Kim, J.Y., Teevan, J., Craswell, N.: Explicit in situ user feedback for web search results. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '16, pp. 829–832. ACM, Pisa (2016). ISBN: 978-1-4503-4069-4. https://doi.org/10.1145/2911451.2914754. https://dl.acm.org/doi/10.1145/2911451.2914754 (Visited on 03/16/2024)
47. Kohavi, R., Longbotham, R., Sommerfield, D., Randal, M. Henne: controlled experiments on the web: survey and practical guide. Data Mining Knowl. Discovery **18**(1), 140–181 (2009). ISSN: 1384-5810, 1573-756X. https://doi.org/10.1007/s10618-008-0114-1. (Visited on 04/07/2024)
48. Koopman, B., Zuccon, G.: Dr Chatgpt tell me what i want to hear: How different prompts impact health answer correctness. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 15012–15022. Association for Computational Linguistics, Singapore (2023). https://doi.org/10.18653/v1/2023.emnlp-main.928. https://aclanthology.org/2023.emnlp-main.928 (Visited on 03/04/2024)
49. Kruschwitz, U., Hull, C.: Searching the enterprise. Foundat. Trends®Inf. Retr. **11**(1), 1–142 (2017). ISSN: 1554-0669, 1554-0677. https://doi.org/10.1561/1500000053. (Visited on 03/28/2024)
50. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Advances in Neural Information Processing Systems, Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.), vol. 33, pp. 9459–9474. Curran Associates, Red Hook (2020)
51. Li, Y., Belkin, N.J.: A faceted approach to conceptualizing tasks in information seeking. Inf. Proc. Manage. **44**(6), 1822–1837 (2008). ISSN: 03064573. https://doi.org/10.1016/j.ipm.2008.07.005. https://linkinghub.elsevier.com/retrieve/pii/S0306457308000836 (Visited on 03/06/2024)
52. Li, Y., Belkin, N.J.: An exploration of the relationships between work task and interactive information search behavior. J. Amer. Soc. Inf. Sci. Technol. **61**(9), 1771–1789 (2010). ISSN: 1532-2882, 1532-2890. https://doi.org/10.1002/asi.21359. https://onlinelibrary.wiley.com/doi/10.1002/asi.21359 (Visited on 03/06/2024)
53. Li, Y., Pan, Q., Wang, S., Yang, T., Cambria, E.: A generative model for category text generation. Inf. Sci. **450**, 301–315 (2018). https://www.sciencedirect.com/science/article/pii/S0020025518302366 (Visited on 02/27/2024)
54. Liu, J.: Deconstructing search tasks in interactive information retrieval: a systematic review of task dimensions and predictors. Inf. Proc. Manag. **58**(3), 102522 (2021). ISSN: 0306-4573. https://doi.org/10.1016/j.ipm.2021.102522. https://www.sciencedirect.com/science/article/pii/S0306457321000315 (Visited on 02/29/2024)
55. Liu, J., Mitsui, M., Belkin, N.J., Shah, C.: Task, information seeking intentions, and user behavior: Toward a multi-level understanding of web search. In: Proceedings of the 2019 Conference on Human Information Interaction and Retrieval. CHIIR '19, pp. 123–132. Association for Computing Machinery, New York, NY (2019). ISBN: 978-1-4503-6025-8. https://doi.org/10.1145/3295750.3298922. https://dl.acm.org/doi/10.1145/3295750.3298922 (Visited on 02/28/2024)
56. Liu, J., Sarkar, S., Shah, C.: Identifying and predicting the states of complex search tasks. In: Proceedings of the 2020 Conference on Human Information Interaction and Retrieval. CHIIR '20, pp. 193–202. ACM, Vancouver, BC (2020). ISBN: 978-1-4503-6892-6. https://doi.org/10.1145/3343413.3377976. https://dl.acm.org/doi/10.1145/3343413.3377976 (Visited on 03/06/2024)

57. Marchionini, G.: Information Seeking in Electronic Environments, vol. 9. Cambridge University Press, Cambridge (1995). https://books.google.com.au/books?hl=en&lr=&id=cYOHgr18DSQC&oi=fnd&pg=PR8&dq=G.+Marchionini.+Information+seeking+in+electronic+environments.&ots=e8oLNGTwZU&sig=7wd1IpxzmQYJgbyudhD3bpnEEJc (Visited on 03/06/2024). https://doi.org/10.1017/CBO9780511626388
58. Matthijs, N., Radlinski, F.: Personalizing web search using long term browsing history. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. WSDM'11, pp. 25–34. ACM, Hong Kong (2011). ISBN: 978-1-4503-0493-1. https://doi.org/10.1145/1935826.1935840. https://dl.acm.org/doi/10.1145/1935826.1935840 (Visited on 03/16/2024)
59. Mittelstadt, B., Wachter, S., Russell, C.: To protect science, we must use LLMs as zero-shot translators. Nat. Human Behav. **7**(11), 1830–1832 (2023). ISSN: 2397-3374. https://doi.org/10.1038/s41562-023-01744-0. https://www.nature.com/articles/s41562-023-01744-0 (Visited on 03/19/2024)
60. Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., Fernández-Leal, Á.: Human-in-the-loop machine learning: a state of the art. Artif. Intell. Rev. **56**(4), 3005–3054 (2023). ISSN: 0269-2821, 1573-7462. https://doi.org/10.1007/s10462-022-10246-w. https://link.springer.com/10.1007/s10462-022-10246-w (Visited on 03/22/2024)
61. Owoicho, P., Dalton, J., Aliannejadi, M., Azzopardi, L., Trippas, J.R., Vakulenko, S.: TREC CAsT 2022: Going beyond user ask and system retrieve with initiative and response generation. In: NIST Special Publication, pp. 500–338 (2022). https://trec.nist.gov/pubs/trec31/papers/Overview_cast.pdf (Visited on 03/06/2024)
62. Pathiyan Cherumanal, S., Tian, L., Abushaqra, F.M., de Paula, A.F.M., Ji, K., Hettiachchi, D., Trippas, J.R., Ali, H., Scholer, F., Spina, D.: Walert: Putting conversational search knowledge into action by building and evaluating a large language model-powered chatbot. In: Proceedings of the ACM Conference on Information Interaction and Retrieval (CHIIR'24) (2024), arXiv-2401. https://ui.adsabs.harvard.edu/abs/2024arXiv240107216P/abstract (Visited on 02/10/2024). https://doi.org/10.1145/3627508.3638309
63. Peltier, C., Becker, M.W.: Individual differences predict low prevalence visual search performance. Cognit. Res. Principl. Impl. (2)(1), 5 (2017). ISSN: 2365-7464. https://doi.org/10.1186/s41235-016-0042-3. http://cognitiveresearchjournal.springeropen.com/articles/10.1186/s41235-016-0042-3 (Visited on 03/16/2024)
64. Radlinski, F., Craswell, N.: A theoretical framework for conversational search. In: Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval. CHIIR '17, pp. 117–126. ACM, Oslo (2017). ISBN: 978-1-4503-4677-1. https://doi.org/10.1145/3020165.3020183. https://dl.acm.org/doi/10.1145/3020165.3020183 (Visited on 03/06/2024)
65. Robinson, A., Thorne, W., Wu, B.P., Pandor, A., Essat, M., Stevenson, M., Song, X.: Bio-SIEVE: Exploring Instruction Tuning Large Language Models for Systematic Review Automation (2023). arXiv: 2308.06610 [cs]. (Visited on 03/25/2024)
66. Roegiest, A., Pinkosova, Z.: Generative information systems are great if you can read. In: Proceedings of the 2024 Conference on Human Information Interaction and Retrieval. CHIIR '24, pp. 165–177. Association for Computing Machinery, Sheffield (2024). ISBN: 9798400704345. https://doi.org/10.1145/3627508.3638345
67. Sahib, N.G., Tombros, A., Stockman, T.: A comparative analysis of the information-seeking behavior of visually impaired and sighted searchers. J. Am. Soc. Inf. Sci. Technol. **63**(2), 377–391 (2012). ISSN: 1532-2890. https://doi.org/10.1002/asi.21696. https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.21696 (Visited on 03/06/2024)
68. Shah, C., White, R., Thomas, P., Mitra, B., Sarkar, S., Belkin, N.: Taking search to task. In: Proceedings of the 2023 Conference on Human Information Interaction and Retrieval. CHIIR '23: ACM SIGIR, pp. 1–13. ACM, Austin, TX (2023). ISBN: 9798400700354. https://doi.org/10.1145/3576840.3578288. https://dl.acm.org/doi/10.1145/3576840.3578288 (Visited on 01/18/2024)
69. Shah, C., White, R.W., Andersen, R., Buscher, G., Counts, S., Sarkar Snigdha Das, S., Montazer, A., Manivannan, S., Neville, J., Ni, X., Rangan, N., Safavi, T., Suri, S., Wan,

M., Wang, L., Yang, L.: Using Large Language Models to Generate, Validate, and Apply User Intent Taxonomies (2024). (Visited on 02/16/2024). https://doi.org/10.48550/arXiv.2309.13063
70. Sitter, S., Stein, A.: Modeling information-seeking dialogues: The conversational roles (COR) model. Rev. Inf. Sci. **1**(1), 165–180 (1996). http://www.fb10.uni-bremen.de/anglistik/langpro/webspace/jb/info-pages/misc/sds/cor.pdf (Visited on 03/06/2024)
71. Soufan, A., Ruthven, I., Azzopardi, L.: Untangling the concept of task in information seeking and retrieval. In: Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval. ICTIR '21, pp. 73–81. ACM, Virtual Event (2021). ISBN: 978-1-4503-8611-1. https://doi.org/10.1145/3471158.3472259. https://dl.acm.org/doi/10.1145/3471158.3472259 (Visited on 02/29/2024)
72. Spink, A.: A user-centered approach to evaluating human interaction with web search engines: an exploratory study. Inf. Process. Manag. **38**(3), 401–426 (2002). ISSN: 03064573. https://doi.org/10.1016/S0306-4573(01)00036-X. (Visited on 04/07/2024)
73. Suh, M.(Mia), Youngblom, E., Terry, M., Cai, C.J.: AI as social glue: Uncovering the roles of deep generative AI during social music composition. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. CHI '21, pp. 1–11. Association for Computing Machinery, New York, NY (2021). ISBN: 978-1-4503-8096-6. https://doi.org/10.1145/3411764.3445219. https://dl.acm.org/doi/10.1145/3411764.3445219 (Visited on 02/26/2024)
74. Taylor, R.S.: The process of asking questions. Am. Document. **13**(4), 391–396 (1962). ISSN: 0096-946X, 1936-6108. https://doi.org/10.1002/asi.5090130405. https://onlinelibrary.wiley.com/doi/10.1002/asi.5090130405 (Visited on 03/05/2024)
75. Teevan, J., Jones, W., Bederson, B.B.: Personal information management. Commun. ACM **49**(1), 40–43 (2006). https://doi.org/10.1145/1107458.1107488
76. Teevan, J., Dumais, S.T., Horvitz, E.: Potential for personalization. ACM Trans. Comput.-Human Interact. **17**(1), 1–31 (2010). ISSN: 1073-0516, 1557-7325. https://doi.org/10.1145/1721831.1721835. https://dl.acm.org/doi/10.1145/1721831.1721835 (Visited on 03/12/2024)
77. Thomas, P., Spielman, S., Craswell, N., Mitra, B.: Large Language Models Can Accurately Predict Searcher Preferences (2023). arXiv: 2309.10621 [cs]. http://arxiv.org/abs/2309.10621 (Visited on 03/05/2024) Published at SIGIR'24: https://dl.acm.org/doi/10.1145/3626772.3657707
78. Trippas, J.R., Spina, D., Sanderson, M., Cavedon, L.: Results presentation methods for a spoken conversational search system. In: Proceedings of the First International Workshop on Novel Web Search Interfaces and Systems. CIKM'15: 24th ACM International Conference on Information and Knowledge Management, pp. 13–15. ACM, Melbourne (2015). ISBN: 978-1-4503-3789-2. https://doi.org/10.1145/2810355.2810356. https://dl.acm.org/doi/10.1145/2810355.2810356 (Visited on 03/06/2024)
79. Trippas, J.R., Spina, D., Cavedon, L., Joho, H., Sanderson, M.: Informing the design of spoken conversational search: Perspective paper. In: Proceedings of the 2018 Conference on Human Information Interaction & Retrieval—CHIIR '18. The 2018 Conference, pp. 32–41. ACM Press, New Brunswick, NJ (2018). ISBN: 978-1-4503-4925-3. https://doi.org/10.1145/3176349.3176387. http://dl.acm.org/citation.cfm?doid=3176349.3176387 (Visited on 03/06/2024)
80. Trippas, J.R., Spina, D., Scholer, F., Awadallah, A.H., Bailey, P., Bennett, P.N., White, R.W., Liono, J., Ren, Y., Salim, F.D., Sanderson, M.: Learning about work tasks to inform intelligent assistant design. In: Proceedings of the 2019 Conference on Human Information Interaction and Retrieval. CHIIR '19. ACM, Glasgow Scotland (2019), pp. 5–14. ISBN: 978-1-4503-6025-8. https://doi.org/10.1145/3295750.3298934. https://dl.acm.org/doi/10.1145/3295750.3298934 (Visited on 03/04/2024)
81. Trippas, J. R., Al Lawati, S. F. D., Mackenzie, J., Gallagher, L. : What do users really ask large language models? An initial log analysis of google bard interactions in the wild. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1–5. Association for Computing Machinery. New York, NY (2024). https://doi.org/10.1145/3626772.3657914

82. Vakkari, P.: Task-based information searching. Ann. Rev. Inf. Sci. Technol. **37**(1), 413–464 (2003). ISSN: 0066-4200, 1550-8382. https://doi.org/10.1002/aris.1440370110. (Visited on 03/28/2024)
83. Wang, S., Scells, H., Koopman, B., Zuccon, G.: Can ChatGPT write a good boolean query for systematic review literature search? In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1426–1436. ACM, Taipei (2023). ISBN: 978-1-4503-9408-6. https://doi.org/10.1145/3539618.3591703. (Visited on 03/25/2024)
84. Wang, J., Ma, W., Sun, P., Zhang, M., Nie, J.Y.: Understanding User Experience in Large Language Model Interactions (2024). arXiv: 2401.08329 [cs]. http://arxiv.org/abs/2401.08329 (Visited on 02/05/2024)
85. White, M.: Enterprise Search: Enhancing Business Performance. Enhancing Business Performance. O'Reilly Media, Sebastopol (2015). ISBN: 978-1-4919-1551-6
86. White, R.W.: Tasks, copilots, and the future of search: a Keynote at SIGIR 2023. ACM SIGIR Forum **57**(2), 4:1–4:8 (2024). ISSN: 0163-5840. https://doi.org/10.1145/3642979.3642985. https://dl.acm.org/doi/10.1145/3642979.3642985 (Visited on 02/10/2024)
87. White, R.W., Dumais, S.T., Teevan, J.: Characterizing the influence of domain expertise on web search behavior. In: Proceedings of the Second ACM International Conference on Web Search and Data Mining. WSDM'09, pp. 132–141. ACM, Barcelona (2009). ISBN: 978-1-60558-390-7. https://doi.org/10.1145/1498759.1498819. https://dl.acm.org/doi/10.1145/1498759.1498819 (Visited on 03/15/2024)
88. Wildemuth, B., Freund, L., Toms, E.G.: Untangling search task complexity and difficulty in the context of interactive information retrieval studies. J. Document, **70**(6), 1118–1140 (2014). https://www.emerald.com/insight/content/doi/10.1108/JD-03-2014-0056/full/html (Visited on 02/29/2024)
89. Wright, S.A., Schultz, A.E.: The rising tide of artificial intelligence and business automation: Developing an ethical framework. Business Horiz. **61**(6), 823–832 (2018). ISSN: 00076813. https://doi.org/10.1016/j.bushor.2018.07.001. https://linkinghub.elsevier.com/retrieve/pii/S0007681318301046 (Visited on 03/06/2024)
90. Wu, W.-C., Kelly, D.: Online search stopping behaviors: An investigation of query abandonment and task stopping. Proc. Am. Soc. Inf. Sci. Technol. **51**(1), 1–10 (2014). ISSN: 0044-7870, 1550-8390. https://doi.org/10.1002/meet.2014.14505101030. (Visited on 04/07/2024)
91. Xie, I.: Dimensions of tasks: influences on information-seeking and retrieving process. J. Document. **65**(3), 339–366 (2009). ISSN: 0022-0418. https://doi.org/10.1108/00220410910952384. https://doi.org/10.1108/00220410910952384 (Visited on 02/29/2024)
92. Xie, Y., Pan, Z., Ma, J., Jie, L., Mei, Q.: A prompt log analysis of text-to-image generation systems. In: Proceedings of the ACM Web Conference 2023. WWW '23: The ACM Web Conference 2023, pp. 3892–3902. ACM, Austin, TX (2023). ISBN: 978-1-4503-9416-1. https://doi.org/10.1145/3543507.3587430. https://dl.acm.org/doi/10.1145/3543507.3587430 (Visited on 02/10/2024)
93. Zamani, H., Trippas, J.R., Dalton, J., Radlinski, F.: Conversational information seeking. Foundat. Trends®Inf. Retr. **17**(3–4), 244–456 (2023). https://www.nowpublishers.com/article/Details/INR-081 (Visited on 03/06/2024)
94. Zendel, O., Culpepper,J.S., Scholer, F., Thomas, P.: Enhancing human annotation: Leveraging large language models and efficient batch processing. In: Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval, pp. 340–345. ACM, Sheffield (2024). ISBN: 9798400704345. https://doi.org/10.1145/3627508.3638322. (Visited on 03/28/2024)

Chapter 5
Improving Generative Information Retrieval Systems Based on User Feedback

Qingyao Ai, Zhicheng Dou, and Min Zhang

Abstract In this chapter, we discuss how to improve Generative Information Retrieval (GenIR) systems based on user feedback. Before describing the approaches, it is necessary to be aware that the concept of "user" has been extended in the interactions with the GenIR systems. Different types of feedback information and strategies are also provided. Then the alignment techniques are highlighted in terms of objectives and methods. Following this, various ways of learning from user feedback in GenIR are presented, including continual learning, learning and ranking in the conversational context, and prompt learning. Through this comprehensive exploration, it becomes evident that innovative techniques are being proposed beyond traditional methods of utilizing user feedback and contribute significantly to the evolution of GenIR in the new era. We also summarize some challenging topics and future directions that require further investigation.

5.1 Introduction

For an information access (IA) system that is built to provide useful information to users, interactions with users are definitely crucial and important. There are two types of user feedback, *explicit feedback* and *implicit feedback*, based on whether the user's opinions or preferences regarding the provided information are expressed

Qingyao Ai and Zhicheng Dou contributed equally to this work.

Q. Ai · M. Zhang (✉)
Department of Computer Science and Technology, Tsinghua University, Beijing, China

Quanchegn Lab, Tsinghua University, Jinan, China
e-mail: aiqy@tsinghua.edu.cn; z-m@tsinghua.edu.cn

Z. Dou
Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
e-mail: dou@ruc.edu.cn

R. W. White, C. Shah (eds.), *Information Access in the Era of Generative AI*,
The Information Retrieval Series 51, https://doi.org/10.1007/978-3-031-73147-1_5

directly through clear statements or indirectly with some signals. (Section 5.1.2 will delve into more details about explicit and implicit feedback.)

Similar to traditional information retrieval (IR), GenIR also leverages users' feedback in various ways to improve the system's capability and performance. However, upon closer examination, it becomes evident that there are distinct factors that specifically contribute to GenIR systems, in terms of types of feedback information and utilization strategies. We also raise the attention to the definition of "user" in the new era.

In this chapter, we first discuss the feedback factors and the differences in the new era in the first section. Subsequently, we provide detailed descriptions and discussions on the alignment with user factors in GenIR in Sect. 5.2. "Alignment" in generative models usually refers to the process of fine-tuning the model to ensure its generated text aligns with specific goals, values, or user intentions, often through human feedback or instruction fine-tuning. Subsequent to this, Sect. 5.3 explores the user feedback learning in GenIR, which involves training and improving the GenIR system through understanding users' intents, interests, or preferences based on their historical feedback. A summary is given in the final section, Sect. 5.4, along with discussions on the challenges and future directions.

5.1.1 Concept of User in the Generative Information Retrieval Era

Over the past years, discussions about the ***user*** in information access systems, including search engines, recommender systems, question-answering platforms, etc., have primarily centered around human beings interacting with these systems. However, in the emerging GenIR era, where the new IR system is designed to connect with human beings, tools, or even other GenIR systems, the concept of the ***user*** has been enlarged to a much broader sense.

A ***user*** of the GenIR system can now include:

- A human being who uses the GenIR system, similar to the *user* in traditional IR system
- A Large Language Model (LLM) agent that can send or receive information to or from the GenIR system or engage in bidirectional information exchange, also refers to the *agent* in publications
- Another system, tool, or application that interacts with the GenIR system, sometimes termed as the *client* in technical context

Interactions from the traditional *users*, *agents*, and *clients* should all be taken into consideration as user feedback, whether from real or virtual ***users*** within the GenIR system.

5.1.2 User Feedback

The GenIR system still maintains two fundamental types of user feedback, *implicit feedback* and *explicit feedback* as usual. However, the scope of feedback information has been significantly broadened.

Consistently, one of the major feedback is the user interaction history. Interactions with the GenIR system encompass queries, questions, clicks, views, purchases, comments, and more. Such information is usually taken as *implicit feedback* information. In contrast, users' explicit annotations such as favorites, likes, ratings, or direct feedback on satisfaction constitute *explicit feedback* information. In GenIR systems, a notable difference lies in the increased availability of *explicit feedback* provided by users through system inputs or prompts. Nowadays, users are accustomed to communicating their specific requests, intentions, and interests to the GenIR system. In many instances, multi-round conversations have become commonplace.

In GenIR systems, feedback information manifests in two distinct forms:

(1) Numerical information, primarily consisting of identifier-level data that identifies the items with which the user has interacted. Sometimes, this information is presented in sequential order. Such information offers a glimpse into the user's behavioral patterns.
(2) Detailed information encompassing multiple modalities. Textual data is the most commonly utilized, including query text, item titles, content, user comments, and questions. As LLM technology rapidly advances, longer natural language expressions are increasingly being leveraged. Multimedia inputs, such as images, music, or videos, sometimes integrated with visual LLM (e.g. [1]), have also garnered significant attention. This multifaceted feedback allows for a richer, more nuanced understanding of the user's preferences and interactions within the GenIR system.

How to leverage such rich user feedback information smoothly in the GenIR system to improve performance is a crucial part of the new LLM era. The next section briefly summarizes the strategies for using user feedback information in GenIR systems.

5.1.3 Strategies for Generative Information Retrieval System Improvement with User Feedback

Introducing user feedback information into GenIR systems can be facilitated through prompt engineering or instruction construction, which is perhaps the most straightforward approach [1–4]. Historical user interactions can be encoded using various types of index, such as title-based indexing, random indexing, independent indexing, sequential indexing, semantic indexing, or collaborative indexing [5].

These indexing inputs can then serve as prompts for the system, as illustrated in the following Fig. 5.1. The prompt strategy is commonly employed in zero-shot or in-context learning scenarios, where LLMs are directly utilized as the information system. By leveraging this approach, GenIR systems can efficiently integrate user feedback to enhance their performance.

The second strategy involves leveraging historical interaction information for fine-tuning the parameters of the LLM. This approach aligns the user or item representations within the pre-trained language model. Typically, this information is utilized as either an identifier index or a text index with item content, reviews, etc. [6–8]. In certain instances, the user-item collaborative information is initially encoded by a traditional IR system to generate embeddings for users or items [9–11]. These embeddings implicitly contain collaborative interaction feedback, which is subsequently utilized in the fine-tuning and alignment process.

The third strategy focuses on capturing the user's implicit or explicit preferences and indicating both vague and specific intents. By integrating these preferences and intents, the GenIR system is able to identify the user's specific task [12]. For instance, an implicit preference associated with a specific search intent for a mobile phone would be linked to a product-search task, while an implicit preference with a more vague intent might lead to a recommendation task.

The fourth strategy is to take user behavior as the action, reward, or even the evaluator within an agent-based GenIR system [13–17]. This approach effectively guides the system in learning the appropriate actions and responses. In such GenIR systems, user feedback information plays a crucial role in the system's learning and refinement process across multiple rounds of interaction. By continuously incorporating user feedback, the system can adapt and improve its performance.

It is anticipated that even more strategies will emerge as research continues. In the subsequent sections, we discuss deeper into these various approaches, exploring the alignment with user preferences and the learning mechanism.

5.2 Alignment with User Factor in GenIR

Alignment techniques have been widely recognized as one of the key components for the construction of effective LLMs. In the first technical report of ChatGPT [18], alignment techniques such as Reinforcement Learning with Human Feedback (RLHF) [19, 20] have already been extensively used in the training and construction of the chat system. Right after the success of ChatGPT, LLM alignment has become one of the most important research directions in the community of Natural Language Processing (NLP) and, as discussed in the later part of this chapter, also has significant potential for building information access systems in the era of generative AI.

Despite the recent surge in interest in alignment technology following the success of ChatGPT, it is important to note that research in this area has been ongoing for many years. In fact, it is difficult to pinpoint the exact moment when alignment

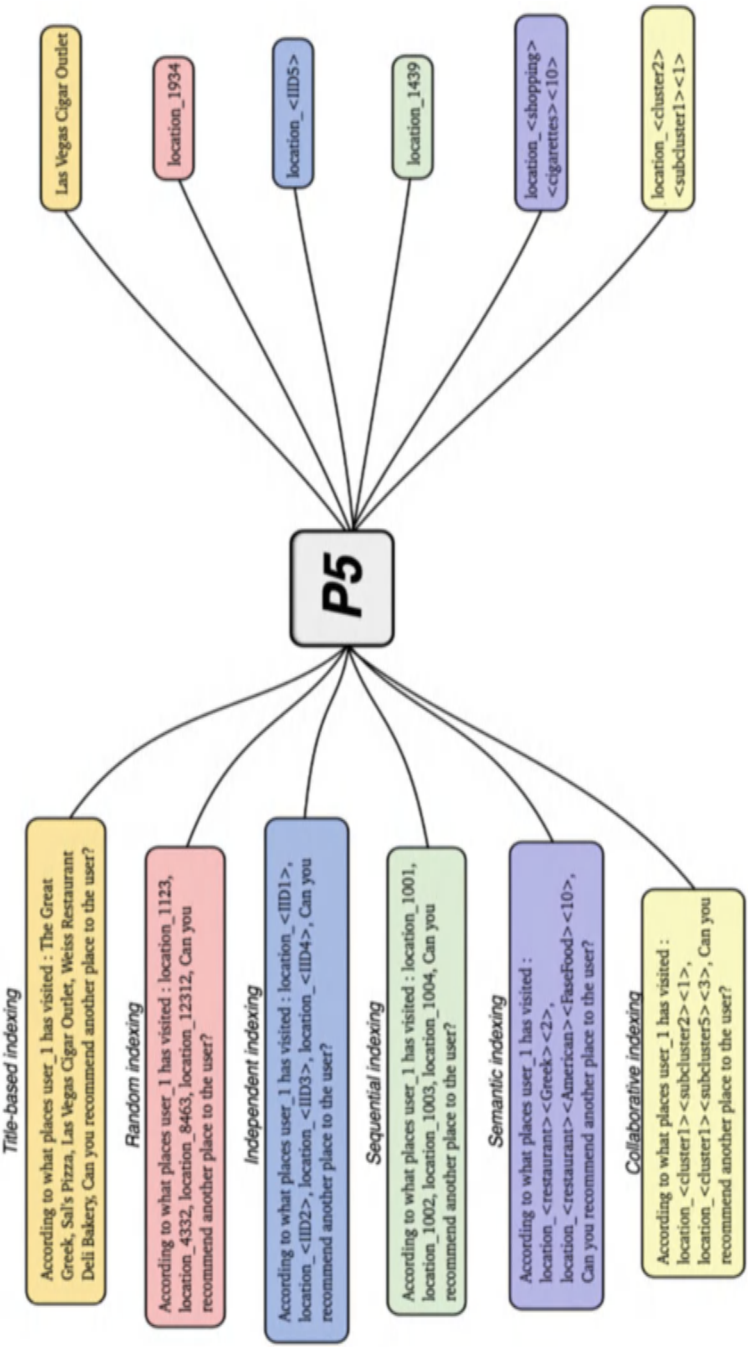

Fig. 5.1 Different types of indexing as prompt input to GenIR systems proposed by Geng et al. [5]

techniques were first introduced to the studies of large language models. In the early days of neural language models, researchers focused mainly on designing more powerful model structures [21–23] and training techniques [24, 25] to enable LLMs to process information and learn patterns from massive amounts of data. The performance of language models in those days was not strong enough to understand user's instructions and generalize to multiple types of tasks. However, as the capabilities of LLMs grows, concerns related to aspects other than task performance, such as ethic, robustness, and have gradually become obstacles that prevent LLMs from applications in real-world scenarios. For example, in 2016, users had successfully tricked a Twitter bot constructed by Microsoft (i.e., Tay) to produce statements that were misogynistic [26]; in 2022, Meta's BlenderBot 3 had been "taught" to be racist right after it was released to the public [27]. Therefore, a greater emphasis, especially on model safety and ethical issues, was placed on aligning LLM outputs with human values and preferences since 2017. It was around this period when a group of alignment methods, including the famous RLHF, have been used in the training of LLMs. As of today, almost all LLMs must go through an alignment process before being launched and released to the public.

In this section, we focus on the introduction and discussion of LLM alignment from the perspectives of information access. Besides the safety and ethical problems of LLMs, there are also unique challenges and needs of LLMs when applied to information access tasks. Those unique challenges also lead to unique methods and research directions that have great potential for information accessing in the era of generative AI. In the following, we first provide a brief introduction of the common objectives for alignments in LLMs and information accessing and then introduce a couple of representative alignment methods in the field. Last but not least, we discuss the connections and differences between alignments and other LLM techniques such as Supervised Fine-Tuning (SFT) from the perspective of information retrieval and access.

5.2.1 Alignment Objective

LLM alignment is a cornerstone in the development of generative AI systems, particularly in the context of ensuring that these models act in ways that are beneficial, safe, and aligned with human values and intentions. On the one hand, as LLMs become more powerful every day, their potential impact on human society increases, making the alignment of these models with ethical standards and user intentions an essential objective [18, 20, 28]. On the other hand, LLM alignment techniques can supplement SFT or other training techniques in equipping LLMs with abilities or characteristics desired for diverse tasks and applications in specific domains [29]. From the perspectives of information access, the objective of LLM alignment techniques is multifaceted, with some parts of it aligning closely with other LLM applications and some parts of it diverging significantly from those widely considered in the development of general LLMs.

5.2.1.1 Objectives Shared by General LLM Applications

Similar to other LLM applications, the usage of LLMs in information accessing needs to prevent the possibility of outputs that are harmful from ethical perspectives or undesirable by user intents. Specifically, such objectives include but are not limited to the following.

Preventing Harmful Outputs One primary objective of LLM alignment techniques, shared by both information accessing and other applications, is to prevent models from generating harmful, biased, or inappropriate content [20] that violates the universal values of human beings. This includes outputs that could be misleading, factually incorrect, or that perpetuate harmful stereotypes. Before the era of generative AI, major information accessing systems usually focus on retrieving existing Web pages or documents created by humans to satisfy the user's information need. Therefore, the prevention of harmful outputs can be done directly through pre-processing such as spam detection and keyword filtering [30, 31]. With LLMs, however, controlling the outputs of information systems becomes significantly more difficult due to their stochastic nature [32]. Alignment techniques that prevent such harmful outputs through the post-training of LLMs have then become the most popular methods used in generative systems.

Aligning with User Intents Another critical aspect of LLM alignment is ensuring that models accurately understand and align with user intentions [29]. This means that LLMs must be adept at interpreting the context and nuances of user queries and generating responses that accurately reflect the user's desired outcome. User intent understanding is at the core of information accessing, and numerous methods have been proposed to solve this problem in the context of traditional matching-based IA systems [33–35] Unfortunately, as the internal knowledge structure of LLMs is still obscure, it is difficult (at least of today) to adopt our knowledge and experience obtained from previous studies directly to generative IA systems. LLM alignment is one of the most direct and practical methods to improve the system's ability in understanding user intents.

Adhering to Ethical Guidelines LLM alignment also involves adhering to ethical guidelines and principles. This encompasses a range of considerations, from ensuring privacy and data security to promoting fairness and avoiding discrimination. In information accessing, these are also of great importance in practice. Popular search engines before the era of generative AI have already been widely criticized for imposing biased exposure to information such as political statements and news [36, 37]. With more powerful yet nontransparent LLMs used in modern IA systems, such issues are becoming more intricate and vital. Developing generative IA systems that can balance fairness with relevance, respect user privacy, and treat sensitive topics with the appropriate level of care requires a deliberate and thoughtful approach. More importantly, as cultures, individuals, and groups may have vastly different views on what is considered appropriate, ethical, or aligned

with their intentions, we need methods that are both effective and efficient in terms of model adaption.

5.2.1.2 Objectives Unique to Information Access

Besides those common alignment objectives shared by general LLMs, information access also has unique challenges that must be solved in order to construct effective generative IA systems. As the ultimate goal of IR and access is to satisfy users' information needs, to the best of our knowledge, the special characteristics needed by generative IA systems can be broadly categorized into two types in existing literature: the need for personalization and the capability of fine-grained information discrimination.[1]

Personalization At its core, personalization is about tailoring the interactions and information delivery of a system to the unique preferences, interests, and needs of an individual user [38]. The key idea is to transfer the user experience from a one-size-fits-all approach to a more intimate and relevant exchange [39, 40]. In general applications of LLMs, this usually means speaking with the languages, styles, and values preferred by each user. In the context of information access, this also means understanding and utilizing the connection between the user's information need with time, locations, application scenarios, and all kinds of user information that potentially affect user's perceptions of information utility. Traditional personalization in information access focuses on the construction of user profiles and the design of algorithms and models that effectively incorporate user information into the analysis of information relevance. In the era of generative AI, while the structures of the models and systems have significantly changed, the needs of those two still exist. LLMs have strong in-context learning ability, which can implicitly create a user model simply by feeding the descriptions of user profile as prompts in the input user queries. Yet existing LLMs can only take inputs with text, images, or other standard multimedia formats, but user profiles go beyond these. How to construct and incorporate hyper-information like user-user, user-item, and item-item interactions effectively under the current model frameworks of LLMs and generative AI is still an open question. Alignment techniques, as flexible and relatively lightweight methods to optimize large generative models, are thus of great potential for personalization in generative information accessing.

Fine-Grained Discrimination With the exponential increase in the availability of digital content, the challenge of information access is no longer just finding relevant information but finding the most appropriate content among a set of potentially relevant items. Most existing studies on retrieval and ranking models are essentially developing better methods to analyze and discriminate input documents based on

[1] Please note that the categorization here is by no means inclusive as this is still an ongoing research topic.

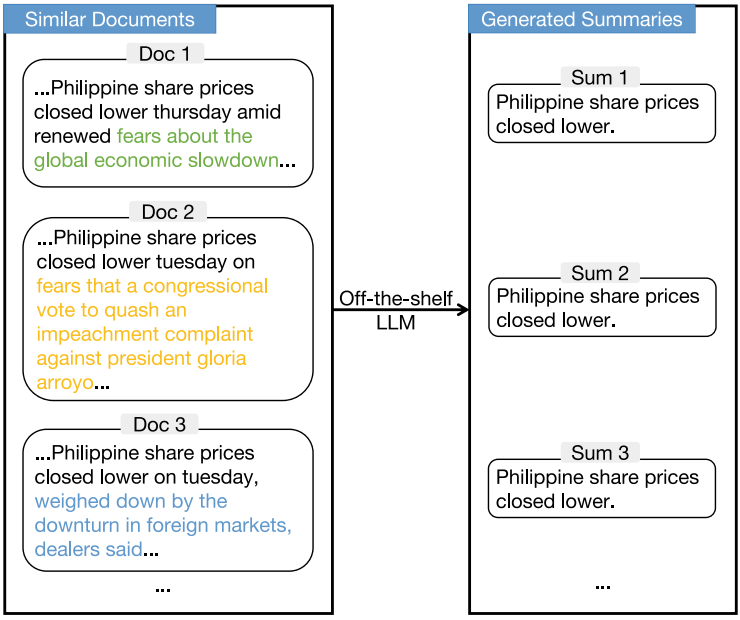

Fig. 5.2 Illustrations of LLMs application in document summarization for similar documents provided by Dong et al. [41]. The distinctive parts of each document are highlighted in different colors

their fine-grained differences and utility given users' queries. In the era of LLMs, while the final outputs of an IA system may no longer be a simple listing of result candidates, the ability of discriminating information in fine grains is still of great importance. It allows the system to navigate in complex data collections, identifying the subtleties that differentiate pieces of information in ways that are significant to the user. Yet the acquisition of such ability is usually not covered in the alignment process of off-the-shelf LLMs. An illustration example is provided by Dong et al. [41] in Fig. 5.2. When the request is the same (i.e., "create a summary of the document" in Fig. 5.2) and the input documents are similar, the off-the-shelf LLM (i.e., Flan-T5 in the figure) tends to produce identical responses to all documents. Such problems could be insignificant in many NLP applications where the quality of outputs is evaluated independently with each other. In information access, however, we often care about the discrimination of input documents more than we care about their absolute relevance or utility. For instance, if we generate identical snippets for similar documents retrieved by search engines, it would remarkably increase the difficulty for users in pinpointing the exact result that answer her needs. Because the ability to produce such discriminative outputs in fine grains can hardly be learned from the standard next token prediction tasks, one may need extra alignment process to enhance the model from this perspective [42].

5.2.2 Alignment Method

Alignment methods are strategies developed to steer the behavior of generative AI models toward desired outcomes. These methods often rely on a framework that involves computing rewards based on model outputs and using these rewards to optimize the model's performance. Specifically, this usually involves two steps: (1) the collection and computation of rewards, and (2) the optimization of model parameters based on the rewards. Most existing alignment methods are designed for general alignment objectives in NLP tasks, but given the great potential and importance of LLMs in future information access, several researchers have also started to investigate how to design alignment methods tailored for the needs of IA tasks. In this section, we first introduce a couple of well-developed reward collection methods in LLM alignments and then briefly describe several standard optimization methods that have been widely used in existing studies.

5.2.2.1 Collection of Rewards

The collection of rewards is a pivotal step in aligning LLMs. It evaluates the model's outputs/responses against certain criteria to determine how well they align with desired outcomes. Very much similar to the design of loss functions in Learning to Rank (LTR) [43], the nature of reward computations in LLM alignment can be broadly grouped into several categories based on the inputs and training paradigms of the reward functions. Specifically, if we borrow the terminology used in LTR literature [44], the reward collection methods can be categorized from two perspectives. From the perspective of reward function input, we have

- Pointwise input: Rewards are computed for LLM outputs based on each individual input data points independently.
- Groupwise (pairwise) input: Rewards are computed for LLM outputs based on a group (or pair) of different input data points together.

From the perspective of reward computation or reward function training, we have:

- Pointwise training: Rewards are computed on or reward functions are trained with each LLM output independently.
- Groupwise (pairwise) training: Rewards are computed on or reward functions are trained with a group (or pair) of LLM outputs together.

An illustration of the differences between those methods is depicted in Fig. 5.3. With this taxonomy, we introduce a couple of popular alignment methods in the following and summarize their types in Table 5.1. Careful readers may notice that all the reward methods here have a prefix "RL," which stands for *reinforcement learning*. While these reward collection methods are independent to the use of learning algorithms (which is discussed in Sect. 5.2.2.2), they are often referred to or analyzed together with reinforcement learning. For simplicity, we use the terminology widely used in the LLM literature to refer to them, but please note that

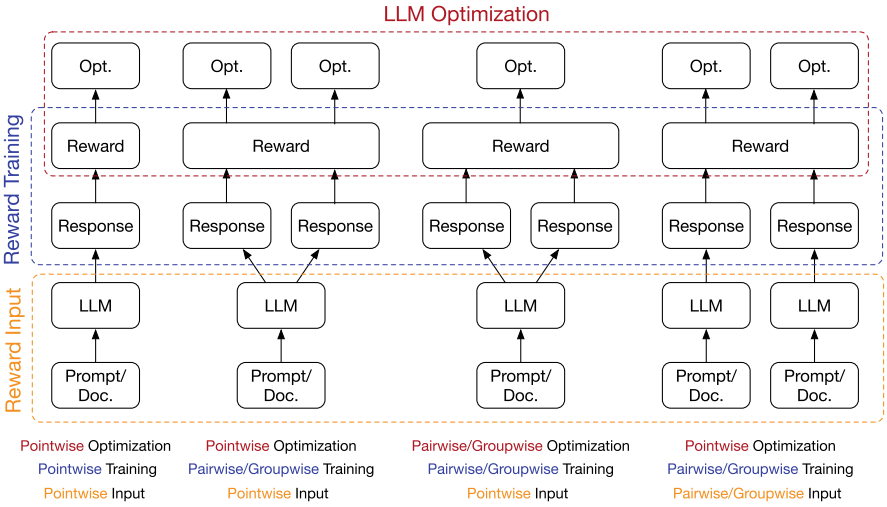

Fig. 5.3 An illustration of example reward collection and optimization methods in LLM alignment

Table 5.1 Example alignment methods and their categorization based on input and computation/training paradigms

	Input		Computation/training	
	Pointwise	Pairwise/groupwise	Pointwise	Pairwise/groupwise
RLHF	[19, 20, 45]		[20]	[19, 20, 45]
RLAIF	[46, 47]		[47]	[46, 47]
RLCF		[41]	[41]	

this does not indicate that alignment methods using these rewards must be developed under reinforcement learning frameworks.

Reinforcement Learning from Human Feedback (RLHF) To the best of our knowledge, RLHF is the most well-known and popular alignment method today. It has been recognized as one of the most important parts of ChatGPT [18]. RLHF collects feedback from human users on different LLM outputs to optimize model parameters accordingly [48, 49]. Since the feedback is directly collected from humans, RLHF is capable of aligning generative AI models with almost all related objectives such as output safety and ethical values. Typical RLHF process involves the generation of multiple output candidates from one or multiple LLMs given a single input. It can be treated as a pointwise input method because RLHF always collect rewards on output candidates generated for a single input (e.g., prompt). Then, with the output candidates generated by LLMs, RLHF further asks human annotators to judge the quality of each output and train reward functions accordingly. The annotation process could be pointwise (through not common in recent LLM literature), i.e., asking the annotator to give a rating to each output separately, or pairwise/groupwise, i.e., asking the annotator to provide a preference

or ranking of multiple outputs together. Accordingly, the final reward function learned from such annotation data is constructed with either pointwise, pairwise, or groupwise training data and thus can be categorized as pointwise, pairwise, or groupwise output methods.

Reinforcement Learning from AI Feedback (RLAIF) Despite the flexibility and effectiveness of RLHF, its needs for human-in-the-loop significantly raise the cost of model alignment and unpredictable variance, particularly when the feedback data are not large or reliable enough, in the optimization process. Therefore, researchers have also investigated extensively on how to conduct model alignment without supervised data. Given the rapid development of LLMs in the last two years, one of the trending alignment methods in both academic and industrial communities is RLAIF [46]. The motivation of RLAIF is to replace humans in the process of RLHF with a powerful LLM that can mimic human behaviors (which we refer to as the AI feedback model) to generate feedback for each model output. Based on this idea, it basically reuses the existing framework of RLHF with minor modifications and align LLMs for different objectives by prompting the AI feedback models with objective-related task descriptions or annotation guidelines (e.g., RLCD [47]). RLAIF is a pointwise input method and, theoretically, could be either pointwise, pairwise, or groupwise from the output perspective. However, as far as we know, none of the existing studies have used RLAIF with the groupwise output paradigm, probably because directly ranking multiple candidates is still a difficult task for modern LLMs [50].

Reinforcement Learning from Contrastive Feedback (RLCF) RLHF and RLAIF are powerful methods that have already been shown to be effective in many NLP tasks, but their applications to optimize alignment objectives specifically important in information access have, unfortunately, been unsuccessful so far. As discussed in Sect. 5.2.1, the ability to discriminate information in fine grains is the key to generate informative and useful outputs in IA scenarios, but studies have found that naively adopting those alignment techniques do not improve the model's performance in IA tasks as we expected [41, 42]. One of the key reasons lies in the input paradigms of RLHF and RLAIF. To generate outputs that are informative and discriminative, LLMs need to understand and capture what makes an input piece of information unique in the corpus or data collections. However, RLHF and RLAIF are developed with the pointwise input paradigm, and it is difficult, if not impossible, to teach LLMs to generate outputs unique to an input without seeing and comparing with other candidate inputs. Therefore, RLCF is proposed to conduct alignment with groupwise input and output paradigms for IR [41]. The idea is to let LLMs generate outputs for different inputs simultaneously and construct reward functions based on the comparison of each output for each input. For example, one can compare query generation or expansion candidates for a single document with those generated for other similar documents to improve LLMs' ability to capture the uniqueness of each document. The original RLCF method computes rewards with retrieval models to enhance the final LLM's effectiveness in IR tasks, but such groupwise input and output paradigms could have potentials in the alignment

of other objectives as well because, as widely acknowledged in LTR literature, groupwise methods have more capacity in complex objective modeling and less variance in parameter optimization.

5.2.2.2 Parameter Optimization

With the rewards collected for particular alignment objectives, the next step is to optimize the parameters of LLMs. Similar to the methods we used for feedback collection, the optimization algorithms for LLM alignments can also be broadly categorized based on their inputs. In this section, we only describe several representative optimization algorithms for LLM alignments, namely, Proximal Policy Optimization (PPO) [51], Direct Preference Optimization (DPO) [52], and ranking-based optimization [53, 54]. Please note that this is still an ongoing research direction, and the methods discussed here are far from covering all potential solutions in the area.

Proximal Policy Optimization (PPO) While PPO is not the first optimization algorithm used for reinforcement learning in LLMs, it is considered one of the most popular methods in LLM literature today, partially thanks to its applications in OpenAI products[2] and ChatGPT [18]. The primary goal of PPO, so as reinforcement learning techniques in general, is to train models to make sequences of decisions by rewarding desired behaviors and penalizing undesired ones. In the context of LLM alignment, PPO can be used with different types of rewards discussed in Sect. 5.2.2.1, and it stands out from other RL algorithms due to its better balance of simplicity, efficiency, and effectiveness, compared to its predecessor such as Trust Region Policy Optimization (TRPO) [55]. The core idea of PPO is to take small steps in policy space to improve the model while ensuring that the new policy is not too different from the old one. This is achieved through a clip mechanism, which limits the size of the policy update at each iteration. The clipped objective function helps the model conduct gradient descent while preventing overly large updates that could lead to performance collapse, a common issue in earlier RL methods that could lead to unstable training processes. To compute such objective functions, PPO needs a reward model that can directly estimate the gain or loss of a particular action (e.g., the output of LLMs). Therefore, it is usually used with pointwise output reward collection methods such as those shown in Table 5.1. RLHF with PPO is widely used as the backbone alignment method of many famous LLMs such as GPTs, Llamas, etc.

Direct Preference Optimization (DPO) A typical alignment method using PPO needs to create a reward model from the collected feedback data to score each LLM output for parameter optimization. While the construction of such reward model is possible for most types of rewards, it does not necessarily fit the nature

[2] https://openai.com/research/openai-baselines-ppo

characteristics of each reward type. It is essentially a pointwise output method that creates independent labels for each output candidate, and converting pairwise or groupwise data (e.g., human preferences over different LLM outputs) to pointwise format usually leads to significant information loss and more variance in model optimization [56]. To this end, DPO [52] is proposed to optimize model parameters directly with preference data. Careful readers may notice that all reward collection methods discussed in Sect. 5.2.2.1 has a prefix "RL." This is partially because most popular alignment methods use reinforcement learning for parameter optimization. In contrast, DPO directly computes model gradients without using reinforcement learning by minimizing the Kullback-Leibler divergence between the ground-truth preference and LLM output distributions. This method is highly similar to standard pairwise or listwise methods used in learning-to-rank literature [57–59]. As pointed out by Rafailov et al. [52], it outperforms popular reinforcement learning methods based on PPO in both effectiveness and robustness. Considering that most alignment objectives (e.g., harmfulness, ethic, etc.) involve significant human subjectivity, preference-based optimization methods could be more promising in theory. Besides, from the research perspective of information access, this also indicates that techniques from classic retrieval and ranking studies may provide important guidelines for the design of future LLM alignment methods.

Ranking-Based Optimization Following similar motivations with DPO, several methods have been proposed to further extend the utilization of pairwise preference reward to listwise reward for LLM alignments. Notable representatives include rank responses to align language models with human feedback (RRHF) [53] and retrieval-augmented fine tuning (RAFT) [54]. Despite data processing and implementation details, RRHF could be treated as a listwise version of DPO. Its core idea is to score multiple responses via a crafted probability function and learns to align the corresponding probability distribution with human preferences through a ranking loss constructed based on the variation of hinge functions [60]. RAFT approaches the problem from a different angle. It ranks multiple LLM response candidates based on preference data (or a reward model learned from preference data) and selects the samples with highest rewards to fine-tune the LLM. It is well acknowledged in IR literature that listwise ranking methods have better potentials in fitting preference data, both in theory and in practice [43]. Therefore, methods like RRHF and RAFT have both theoretical and empirical advantages over standard PPO and DPO methods in model alignments. While such advantages are not fully explored in the general tasks such as dialog generation and machine translation, they could be important for the application of generative models in information access since many IA tasks exhibit natural needs of response discrimination and ranking.

5.3 Learning from User Feedback in GenIR

As introduced in the previous section, in the training procedure of LLMs, user feedback is very important to align the values of LLMs with humans. RLHF is

widely adopted as the final training stage of LLMs. Besides LLMs, user feedback is vital in IR systems. It is commonly used as the final optimization target. For example, the CTR task [61–63] aims to predict the click-through rate, and the ranking models are usually tuned toward the user click signals [64]. Besides, some user-centric tasks such as recommendation and personalized search collect and utilize user history feedback to provide tailored results for users' current information needs [65–67]. For example, personalized search models apply the query attention technique to aggregate user search and click histories to build user preferences under their current queries [68].

In the era of LLMs, many personalized search and recommendation [69] approaches devise LLMs to understand user histories and construct user interests. For example, inspired by the memorization mechanism in cognitive science, Zhou et al. [70] designed several memory modules including sensory memory, short-term memory, and long-term memory to facilitate LLMs to retrieve relevant user histories to current intents. In the recommendation area, LLMs are usually adopted to enrich user histories since they store extensive world knowledge [69]. Recently, LLM-based agents [71] have attracted much attention from academia and industry. These agents have abilities to memorize past behaviors, make plans to achieve final tasks, and take action under current situations. It is worth exploring to involve user feedback in search agents to solve IR tasks.

5.3.1 Continual Learning

IR systems are designed to retrieve relevant information based on user queries. As users interact with these systems, they generate valuable data, such as search queries, clicked results, dwell time on pages, and explicit feedback like ratings or comments, that can be used to improve the system's understanding of user intent and preference. By leveraging the collected data, IR systems can progressively refine their retrieval algorithms, leading to more accurate and personalized search results. This process forms the basic paradigm of continual learning [72] in an IR system. Many methods have been proposed to incorporate user feedback data into optimizing a traditional IR system [73, 74].

Continual learning is also vital for generative systems like LLMs to be regularly updated to include the latest human knowledge and feedback [75, 76]. As introduced by Wu et al. [75], continual learning could be applied with different training stages, including pre-training, fine-tuning, and alignment. Traditional IR ranking models are relatively small and can be easily updated in a batch manner. The separated document index could be updated dynamically when new documents are available, and hence, it is relatively easier for the entire system to update continually. Contrarily, generative IR models are usually large, and all information about the documents and the ranking are embedded in the same generative model. It is much more challenging to update such systems. For example, LLMs have the

"catastrophic forgetting" problem [76]: the performance of the old task based on previous knowledge domains will degrade when new user data are fed.

5.3.2 Learning and Ranking in Conversation Context

In the interaction with conversational search systems, users may generate various types of feedback, such as asking follow-up questions based on the system's responses, expressing dissatisfaction with the system's responses, and providing clarification to the system's inquiries. These natural language-based explicit user feedbacks are crucial for helping the conversational search system continuously meet user needs and optimize its performance. LLMs possess powerful capabilities for understanding and generating dialogue, offering significant opportunities for better comprehension of user feedback in conversational search.

In conversational search, user questions are usually ambiguous and can only be correctly understood based on the conversation context. Traditional methods are struggled in dealing with the long and complex conversation context, resulting in unsatisfactory retrieval performance. In contrast, LLMs show outstanding capability in conversation understanding and therefore can largely improve the accuracy of conversational search intent understanding. Mao et al. [77] proposed a prompting framework to leverage LLMs to perform conversational query rewriting. They developed three aggregation methods to aggregate the generated rewrites and hypothetical responses from LLMs to form a better search intent representation for conversational search. Similarly, Ye et al. [78] also proposed to utilize LLMs to generate informative query rewrites through well-designed instructions. Their results showed that the search performance can be largely improved after utilizing the generated contents from LLMs. Furthermore, LLMs can also be used to mimic users' search behaviors and generate more high-quality search session data. Conversational search systems need massive session-level relevance data for improvements, and LLMs can significantly facilitate the data curation process. One of such related works is ConvAug [79], which is a cognition-based framework that leverages LLMs to generate more conversational search sessions. These pseudo sessions can help conversational retrievers capture the diverse nature of conversational contexts to be more effective and robust.

Besides, in the interaction process of conversational search, the user's responses to the system responses (e.g., clarification questions and inaccurate responses) are also crucial for capturing users' real information needs and unique preferences. Recently, the Text Retrieval Conference (TREC) organized an interactive knowledge assistance track (iKAT) [80] for studying collaborative conversational information-seeking systems that can customize and personalize their response based on what they learn about and from the user. Existing works [81] have demonstrated the strong performance of LLMs in aggregating and inferencing users' references. Therefore, LLMs have a large potential to improve the utilization of this type of valuable user initiative feedback to model the user profiles and provide a more accurate and

personalized search experience. LLMs can also be employed to identify the type of users' responses, such as distinguishing whether the response is a new question, a reply to a clarification request, or a hint for correcting a previous answer. We do not need to train a separate model for this intent identification. Instead, we can stream the modeling of all interaction processes in conversational search through prompting with LLMs. The massive knowledge about conversation patterns and the world of LLMs also makes it a promising end-to-end foundation to be an end-to-end foundation model for personalized conversational search systems.

5.3.3 Prompt Learning

LLMs have demonstrated excellent performance in language understanding, making them also promising for learning user feedback, particularly in the area of query refinement. In search engines and similar platforms, understanding the context and intent behind user queries is crucial for delivering accurate and relevant results. We consider two possible ways of applying LLMs to query refinement.

Directly Prompting LLMs for Query Refinement Given the substantial computational resources required for fine-tuning LLMs, a more straightforward approach is prompt learning. This method entails describing the task in text and prompting the models to solve it. Upon gathering user feedback, LLMs can analyze the feedback, comprehend the underlying meaning, and suggest refinements for the user input query, thereby enhancing retrieval performance. Previous studies [82–84] have applied LLMs to query rewriting. The results indicate that LLMs can generate effective user queries, particularly when provided with few-shot demonstrations. Furthermore, LLMs have shown superior performance in conversational query rewriting [77], attributable to the availability of more comprehensive contextual information. These findings indicate the significant potential of applying LLMs to query refinement.

Distilling Knowledge from LLMs to Smaller Models In practice, it is still costly to use LLMs in real applications. Under this circumstance, training a small model specifically for query refinement emerges as a more favorable approach. This can be achieved by employing LLMs to refine queries based on user feedback, subsequently utilizing these refined queries as labels to train a specialized model. This strategy not only reduces computational overhead but also maintains the efficacy of the learning process, thereby offering a pragmatic solution for real-world applications.

5.4 Summary

In this chapter, we delve into how user feedback can enhance the GenIR system. Firstly, we clarify the concept of "user" and subsequently explore the diverse types and forms of user feedback information. Furthermore, we outline four established strategies that leverage user feedback effectively. Secondly, we provide a detailed account of the crucial technique of alignment in the GenIR context, discussing both the alignment objective and various methods employed. Finally, we highlight the significance of user feedback learning in GenIR, encompassing human-in-the-loop approaches, continuous learning, learning and ranking within conversational contexts, as well as prompt learning. Through this comprehensive exploration, it becomes evident that innovative techniques are being proposed beyond traditional methods of utilizing user feedback and contribute significantly to the evolution of GenIR in the new era.

There are some challenging topics and future directions that we believe need further exploration, such as:

- **User intention understanding within the GenIR system**. For example, how do we precisely determine the user's true intent? How do we manage shifts in user intentions during multi-turn interactions or conversations with the GenIR system? When we broaden the concept of ***user*** to also include agents/clients that interact with the GenIR system, could this lead to self-feedback loops within the GenIR system and a bias toward artificial intentions?
- **User behavior analysis and understanding with "less but rich feedback."** As the end user interacts with generated responses, we may receive less feedback than traditional IR systems (e.g., clicks on search engine results pages). On the other hand, the feedback is richer (e.g., an explicit feedback in the conversation like "Thank you, that's really helpful" or a detailed follow-up indicating continued engagement when the information need is not met). Studying the utilization of limited yet in-depth user interaction behaviors in the GenIR system is valuable. There are additional research questions, such as: How do we align personalized models using limited user data? How can we efficiently fine-tune and store personalized generative models?
- **User-centric evaluation of the GenIR system**. For instance, how do we measure user satisfaction when engaging with complex tasks during interactions with the GenIR system? Is personalized evaluation feasible and essential?
- **Privacy protection within the GenIR system**. Particularly, we need to consider how to ensure privacy is maintained when utilizing user feedback in personalized generative models.

References

1. Liu, Y., Wang, Y., Sun, L., Yu, P.S.: Rec-gpt4v: Multimodal recommendation with large vision-language models. arXiv preprint arXiv:2402.08670 (2024)
2. Dai, S., Shao, N., Zhao, H., Yu, W., Si, Z., Xu, C., Sun, Z., Zhang, X., Xu, J.: Uncovering ChatGPTâs capabilities in recommender systems. In: Proceedings of the 17th ACM Conference on Recommender Systems, pp. 1126–1132 (2023)
3. Liu, J., Liu, C., Lv, R., Zhou, K., Zhang, Y.: Is ChatGPT a good recommender? A preliminary study. arXiv preprint arXiv:2304.10149 (2023)
4. Wang, L., Lim, E.-P.: Zero-shot next-item recommendation using large pretrained language models. arXiv preprint arXiv:2304.03153 (2023)
5. Geng, S., Liu, S., Fu, Z., Ge, Y., Zhang, Y.: Recommendation as language processing (RLP): A unified pretrain, personalized prompt & predict paradigm (p5). In: Proceedings of the 16th ACM Conference on Recommender Systems, pp. 299–315 (2022)
6. Hou, Y., Zhang, J., Lin, Z., Lu, H., Xie, R., McAuley, J., Zhao, W.X.: Large language models are zero-shot rankers for recommender systems. In: European Conference on Information Retrieval
7. Rajput, S., Mehta, N., Singh, A., Hulikal Keshavan, R., Vu, T., Heldt, L., Hong, L., Tay, Y., Tran, V., Samost, J., et al.: Recommender systems with generative retrieval. In: Advances in Neural Information Processing Systems 36 (2024)
8. Zhai, J., Zheng, X., Wang, C.-D., Li, H., Tian, Y.: Knowledge prompt-tuning for sequential recommendation. In: Proceedings of the 31st ACM International Conference on Multimedia, pp. 6451–6461 (2023)
9. Liao, J., Li, S., Yang, Z., Wu, J., Yuan, Y.,Wang, X., He, X.: Llara: Aligning large language models with sequential recommenders. arXiv preprint arXiv:2312.02445 (2023)
10. Luo, S., Yao, Y., He, B., Huang, Y., Zhou, A., Zhang, X., Xiao, Y., Zhan, M., Song, L.: Integrating large language models into recommendation via mutual augmentation and adaptive aggregation. arXiv preprint arXiv:2401.13870 (2024)
11. Petrov, A.V., Macdonald, C.: Generative sequential recommendation with GPTRec. arXiv preprint arXiv:2306.11114 (2023)
12. Zhang, J., Xie, R., Hou, Y., Zhao, W.X., Lin, L., Wen, J.-R.: Recommendation as instruction following: A large language model empowered recommendation approach. arXiv preprint arXiv:2305.07001 (2023)
13. Zhang, A., Sheng, L., Chen, Y., Li, H., Deng, Y., Wang, X., Chua, T.-S.: On generative agents in recommendation. arXiv preprint arXiv:2310.10108 (2023)
14. Huang, X., Lian, J., Lei, Y., Yao, J., Lian, D., Xie, X.: Recommender ai agent: Integrating large language models for interactive recommendations. arXiv preprint arXiv:2308.16505 (2023)
15. Shu, Y., Gu, H., Zhang, P., Zhang, H., Lu, T., Li, D., Gu, N.: Rah! RecSys-assistant-human: A human-central recommendation framework with large language models. arXiv preprint arXiv:2308.09904 (2023)
16. Wang, L., Zhang, J., Chen, X., Lin, Y., Song, R., Zhao, W.X., Wen, J.-R.: RecAgent: A novel simulation paradigm for recommender systems. arXiv preprint arXiv:2306.02552 (2023)
17. Wang, Y., Jiang, Z., Chen, Z., Yang, F., Zhou, Y., Cho, E., Fan, X., Huang, X., Lu, Y., Yang, Y.: RecMind: Large language model powered agent for recommendation. arXiv preprint arXiv:2308.14296 (2023)
18. OpenAI, :, Achiam, J., Adler, S., Agarwal, S., et al. GPT-4 Technical Report (2023)
19. Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S., Amodei, D.: Deep reinforcement learning from human preferences. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Red Hook (2017)
20. Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., Kravec, S., Lovitt, L.,

Nanda, N., Olsson, C., Amodei, D., Brown, T., Clark, J., McCandlish, S., Olah, C., Mann, B., Kaplan, J.: Fine-tuning language models from human preferences (2022)
21. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Burges, C.J., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems, vol. 26. Curran Associates, Red Hook (2013)
22. Mikolov, T., Kombrink, S., Burget, L., Černocký, J., Khudanpur, S.: Extensions of recurrent neural network language model. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5528–5531 (2011). https://doi.org/10.1109/ICASSP.2011.5947611
23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Red Hook (2017)
24. Peters, M.E., Ammar, W., Bhagavatula, C., Power, R.: Semi-supervised sequence tagging with bidirectional language models (2017)
25. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2019)
26. Vincent, J.: Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day
27. Silva, C.: It took just one weekend for Meta's new AI chatbot to become racist. Accessed 2024-02-19
28. Zhang, Z., Lei, L., Wu, L., Sun, R., Huang, Y., Long, C., Liu, X., Lei, X., Tang, J., Huang, M.: SafetyBench: Evaluating the Safety of Large Language Models with Multiple Choice Questions (2023)
29. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P.F., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems, vol. 35, pp. 27730–27744. Curran Associates, Red Hook (2022)
30. Spirin, N., Han, J.: Survey on web spam detection: principles and algorithms. SIGKDD Explor. Newsl. 13(2), 50–64 (2012) https://doi.org/10.1145/2207243.2207252
31. Chirita, P.-A., Diederich, J., Nejdl, W.: MailRank: using ranking for spam detection. In: Proceedings of the 14th ACM International Conference on Information and Knowledge Management. CIKM '05, pp. 373–380. Association for Computing Machinery, New York, NY (2005). https:/doi.org/10.1145/1099554.1099671
32. Wolf, Y., Wies, N., Avnery, O., Levine, Y., Shashua, A.: Fundamental Limitations of Alignment in Large Language Models (2024)
33. Cheng, Z., Gao, B., Liu, T.-Y.: Actively predicting diverse search intent from user browsing behaviors. In: Proceedings of the 19th International Conference on World WideWeb.WWW'10, pp. 221–230. Association for Computing Machinery, New York, NY (2010). https://doi.org/10.1145/1772690.1772714
34. Ashkan, A., Clarke, C.L.A., Agichtein, E., Guo, Q.: Classifying and characterizing query intent. In: Boughanem, M., Berrut, C., Mothe, J., Soule-Dupuy, C. (eds.) Advances in Information Retrieval, pp. 578–586. Springer, Berlin (2009)
35. Su, N., He, J., Liu, Y., Zhang, M., Ma, S.: User intent, behaviour, and perceived satisfaction in product search. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. WSDM '18, pp. 547–555. Association for Computing Machinery, New York, NY (2018). https://doi.org/10.1145/3159652.3159714
36. Trielli, D., Diakopoulos, N.: Partisan search behavior and google results in the 2018 U.S. midterm elections. Inf. Commun. Soc. **25**(1), 145–161 (2022) https://doi.org/10.1080/1369118X.2020.1764605

37. Epstein, R., Robertson, R.E.: The search engine manipulation effect (seme) and its possible impact on the outcomes of elections. Proc. Natl. Acad. Sci. **112**(33), 4512–4521 (2015) https://doi.org/10.1073/pnas.1419828112
38. Teevan, J., Dumais, S.T., Horvitz, E.: Potential for personalization. ACM Trans. Comput.-Hum. Interact. 17(1), 1–31 (2010) https://doi.org/10.1145/1721831.1721835
39. Sieg, A., Mobasher, B., Burke, R.: Web search personalization with ontological user profiles. In: Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management. CIKM '07, pp. 525–534. Association for Computing Machinery, New York, NY (2007). https://doi.org/10.1145/1321440.1321515
40. Bennett, P.N., White, R.W., Chu, W., Dumais, S.T., Bailey, P., Borisyuk, F., Cui, X.: Modeling the impact of short- and long-term behavior on search personalization. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '12, pp. 185–194. Association for Computing Machinery, New York, NY (2012). https://doi.org/10.1145/2348283.2348312
41. Dong, Q., Liu, Y., Ai, Q., Wu, Z., Li, H., Liu, Y., Wang, S., Yin, D., Ma, S.: Aligning the capabilities of large language models with the context of information retrieval via contrastive feedback. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR) (2024)
42. Yoon, C., Kim, G., Jeon, B., Kim, S., Jo, Y., Kang, J.: Ask Optimal Questions: Aligning Large Language Models with Retriever's Preference in Conversational Search (2024)
43. Liu, T.-Y.: Learning to rank for information retrieval. Foundat. Trends® Inf. Retriev. **3**(3), 225–331 (2009) https://doi.org/10.1561/1500000016
44. Ai, Q., Wang, X., Bruch, S., Golbandi, N., Bendersky, M., Najork, M.: Learning groupwise multivariate scoring functions using deep neural networks. In: Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval. ICTIR '19, pp. 85–92. Association for Computing Machinery, New York, NY (2019). https://doi.org/10.1145/3341981.3344218
45. Ziegler, D.M., Stiennon, N.,Wu, J., Brown, T.B., Radford, A., Amodei, D., Christiano, P., Irving, G.: Fine-Tuning Language Models from Human Preferences (2020)
46. Lee, H., Phatale, S., Mansoor, H., Mesnard, T., Ferret, J., Lu, K., Bishop, C., Hall, E., Carbune, V., Rastogi, A., Prakash, S.: RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback (2023)
47. Yang, K., Klein, D., Celikyilmaz, A., Peng, N., Tian, Y.: RLCD: Reinforcement Learning from Contrast Distillation for Language Model Alignment (2023)
48. Stiennon, N., Ouyang, L., Wu, J., Ziegler, D.M., Lowe, R., Voss, C., Radford, A., Amodei, D., Christiano, P.: Learning to summarize from human feedback (2022)
49. Köpf, A., Kilcher, Y., Rütte, D., Anagnostidis, S., Tam, Z.R., Stevens, K., Barhoum, A., Nguyen, D., Stanley, O., Nagyfi, R., ES, S., Suri, S., Glushkov, D., Dantuluri, A., Maguire, A., Schuhmann, C., Nguyen, H., Mattick, A.: OpenAssistant conversations—democratizing large language model alignment. In: Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems, vol. 36, pp. 47669–47681. Curran Associates, Red Hook (2023)
50. Sun, W., Yan, L., Ma, X., Wang, S., Ren, P., Chen, Z., Yin, D., Ren, Z.: Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agents (2023)
51. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal Policy Optimization Algorithms (2017)
52. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. In: Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems, vol. 36, pp. 53728–53741. Curran Associates, Red Hook (2023)
53. Yuan, Z., Yuan, H., Tan, C., Wang, W., Huang, S., Huang, F.: RRHF: Rank Responses to Align Language Models with Human Feedback without tears (2023)

54. Dong, H., Xiong, W., Goyal, D., Zhang, Y., Chow, W., Pan, R., Diao, S., Zhang, J., Shum, K., Zhang, T.: RAFT: Reward rAnked FineTuning for Generative Foundation Model Alignment (2023)
55. Schulman, J., Levine, S., Moritz, P., Jordan, M.I., Abbeel, P.: Trust Region Policy Optimization (2017)
56. Chu, Z., Ai, Q., Tu, Y., Li, H., Liu, Y.: PRE: A Peer Review Based Large Language Model Evaluator (2024)
57. Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., Hullender, G.: Learning to rank using gradient descent. In: Proceedings of the 22nd International Conference on Machine Learning, pp. 89–96 (2005)
58. Ai, Q., Bi, K., Guo, J., Croft, W.B.: Learning a deep listwise context model for ranking refinement. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. SIGIR '18, pp. 135–144. Association for Computing Machinery, New York, NY (2018). https://doi.org/10.1145/3209978.3209985
59. Bruch, S., Wang, X., Bendersky, M., Najork, M.: An analysis of the softmax cross entropy loss for learning-to-rank with binary relevance. In: Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval, pp. 75–78 (2019)
60. Liu, Y., Liu, P., Radev, D., Neubig, G.: Brio: Bringing order to abstractive summarization. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2890–2903 (2022)
61. Chuklin, A., Markov, I., De Rijke, M.: Click Models for Web Search. Springer, Berlin (2022)
62. Zhou, G., Zhu, X., Song, C., Fan, Y., Zhu, H., Ma, X., Yan, Y., Jin, J., Li, H., Gai, K.: Deep interest network for click-through rate prediction. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. KDD '18, pp. 1059–1068. Association for Computing Machinery, New York, NY (2018). https://doi.org/10.1145/3219819.3219823
63. Gu, L.: Ad click-through rate prediction: A survey. In: Database Systems for Advanced Applications. DASFAA 2021 International Workshops: BDQM, GDMA, MLDLDSA, MobiSocial, and MUST, Taipei, April 11–14, 2021, Proceedings 26, pp. 140–153. Springer, Berlin (2021)
64. Dou, Z., Song, R., Yuan, X., Wen, J.-R.: Are click-through data adequate for learning web search rankings? In: Proceedings of the 17th ACM Conference on Information and Knowledge Management. CIKM '08, pp. 73–82. Association for Computing Machinery, New York, NY (2008). https://doi.org/10.1145/1458082.1458095
65. Liu, J., Dolan, P., Pedersen, E.R.: Personalized news recommendation based on click behavior. In: Proceedings of the 15th International Conference on Intelligent User Interfaces. IUI '10, pp. 31–40. Association for Computing Machinery, New York, NY (2010). https://doi.org/10.1145/1719970.1719976
66. Isinkaye, F.O., Folajimi, Y.O., Ojokoh, B.A.: Recommendation systems: Principles, methods and evaluation. Egypt. Inform. J. **16**(3), 261–273 (2015)
67. Qiu, F., Cho, J.: Automatic identification of user interest for personalized search. In: Proceedings of the 15th International Conference on World Wide Web, pp. 727–736 (2006)
68. Ge, S., Dou, Z., Jiang, Z., Nie, J.-Y., Wen, J.-R.: Personalizing search results using hierarchical RNN with query-aware attention. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management. CIKM '18, pp. 347–356. Association for Computing Machinery, New York, NY (2018). https://doi.org/10.1145/3269206.3271728
69. Wu, L., Zheng, Z., Qiu, Z., Wang, H., Gu, H., Shen, T., Qin, C., Zhu, C., Zhu, H., Liu, Q., Xiong, H., Chen, E.: A Survey on Large Language Models for Recommendation (2024). https://arxiv.org/abs/2305.19860
70. Zhou, Y., Zhu, Q., Jin, J., Dou, Z.: Cognitive personalized search integrating large language models with an efficient memory mechanism. arXiv preprint arXiv:2402.10548 (2024)
71. Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., et al.: A survey on large language model based autonomous agents. arXiv preprint arXiv:2308.11432 (2023)

72. Wang, L., Zhang, X., Su, H., Zhu, J.: A comprehensive survey of continual learning: theory, method and application. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**, 5362–5383 (2024)
73. Chapelle, O., Zhang, Y.: A dynamic Bayesian network click model for web search ranking. In: Proceedings of the 18th International Conference on World Wide Web, pp. 1–10 (2009)
74. Dou, Z., Song, R., Yuan, X., Wen, J.-R.: Are click-through data adequate for learning web search rankings? In: Proceedings of the 17th ACM Conference on Information and Knowledge Management, pp. 73–82 (2008)
75. Wu, T., Luo, L., Li, Y.-F., Pan, S., Vu, T.-T., Haffari, G.: Continual Learning for Large Language Models: A Survey (2024). https://arxiv.org/abs/2402.01364
76. Shi, H., Xu, Z., Wang, H., Qin, W., Wang, W., Wang, Y., Wang, H.: Continual learning of large language models: A comprehensive survey. arXiv preprint arXiv:2404.16789 (2024)
77. Mao, K., Dou, Z., Mo, F., Hou, J., Chen, H., Qian, H.: Large language models know your contextual search intent: A prompting framework for conversational search. In: Bouamor, H., Pino, J., Bali, K. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6–10, 2023, pp. 1211–1225. Association for Computational Linguistics, Stroudsburg (2023). https://aclanthology.org/2023.findings-emnlp.86
78. Ye, F., Fang, M., Li, S., Yilmaz, E.: Enhancing conversational search: Large language model-aided informative query rewriting. In: Bouamor, H., Pino, J., Bali, K. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023, pp. 5985–6006. Association for Computational Linguistics, Stroudsburg (2023). https://aclanthology.org/2023.findings-emnlp.398
79. Chen, H., Dou, Z., Mao, K., Liu, J., Zhao, Z.: Generalizing conversational dense retrieval via LLM-cognition data augmentation. arXiv preprint arXiv:2402.07092 (2024)
80. https://www.trecikat.com/ (2023)
81. Li, L., Zhang, Y., Liu, D., Chen, L.: Large language models for generative recommendation: A survey and visionary discussions. arXiv preprint arXiv:2309.01157 (2023)
82. Gao, L., Ma, X., Lin, J., Callan, J.: Precise zero-shot dense retrieval without relevance labels. CoRR **abs/2212.10496** (2022)
83. Yu, W., Iter, D., Wang, S., Xu, Y., Ju, M., Sanyal, S., Zhu, C., Zeng, M., Jiang, M.: Generate rather than retrieve: Large language models are strong context generators. In: 11th International Conference on Learning Representations, ICLR 2023 (2023)
84. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E.H., Le, Q., Zhou, D.: Chain of thought prompting elicits reasoning in large language models. In: Advances in Neural Information Processing Systems (2020)

Chapter 6
Generative Information Retrieval Evaluation

Marwah Alaofi, Negar Arabzadeh, Charles L. A. Clarke, and Mark Sanderson

Abstract In this chapter, we consider generative information retrieval (IR) evaluation from two distinct but interrelated perspectives. First, Large Language Models (LLMs) themselves are rapidly becoming tools for evaluation, with current research indicating that LLMs may be superior to crowdsource workers and other paid assessors on basic relevance judgment tasks. We review past and ongoing related research, including speculation on the future of shared task initiatives, such as the Text Retrieval Conference (TREC), and a discussion on the continuing need for human assessments. Second, we consider the evaluation of emerging LLM-based Generative Information Retrieval (GenIR) systems, including Retrieval-Augmented Generation (RAG) systems. We consider approaches that focus both on the end-to-end evaluation of GenIR systems and on the evaluation of a retrieval component as an element in a RAG system. Going forward, we expect the evaluation of GenIR systems to be at least partially based on LLM-based assessment, creating an apparent circularity, with a system seemingly evaluating its own output. We resolve this apparent circularity in two ways: (1) by viewing LLM-based assessment as a form of "slow search," where a slower IR system is used for evaluation and training of a faster production IR system, and (2) by recognizing the continuing need to ground evaluation in human assessment, even if the characteristics of that human assessment must change.

6.1 Introduction

Both the structure of GenIR systems and the capabilities of LLMs are evolving rapidly. It would appear from an evaluation perspective that GenIR presents both challenges and opportunities both concrete and speculative.

M. Alaofi · M. Sanderson
RMIT University, Melbourne, VIC, Australia

N. Arabzadeh · C.L.A. Clarke (✉)
University of Waterloo, Waterloo, ON, Canada

R. W. White, C. Shah (eds.), *Information Access in the Era of Generative AI*,
The Information Retrieval Series 51, https://doi.org/10.1007/978-3-031-73147-1_6

- **Challenges** stem from evaluating the prosodic form of GenIR output: a written synthesis of answers and, sometimes, hallucinated text replacing the classic search response, a ranking of documents.
- **Opportunities** arise from the prospect of automating components of the methodology to evaluate current document retrieval systems. With the apparent ability of generative methods to simulate human actions, we speculate on a range of potential rapid assessments of the worth of a technology prior to actual user trials.

As with any document written at the start of a revolution, it is too early to say what will come. The functionalities and limitations of GenIR are not yet well understood. In many cases, we can only provide a sketch of ongoing research and emerging opportunities. In general, we err on the side of describing future potential rather than surveying the current state of the art, since the latter has changed significantly even between the time we first wrote these words and this, our final proofreading pass. We examine past work to try to contextualize challenges, opportunities, and speculations in more detail.

We interpret *GenIR evaluation* in two ways: (1) the use of generative methods to aid evaluation practices in IR, such as generating document relevance labels, and (2) evaluating the output of a GenIR system, which is likely employing some form of RAG architecture. Figure 6.1 provides a brief description of this chapter's sections and outlines their subsections. We start the chapter by reflecting on past assumptions and challenges within evaluation practices and explore how LLMs can challenge these assumptions and contribute to the development of better practices (Sect. 6.2). We then address the challenges associated with evaluating the output of GenIR systems (Sect. 6.3). Across both sections, we speculate on possible challenges.

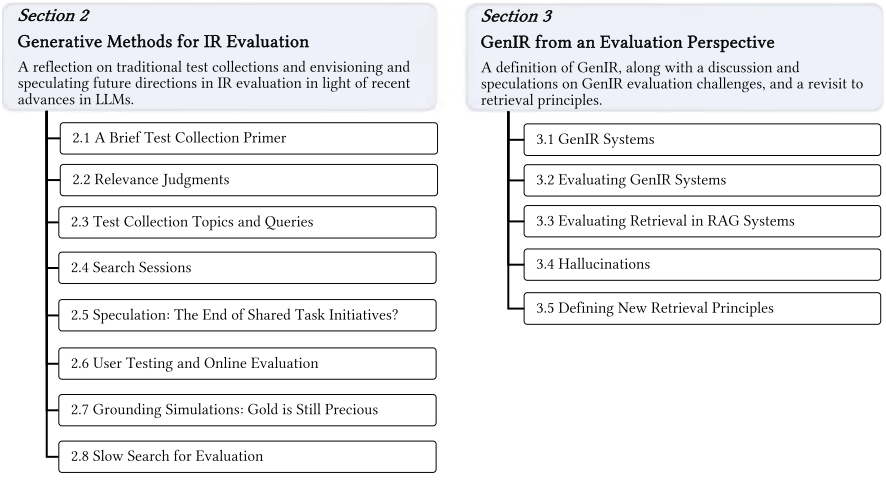

Fig. 6.1 An overview of the main two sections and their subsections

6.2 Generative Methods for IR Evaluation

The arrival of GenIR systems prompts the question of whether the traditional ranking of search results, known as the ten blue links,[1] will be replaced. Ranking, however, remains a common means of search result presentation, which is likely to persist in some form either as an internal component of a RAG system or in applications where the purpose of a search is to identify items as part of seeking information, e.g., mapping applications returning locations in rank order, music applications matching songs, job search engines sorting employment opportunities in order, etc. In this section, we detail the impact of LLMs on the common offline document ranking evaluation methodology, a test collection [65]. We first provide a brief primer to test collections before detailing the impact of LLMs on a number of test collection components and approaches to testing. We then outline the impact of LLMs on the capturing of relevance judgments, on the creation of topics and queries for test collections, and on search sessions, before speculating on the role of shared tasks initiatives in the future. The impact of LLMs on wider user testing is described, before the section concludes with a discussion of the continued role of human labeling in IR evaluation.

6.2.1 A Brief Test Collection Primer

A typical test collection includes a set of search topics (expressed through queries), a corpus of documents, and relevance judgments that record the documents that are relevant to the topics, often referred to as *qrels*. To test an IR system, queries from the collection are run through the system, and its ability to locate relevant documents is measured. Evaluation using a test collection is fully automated, allowing systems to be optimized at a low cost to the experimenter. However, there is a substantial cost involved in creating test collections.

According to Voorhees, offline evaluation practices have mainly operated with the following simplifying assumptions:

- "*Relevance can be approximated by topical similarity, which implies all relevant documents are equally desirable; relevance of one document is independent of the relevance of any other document; and the user information need is static.*
- *a single set of judgments for a topic is representative of the user population.*
- *(essentially) all relevant documents for a topic are known*", Voorhees [80, p. 47]. To this list, we might add that
- there is one representation of an information need.

[1] This phrase seems to emerge around 2007 from the makers of Ask Jeeves—one of the earliest question-answering systems—seeking to contrast their system's written output with what they saw as a traditional search response [38].

Under these assumptions, costs associated with constructing test collections are manageable but still substantial enough to make it nearly impossible for individual academics or research groups to generate such large comprehensive collections by themselves. Consequently, several initiatives were started to share the costs and labor involved in their creation, such as TREC [79], Cross Language Evaluation Forum (CLEF) [59], and National Institute of Informatics Testbeds and Community for Information Access Research (NTCIR) [54]. The challenge of constructing test collections meant that researchers predominantly relied on those produced by these initiatives. The subsequent sections discuss the challenges in more detail and present opportunities for using LLMs to address these challenges.

Early test collections consisted of a few hundred to a few thousand documents.[2] Creating relevance judgments for all documents in such a collection was practically possible. For instance, when creating Cranfield II, Cleverdon employed a team of individuals to manually scan the entire collection to identify all relevant documents for each topic. When building the Library and Information Abstracts (LISA) test collection during the 1980s, one person was employed to search the physical issues of a journal to find relevant documents, which was supplemented with some online search.[3] The scale of test collections was limited by the cost of creating relevance judgments. There was a need among IR researchers to find a way *to produce larger test collections while at the same time locate as many relevant documents as possible* [65, p. 271].

Multiple strategies were explored, though not practically implemented, for creating larger collections, most notably Spärck Jones and Van Rijsbergen [69] introduced document pooling. This technique, which involves sampling documents for relevance assessment through multiple participating searches (now *runs*), subsequently became the conventional method for building IR test collections and is the standard within TREC. Although this approach has its limitations, primarily due to missing some relevant documents [88], which in turn raises concerns about the reusability of collections [17, 81], it has facilitated the expansion of test collections, giving us access to test collections with massive corpora, such as the ClueWeb series, with millions of documents.

6.2.2 Relevance Judgments

Recent studies have demonstrated that LLMs can be used to produce relevance judgments (or labels, distinguishing them from those generated by humans). In May

[2] Readers can refer to the Web site hosted by the University of Glasgow, which archives some of the early test collections, to gain a sense of their modest scale: https://ir.dcs.gla.ac.uk/resources/test_collections/.

[3] See the readme file for further information: https://ir.dcs.gla.ac.uk/resources/test_collections/lisa/.

2023, researchers at Microsoft Bing announced their use of GPT-4 in generating relevance labels, which was later shared in a paper [72]. The LLM generated labels were found to be as accurate as labels created by crowdsourced workers and were being used to train the production system of Bing. Around the same time, Faggioli et al. [35] reported promising results from using LLMs for generating relevance labels. Although these findings have not been extensively tested and may have limitations, they prompt a reevaluation of the need for document pooling, originally adopted to manage costs associated with human labor. That is, it might be now feasible to create complete relevance judgments on a large scale or, at least, create deeper pools as the cost of generating relevance labels has substantially decreased.

The use of LLMs to reduce the cost of relevance judgments echoes a significant historical shift in the value of a material we now take for granted: aluminum. In middle school, American children learn that on the top of the Washington Monument is a relatively small pyramid of solid aluminum. At the time it was placed there, in 1884, aluminum was as rare, and as precious, as silver. The pyramid was the largest piece of solid aluminum in the world. Two years later, Paul Héroult and Charles Hall independently invented a process that would eventually make aluminum cheap enough that when buying an aluminum can of drink, most of the price pays for the contents, not the container. IR is having its Hall-Héroult moment: human judgment was once a rare and precious resource. Now, it appears we can simply ask the large language models (LLM) anything we might ask a human searcher or assessor, but at a much lower cost. This opens many new opportunities for evaluation.

One opportunity is to tailor the definition of relevance to be more specific, including additional dimensions of information utility to different users. Voorhees et al. [81] highlight the score saturation problem in the TREC Deep Learning Track (2021), where many systems are already capable of retrieving ten relevant documents for a wide range of queries from large corpora, calling for *"different metrics or a more focused definition of relevance"* [81]. Relevance can vary across users and contexts, and it is often assessed based on topicality without considering other dimensions, such as understandability. This underscores the need to consider other dimensions of relevance to create test collections that can distinguish among systems. For example, a document might be topically relevant to a query but could exhibit different levels of utility to users based on their domain expertise or operating contexts. It now seems feasible to explore the utility of LLMs to make relevance labels more specific, enabling a detailed and most importantly realistic system evaluation.

Another potential benefit of using LLMs to produce relevance labels is their consistency in the generated labels for documents. Unlike humans, LLMs do not get tired as they generate more labels, nor are they influenced by judgments previously made. There is evidence that there is a great level of inconsistency in human relevance assessment [66, 67], whether due to forgetting earlier decisions, re-calibrating assessments based on the documents already seen, or simply making errors, leading to varying assessments even for almost identical documents [13, 67]. Using a recognized model of LLM with controlled parameters to ensure a deterministic behavior would enable consistency and reproducibility of relevance labels.

While it might be conceivable that the need for collecting these relevance labels and distributing them in test collections could diminish, given that system effectiveness can now be evaluated dynamically and at substantially lower than those incurred using human labor (see [72, Figure 5] for cost-accuracy relative comparison), this approach would undermine the core principle of having test collections serving as static, shared, and reusable resources for system evaluation. See further discussion in Sect. 6.2.5.

6.2.3 Test Collection Topics and Queries

In test collections, the convention is that each information need (search topic) is represented using a *single* query. The queries are generated either by (1) consulting a group of people to generate queries given information need statements or by (2) obtaining a sample from a query log. Going beyond one query to represent a broader spectrum of users employing different query variants was expensive and thought to be unnecessary. However, research suggests that when seeking a common information need, users tend to use a large number of query formulations (often referred to as *query variants*). In studies of user populations, over 50 variants were found per information need [11, 47]. Previous research has demonstrated that factors—such as the used device [26, 42], domain expertise [52, 83], age [14, 75], and language proficiency [25]—influence query formulation. These consequentially impact the quality of search results and overall user satisfaction. Culpepper et al. [32] showed that the impact of query variants on system effectiveness is substantially greater than that due to topic or ranking models. Yet the effect of query variation on IR system effectiveness is often overlooked. Evaluations typically rely on test collections with single queries, leaving the performance of systems for a broader range of users largely unexamined. Given recent studies demonstrating an important role of query variants in system evaluation, how such variants might be generated in a cost-effective manner is a challenge that LLMs may be able to help with.

Unlike with relevance judgments where LLMs have been shown to be a valid substitute for human labels, the work on query variants is more in its infancy. Using artificially created and manually verified query variants, Penha et al. [56] showed a significant drop in the effectiveness of both neural and transformer-based retrieval models. Likewise, Alaofi et al. [3] undertook an empirical investigation into the effects of query variants on a commercial search engine and some inverted indexes. Their research revealed inconsistency in search results across different query variants and shed light on the impact of variants on document retrievability. Similarly, inconsistencies in search results were also demonstrated in the context of searches conducted by children [58].

Crowdsourcing and click graphs have been used to gather query variants. However, both methods have their limitations: crowdsourcing is expensive to scale, and click graphs are noisy and lack information about users. User simulation has been a prevalent instrument in IR, but its application for generating query variants

has not been as extensively explored. For example, Penha et al. [56] proposed a taxonomy for query variants and use multiple techniques to artificially create query variants. More recently, research has shown that LLMs can, to a limited extent, reproduce human query variants, yielding a similar pool of documents of that obtained by using human generated query variants [4]. Engelmann et al. [34] also used LLMs to simulate query variants in an interactive manner, taking into account user sessions and the results seen as feedback for the query generation process. This approach yields more effective search sessions, but does not necessarily reflect how humans engage with search sessions. Another line of research explores using LLMs to generate queries but not for simulations but as a way of generating more query variants to train better rankers (e.g., [15]), generate query expansions (e.g., [48]), and improve document retrievability (e.g., [57]). Giving the ability of LLMs to align its generation to certain properties, an important question arises regarding how effectively they can align with how humans engage with information seeking tasks, reflecting the diverse user properties identified in the literature as influencing query formulation.

6.2.4 Search Sessions

There has long been a recognition that there is more to evaluation than the initial query that establishes a search. Many attempts [8] to extend offline evaluation to include sessions have been tried [20], but as with most efforts to "shift the dial" of offline evaluation, those efforts have not been successful in starting a new standard.

Many of the reasons underpinning the lack of movement in the design of offline evaluation has been a question of cost. The current approach to evaluation while expensive to set up is cheap to use when built. Most approaches to extending the evaluation of search have been more expensive to create or require higher ongoing costs to use. The arrival of generative methods and the ability of generative systems to apparently simulate human behavior to a convincing degree suggests a shifting of the dial. This has already been demonstrated with relevance assessments, but it may also be possible to have viable simulations of interactive sessions with a search engine including effective simulations of document selection and query reformulation, as well as simulations that determine when a search would stop seeking more documents.

6.2.5 Speculation: The End of Shared Task Initiatives?

The arrival of generative systems has the potential to completely redefine how evaluation is conducted in the field of information access. Much of this chapter has focused on existing innovations and future speculations on what might be possible using generative methods. It is worth asking if generative systems may also alter

the way researchers behave. For decades, the field of IR has been characterized by the creation and sharing of large resources that can be used for evaluation. Key among these resources is the test collection. Initially something that was just created by one group and shared with others, this evolved into large-scale shared evaluation tasks starting with TREC in the early 1990s. The tasks were formed in order to share the work required to build large evaluation resources. However, if it is possible to construct evaluation resources individually through the use of generative methods, one might question if these large-scale collaborative evaluation exercises will continue. The costs to researchers of building bespoke datasets with human generated labels has come down substantially, thanks to the rise of crowdsourcing services. Consequently, participation rates at exercises around the world have dropped substantially in recent years. The rise of generative methods simulating the behavior of users and data labelling may be the final nail in the coffin of these long-standing mainstays of our research ecosystem.

6.2.6 User Testing and Online Evaluation

Traditional IR systems returned just a ranked list of documents (see, e.g., Harman's review of pre-Web systems [41]). Over time, the sophistication of ranked output grew. The way ranked documents were displayed depended on the text of the query, thanks to snippets [73], a summary composed of query-focused content extracted from the body of the document [27]. Commercial search engines further augmented the output with direct answers [84], quick links [21, 39], entity cards [16, 53], query suggestions [19], and other components [55]. The sophistication of the output prompted work on so-called whole-page relevance [10], but in the academic community, this approach was not widely adopted, most likely due to the costs of using it.

In the speculations detailed so far in this chapter, the main focus has been on the way that offline evaluation is being redefined through the use of generative methods to label documents as relevant and to generate queries arising from an information need. However, there may be the potential for such replacements to expand into other aspects of evaluation. Hämäläinen et al. [40] detailed how LLMs could be used to simulate many qualitative human responses to the use of and the reactions to systems employed in usability experiments finding that LLMs can "*yield believable accounts of HCI experiences*." It may be possible to revisit whole-page relevance evaluation using generative methods.

6.2.7 Grounding Simulations: Gold Is Still Precious

Evaluation outcomes of systems using test collections reflect "anticipated" real-world performance. Although these test collections appear concrete, featuring

human queries and relevance judgments, they are fundamentally abstract and considerably simplified *simulations* of real-world search scenarios. Use of so-called offline evaluation imagines a simplified process of a searcher browsing the sorted list, top to bottom, identifying relevant pages, and at some point stopping [5, 50, 51, 68, 82]. The extent to which evaluation outcomes reflect actual user satisfaction is crucial; yet it has not received much attention within the community. It was not until 2006 that Turpin and Scholer [76] demonstrated that test collections may poorly reflect reality. This was further investigated by Al-Maskari and Sanderson [2].

The use of LLMs to simulate users in creating test collections raises questions about the validity of this simulation and necessitates further exploration of how well LLMs are aligned with real users. Before going further in simulation and drawing conclusions about how well systems perform, we need to first substantiate the validity of our user simulations. This requires datasets, tools, metrics, and procedures.

User relevance judgments and queries are abundantly available through numerous iterations of shared tasks. Consequently, the approximation of queries and relevance labels to human-generated ones can be examined. However, if personalized relevance labels are to be simulated, for example, taking into account other dimensions of information utility, then we have almost no way to validate their performance since such ground truth data is not widely available. For instance, in a context where we would like to evaluate how well a system performs in response to an expert user as opposed to a non-expert, such data is not readily available. Similarly, when simulating query variants issued by multiple users, very few sources of data are available for validation, and demographic data is often missing. Real human data that fits the definition of gold [9], where both the query and relevance assessments are produced by a diverse set of humans operating in different contexts and demographic data is collected, are highly needed in order to facilitate the research of simulation validation.

In terms of measuring the accuracy of simulations, that is how closely the LLM aligns with human searchers, one can consider if the simulated data exhibits similar properties to human-generated data or leads to comparable conclusions [12], as exact matches may not be feasible in tasks involving language, where queries can be formulated in various ways. Statistical properties of queries, such as length and complexity, can serve as indicators. Other metrics may assess the impact of simulated data compared to human-generated data. For instance, do generated queries demonstrate similar effectiveness to human queries and/or produce similar pools of documents? Do relevance labels result in the same system rankings as if those produced by humans are used?

6.2.8 Slow Search for Evaluation

In 2023, researchers proposed replacing human relevance assessments with LLM assessments [35, 72]. A common objection to these proposals recognizes their

circularity. Using automated methods to assess other automated methods is not without its dangers, if, as is common in this chapter, one looks at historical precedence for current events. One could look at the way in which automated relevance assessments were attempted earlier in the history of IR. A classic example is pseudo relevance feedback [31]. This is a technique that assumes a query from a user will be sufficiently accurate that one can make the assumption that top ranked documents returned by that initial query are themselves likely to be relevant. The text of those documents can then be used in an internal reformulation of the query to produce better results. While rarely seen in commercial systems, pseudo relevance feedback is a well-known technique.

If LLM assessment is sufficient to replace a human assessment, then why not treat the LLM as a ranker, ranking items according to their LLM assessed relevance? If an LLM-based evaluation is generating the labels for evaluation, ranking by those labels always produces an ideal result.

One way to avoid this circularity is to consider the difference in time and resources needed by a production GenIR system vs. the time and resources required for LLM-based evaluation. For evaluation purposes, we can take all the time we need to find the best response and then use that response to evaluate the efficiency vs. effectiveness trade-off between, for example, a production system that responds in 100 ms and one that responds in 500 ms. From the standpoint of an efficiency vs. effectiveness trade-off, for the purposes of evaluation, we can essentially ignore efficiency.

The trade-off between retrieval efficiency and effectiveness has long been a subject of academic research [6, 7, 18, 28, 87] and a key consideration for commercial search engines, which aim for an average query latency in hundreds of milliseconds [7, 49]. However, in the past, we have had relatively few methods for tuning the trade-off between efficiency and effectiveness beyond a narrow range. Efficiency vs. effectiveness trade-offs might be measured in terms of tiny percentages of effectiveness improvements at the cost of milliseconds of query latency, but we could never improve effectiveness enough to justify a latency of seconds or longer

Teevan et al. [71] in advocating for "Slow Search" write, "*With even just a little extra time to invest, search engines can relax existing restrictions to improve search result quality. For example, complex query processing can be done to identify key concepts in the query, and multiple queries derived from the initial query can be issued to broaden the set of candidate documents to cover different aspects of the query.*"

Unfortunately, it was never fully demonstrated that investing more time would ever achieve these goals. We had no way to operationalize the proposal of Teevan et al. [71]. If a search engine is fast, the searcher can quickly see if the results are not relevant and immediately reformulate their query [62]. If a query is missing a key concept, the searcher can add it. Low latency is an important feature of search engines, since it facilitates rapid interaction. We can only justify higher latency if rapid interaction is not required.

We have now entered an era where deriving multiple queries and other complex query processing might genuinely improve the results in more than a trivial way. With more time, our GenIR system might prompt an LLM to make relevant judgments, determine what aspects of a document make it relevant, and automatically refine queries in light of these determinations. A GenIR system might compare one item against another, until it identifies the best overall result. In some cases, it might be worth the time of the searcher to wait for this result, but if not, it can still be used to evaluate the faster result actually returned to the searcher.

In some sense, evaluation has always been slow search with a human-in-the-loop. In a traditional TREC ad hoc task, we build a pool, and humans assess items in the pool, creating an ideal response. Now we can use an LLM to replace these humans. However, unless we determine that taking all the time we need always produces the best possible response, we still need a way to evaluate the results of slow search. If the quality of LLM assessment can reach the level of traditional human assessment, do we consider this as our peak achievement? Or do we recognize that there is still room for improvement by involving humans to perhaps monitor LLMs or revisit our ideal definition of relevance?

6.3 Generative Information Retrieval from an Evaluation Perspective

In the previous section, we considered the use of generative methods to aid evaluation practices in current IR systems, in particular for generating document relevance labels. In this section, we consider the evaluation of emerging IR systems that may not adhere to conventional assumptions about ranking and result presentation.

6.3.1 Generative Information Retrieval Systems

Current and potential capabilities of GenIR systems were engendered by the increasing capabilities of LLMs, especially their ability to conduct zero-shot natural language tasks, including summarization, query understanding, and query expansion. Most GenIR systems replace the query and ranked list with a conversation and a written synthesis of information, similar to that shown in Fig. 6.2. At the time of writing, these systems include Perplexity[4] and newer versions of Bing.[5] The TREC 2024 RAG Track, which supersedes the Deep Learning Track, also assumes this interface format.[6] The searcher poses a question in a potentially longer,

[4] perplexity.ai.

[5] bing.com.

[6] trec-rag.github.io.

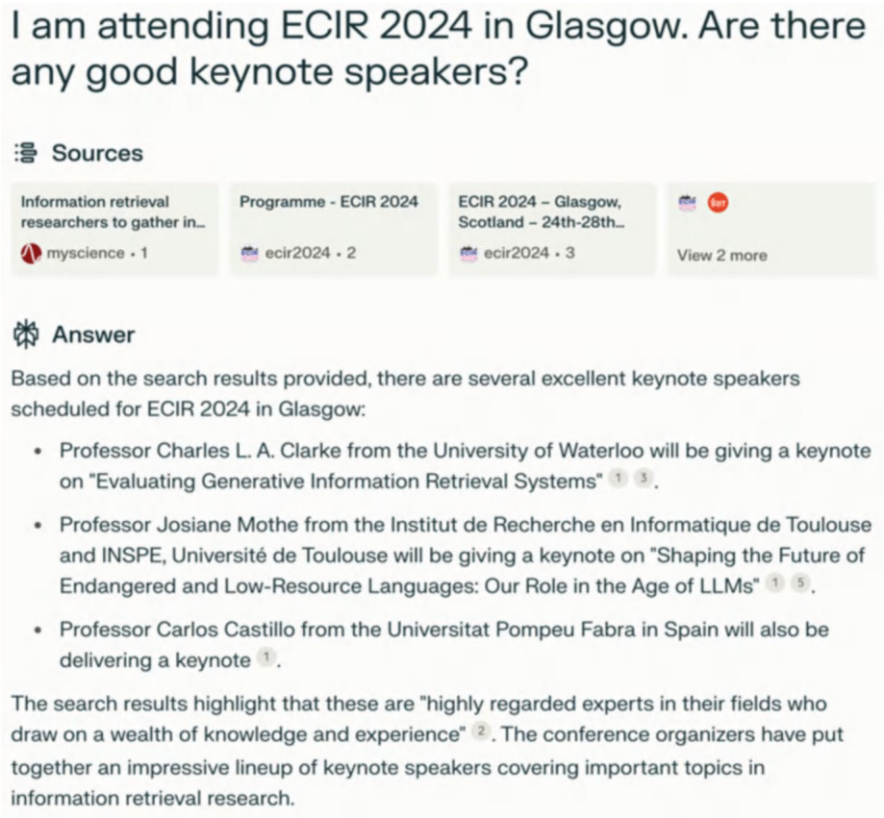

Fig. 6.2 A GenIR user interface from March 2024 (from `perplexity.ai`)

more natural, and conversational form. The system responds with a single coherent answer, which may be supported by links to sources. Gienapp et al. [37] view a GenIR system as a "synthetical" search engine that searches for sources, "compiles them, synthesizes missing information, presents it coherently, and grounds its claims in the retrieved sources." The system provides searchers with a single unified answer "that covers a complex topic with in-depth analysis from varied perspectives" (Fig. 6.3). Such interactions and outputs will require us to seek a new evaluation model.

For evaluation purposes, we need not make any assumptions about the internal architecture of a GenIR system, which may simply be a single large neural model. In this case, our evaluation must focus on the end-to-end interaction. A query or question is entered by the searcher, and the system responds with an answer, which may reflect a larger conversational context, including personalization. Under this view, our core search metric becomes the following: How good is this overall response? In traditional IR evaluation, the focus was often on the output of a ranker. While whole-page relevance was a factor in evaluation, especially in industry

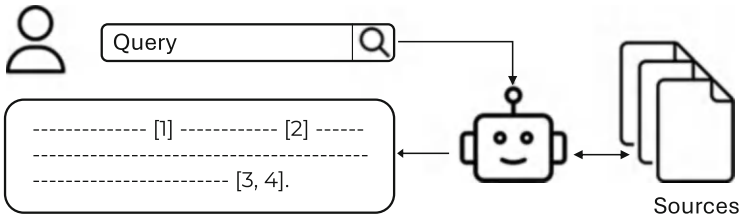

Fig. 6.3 A GenIR system as a synthetical search engine

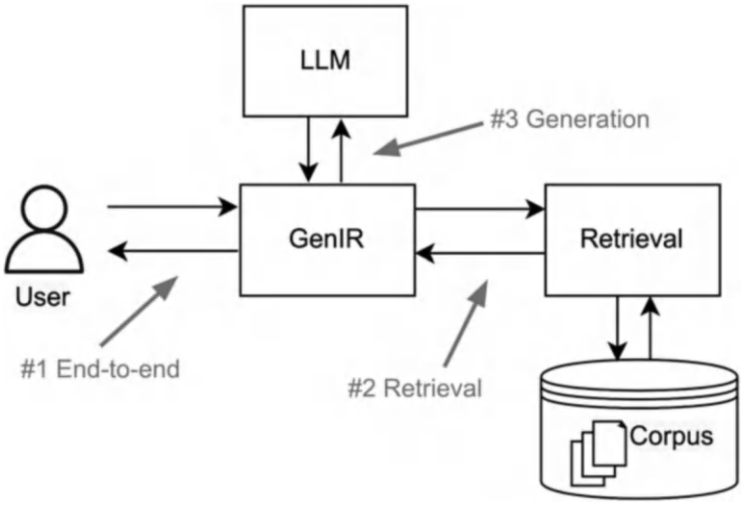

Fig. 6.4 RAG Architecture overview

contexts [10], it was one of many factors. If we view a GenIR system as a black box, whole-page relevance becomes a central factor.

While we can view a GenIR system as a black box for evaluation purposes, a RAG architecture [36] often underlies a GenIR system. In Fig. 6.4, we have simplified the architecture of a RAG system to its key components. At the front end, a searcher interacts with a *generative component*, which in turn interacts with both an *LLM* and a *retrieval component*. The retrieval component is used to search a corpus, which is assumed as a source of ground truth—although, like any IR system, the corpus itself may contain spam and documents of varying quality. Information provided by the RAG system to the searcher requires support from the corpus. The generative component interacts with the LLM for purposes of query understanding, query expansion, summarization, and similar tasks, while it interacts with the retrieval component through keyword or other queries to find sources for its response. The system may interact with both the LLM and the retrieval component multiple times before responding to a user's query, where the overall approach may

be retrieve-then-generate, generate-then-retrieve [1], or a combination of multiple generation and retrieval steps.

A competing definition of "Generative Information Retrieval" systems describes a GenIR system as one that does not generate the answer to the searchers query; instead, it replaces a traditional search engine by using a neural model to directly generate the identifiers of documents that answer the query [23, 24, 60, 70, 85]. While these systems are "generative" in the sense that a neural model is directly generating document identifiers, from an evaluation perspective, they are no different from any other retrieval component that returns document identifiers, except in one respect. Since the document identifiers are generated, it is conceivable for such systems to "hallucinate" document identifiers that do not exist. Nonetheless, for the purpose of our discussion, we view them as a type of retrieval component.

6.3.2 Evaluating Generative Information Retrieval Systems

As shown in Fig. 6.4, a RAG system may be evaluated at three points: (1) at the front end, where we are evaluating the end-to-end performance of the system; (2) at the top of the retrieval component, where we are evaluating the retrieval component in the context of the overall GenIR system; and (3) at the point of interaction with the LLM. In the context of the overall system, the retrieval component (#2) is essentially a subordinate system returning a ranked list of items for the generative component. While a human searcher may eventually be given links to items in the corpus, these will be selected by the generative component. Evaluating interactions with the LLM (#3) falls slightly outside our scope into the broader topic of NLP evaluation, including the evaluation of summarization and information extraction.

Evaluation of an end-to-end GenIR system (#1) introduces challenges beyond those of traditional IR evaluation. Gienapp et al. [37] argue that the key difference between a traditional search engine and a GenIR system is that the GenIR system is essentially searching an infinite corpus of all possible responses that could be synthesized by the system [33]. Traditional IR test collections, such as those created by the TREC, try to be reusable, with a nearly complete set of relevance judgments. With a finite corpus, this approach is conceptually possible; with an infinite corpus, it is not.

One approach to evaluating the retrieval component (#2) would be to evaluate it as a traditional search engine. Its role is to execute a query over a corpus of items and return a ranked list of those items. To evaluate the retrieval component of a RAG system, we may be able to adapt existing offline evaluation methods. Even if the interface seen by the searcher is no longer "ten blue links," internally, we can imagine a similar interface between the generative component and the retrieval component, although the browsing models assumed by offline evaluation metrics no longer apply. These browsing models often assume that the searcher has limited patience [51] or that the searcher will stop scanning the ranked list after a relevant

item is found [22]. A generative component might be assumed to dig deeper into the ranked list and seek information from more sources.

6.3.3 Evaluating Retrieval in RAG Systems

RAG systems include a retrieval component (Fig. 6.4), which supports retrieval over a corpus that provides ground truth for our GenIR system. For evaluation purposes, we might treat the retrieval component as an old-fashioned search engine, even if it itself includes generative components. A query goes into the retrieval component, and a ranked list comes out. However, since this response is entirely internal to the GenIR system, it need not only be a ranked list. It could be richer and more complex. The output of the retrieval component must be tailored to the needs of the overall system, and not to the needs of a human searcher.

If we view the retrieval component as an old-fashioned search engine, returning a ranked list, we might employ traditional evaluation methods. If we think about the GenIR system as internally browsing down the output of the retrieval component, we could use NDCG@10 as our metric. However, the GenIR has more "patience" than a human searcher, so the Normalized Discounted Cumulative Gain (NDCG) discount function might not be the right one to use.

The purpose of the retrieval component is to return the items that the overall GenIR system needs to craft its response. Traditional ranking stacks often use a BM25-based first stage that returns a large collection of items, maybe 1000, for re-ranking by a second-stage ranker [87]. The output of this second stage is then filtered, re-ranked, and processed by more stages until a final stage produces a ranked list that can be shown to the searcher. A typical metric for the first stage is recall@1000. Perhaps recall might be a better metric for evaluating the retrieval component, since the overall GenIR system essentially acts as the upper stages.

6.3.4 Hallucinations

Even when supported by a retrieval component, GenIR systems might generate factually inaccurate or misleading responses. In traditional IR evaluation, we assume that the corpus is curated and can be trusted. If we cannot trust it, then we filter it for spam and other misinformation. While in traditional Web search some pages are higher quality than others, the output of the search engine is a list of pages, which the searcher can ultimately inspect for themselves. They are not depending on the search engine to summarize the information for them.

Since GenIR systems can hallucinate [74], it is not sufficient to filter the corpus for spam and misinformation. We must also evaluate the accuracy of the end-to-end response. The final generated response can be false or contain falsehoods, even if the retrieved material is true. Fact-checking must become a standard component of GenIR evaluation.

The situation has already happened[7] "in the wild." A chatbot on the Air Canada Web site incorrectly advised a customer, Jake Moffatt, that he could receive a reduced bereavement rate by submitting a claim within 90 days of ticket issue. The response from the chatbot included a link to a static page on the company's Web site that provided the correct information, indicating that the claim had to be submitted in advance of ticket issue. Air Canada refused the Moffatt's claim. Moffatt took the matter to the Civil Resolution Tribunal of the province of British Columbia who allowed the claim, writing:

> Air Canada argues it cannot be held liable for information provided by one of its agents, servants, or representatives—including a chatbot. It does not explain why it believes that is the case. In effect, Air Canada suggests the chatbot is a separate legal entity that is responsible for its own actions. This is a remarkable submission. While a chatbot has an interactive component, it is still just a part of Air Canada's website. It should be obvious to Air Canada that it is responsible for all the information on its website. It makes no difference whether the information comes from a static page or a chatbot.
> I find Air Canada did not take reasonable care to ensure its chatbot was accurate. While Air Canada argues Mr. Moffatt could find the correct information on another part of its website, it does not explain why the webpage titled "Bereavement travel" was inherently more trustworthy than its chatbot. It also does not explain why customers should have to double-check information found in one part of its website on another part of its website.

While technical details of the chatbot are not available, we can view it as a RAG system since it returned both a generated answer and a link intended to support the answer. While this is a minor matter from a legal standpoint, it demonstrates that a RAG system can generate materially false information, even when supported by retrieved information that is correct. Extracted Web page summaries have long been a feature of Web search results [27]. While extracted summaries may not always provide the information the searcher requires, they generally provide an accurate quote from the page or its metadata.

The accuracy of a traditional search engine depends on the accuracy of the information in its corpus. The search engine may not be able to find relevant information, but when it does, it does not alter or interfere with it. If the corpus contains misinformation, we attempt to filter it. For evaluation purposes, we measure the quality of the filter. Since a GenIR system can hallucinate misinformation, we must now evaluate the accuracy of its output, along with relevance and other traditional considerations.

[7] https://www.canlii.org/en/bc/bccrt/doc/2024/2024bccrt149/2024bccrt149.html.

6.3.5 *Defining New Retrieval Principles*

Such is the ubiquity of documents in retrieval system design and evaluation, many of the fields key principles are grounded in documents. We briefly detail three of the best-known retrieval: Robertson's Probability Ranking Principle (PRP) [63], Jardine and van Rijsbergen [44]'s Cluster Hypothesis [44], and Craswell et al. [30]'s Cascade Model [30].

Robertson's PRP is widely viewed as a fundamental goal of ranking in IR. It is most commonly expressed as: "*If an IR system's response to each query is a ranking of the documents in the collection in order of decreasing probability of relevance, then the overall effectiveness of the system to its users will be maximized.*" The notion of an ideal ranking, which is built into traditional evaluation metrics such as NDCG [45], depends on the PRP that the best result is to order items according to their probability of relevance.

The cluster hypothesis was defined twice, first as '*It is intuitively plausible that the associations between documents convey information about the relevance of documents to requests.*.' Later, van Rijsbergen [77, Chapter 3] simplified the hypothesis as "*closely associated documents tend to be relevant to the same requests*." The hypothesis inspired many later approaches to the clustering of documents [43, 78] as well as result diversification [46].

Seeking a simplified model of user behavior, Craswell et al. [30] examined large user interaction logs in an attempt to capture a broad form of behavior of user interaction. They produced the cascade model: "*where users view results from top to bottom and leave as soon as they see a worthwhile document*." This model has underpinned a great many modern evaluation measures and also inspired many subsequent studies developing extensions to this model.

All three ideas assume the fundamental unit in retrieval is the document. In the case of GenIR, the entirety of the system's end-to-end response should be relevant, and nothing should be redundant; the boundaries between documents hold far less importance. Everything in the response should be there for a reason, and in many cases, the response should include more than just the bare answer. The response might link to background articles that support the response. It might provide opposing perspectives. It might suggest cheaper or higher-quality alternatives to a product. It might synthesize similar responses from multiple sources into a single sentence. It might ask for clarification or disambiguation.

We might ask what replaces these principles in a GenIR system. One idea is provided by the work of Rajput et al. [61]. They propose *nuggets* as a basis for evaluation, where we might think of nuggets as an atomic unit of relevance, e.g., some fact, relationship, or concept that a perfectly relevant document would contain [29, 64]. Rajput et al. [61] propose to build a reusable test collection in a two-phase process. In the first phase, human assessors would identify and extract nuggets from relevant documents. In the second phase, these nuggets would be automatically matched against unjudged documents to measure relevance, providing

a reusable test collection that does not depend on a fixed corpus with relevance labels for individual items.

While they provide experimental support demonstrating both the feasibility and benefits of this approach, it was not widely adopted for either academic or industry assessment. Possible reasons include the need for reliable and trained assessors to identify nuggets, as well as the need to automatically match the nuggets against documents. In 2012, they could only suggest a surface-level, lexical approach to matching, and of course, humans are expensive. Crowdsourcing might reduce the cost but might increase noise and decrease reliability.

In 2024, an LLM might be expected to reliably and cheaply extract nuggets and match them against documents. All it takes is a few calls to an API, costing fractions of a cent per call. It is now almost trivial to realize the vision of Rajput et al. [61], and this proposal is just one of many such proposals in the literature. All the proposals for IR evaluation in terms of diversity, novelty, fairness, completeness, conciseness, effort, or whatever are now both cheap and straightforward to implement.

We can already see nugget-based evaluation emerging as a basis for GenIR evaluation. For example, the new TREC 2024 RAG track[8] takes a nugget-based approach. To formulate a general principle, we turn to Zhai et al. [86]. They propose *subtopic evaluation*, which is closely related to nugget-based evaluation. Evaluation with subtopics is "based on *dependent relevance*, instead of *independent relevance*, as has been assumed in most traditional retrieval methods. The subtopic retrieval problem has to do with finding documents that cover as many different subtopics as possible." To extend this idea to GenIR, we might articulate a principle that the system's response should cover as many nuggets or subtopics as possible.

6.4 Conclusions

Evaluation lies at the core of so much of IR research. If there is any aspect that separates this field from others, it is focus on high-quality evaluation of systems. In this chapter, we examined the impact of LLMs on the evaluation of IR both from the perspective of exploiting the models to speed up traditional evaluation methodologies and to consider the more challenging prospect of evaluating a fully generated response following a conversational interaction. There are some clear early wins such as the revelation that LLMs can be used to generate relevance labels; however, as with any technology when it is first introduced, the boundaries of what the technology can achieve—and more importantly what it cannot—are still being drawn. We have attempted to describe what currently sits within those boundaries, what is yet to be known, and what might change in our field.

[8] trec-rag.github.io.

References

1. Abbasiantaeb, Z., Aliannejadi, M.: Generate then Retrieve: Conversational Response Retrieval Using LLMs as Answer and Query Generators (2024)
2. Al-Maskari, A., Sanderson, M.: A review of factors influencing user satisfaction in information retrieval. J. Am. Soc. Inf. Sci. (JASIST) **61**(5), 859–868 (2010). https://doi.org/10.1002/asi.21300
3. Alaofi, M., Gallagher, L., McKay, D., Saling, L. L., Sanderson, M., Scholer, F., Spina, D., White, R.W.: Where do queries come from?. In: Amigó, E., Castells, P., Gonzalo, J., Carterette, B., Culpepper, J.S., Kazai, G. (eds.) The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11–15, 2022, pp. 2850–2862, ACM, New York (2022). https://doi.org/10.1145/3477495.3531711
4. Alaofi, M., Gallagher, L., Sanderson, M., Scholer, F., Thomas, P.: Can generative LLMs create query variants for test collections? an exploratory study. In: Chen, H., Duh, W.E., Huang, H., Kato, M.P., Mothe, J., Poblete, B. (eds.) Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23–27, 2023, pp. 1869–1873. ACM, New York (2023). https://doi.org/10.1145/3539618.3591960
5. Arabzadeh, N., Kmet, O., Carterette, B., Clarke, C.L.A., Hauff, C., Chandar, P.: A is for Adele: an offline evaluation metric for instant search. In: Yoshioka, M., Kiseleva, J., Aliannejadi, M. (eds.) Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2023, Taipei, Taiwan, 23 July 2023, pp. 3–12. ACM, New York (2023). https://doi.org/10.1145/3578337.3605115
6. Arabzadeh, N., Yan, X., Clarke, C.L.A.: Predicting efficiency/effectiveness trade-offs for dense vs. sparse retrieval strategy selection. In: Demartini, G., Zuccon, G., Culpepper, J.S., Huang, Z., Tong, H. (eds.) CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1–5, 2021, pp. 2862–2866. ACM, New York (2021). https://doi.org/10.1145/3459637.3482159
7. Arapakis, I., Bai, X., Cambazoglu, B.B.: Impact of response latency on user behavior in web search. In: Geva, S., Trotman, A., Bruza, P., Clarke, C.L.A., Järvelin, K. (eds.) The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast, QLD, Australia—July 06–11, 2014, pp. 103–112. ACM, New York (2014). https://doi.org/10.1145/2600428.2609627
8. Azzopardi, L., Järvelin, K., Kamps, J., Smucker, M.D.: Report on the SIGIR 2010 workshop on the simulation of interaction. SIGIR Forum **44**(2), 35–47 (2010). https://doi.org/10.1145/1924475.1924484
9. Bailey, P., Craswell, N., Soboroff, I., Thomas, P., de Vries, A.P., Yilmaz, E.: Relevance assessment: are judges exchangeable and does it matter. In: Myaeng, S., Oard, D.W., Sebastiani, F., Chua, T., Leong, M. (eds.) Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, Singapore, July 20–24, 2008, pp. 667–674. ACM, New York (2008). https://doi.org/10.1145/1390334.1390447
10. Bailey, P., Craswell, N., White, R.W., Chen, L., Satyanarayana, A., Tahaghoghi, S.M.M.: Evaluating whole-page relevance. In: Crestani, F., Marchand-Maillet, S., Chen, H., Efthimiadis, E.N., Savoy, J., (eds.) Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2010, Geneva, Switzerland, July 19–23, 2010, pp. 767–768. ACM, New York (2010). https://doi.org/10.1145/1835449.1835606
11. Bailey, P., Moffat, A., Scholer, F., Thomas, P.: Retrieval consistency in the presence of query variations. In: Kando, N., Sakai, T., Joho, H., Li, H., de Vries, A.P., White, R.W. (eds.) Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7–11, 2017, pp. 395–404. ACM, New York (2017). https://doi.org/10.1145/3077136.3080839

12. Balog, K., Zhai, C.: User simulation for evaluating information access systems (2023). https://doi.org/10.48550/arXiv.2306.08550
13. Bernstein, Y., Zobel, J.: Redundant documents and search effectiveness. In: Herzog, O., Schek, H., Fuhr, N., Chowdhury, A., Teiken, W. (eds.) Proceedings of the 2005 ACM CIKM International Conference on Information and Knowledge Management, Bremen, Germany, October 31–November 5, 2005, pp. 736–743. ACM, New York (2005). https://doi.org/10.1145/1099554.1099733
14. Bilal, D., Gwizdka, J.: Children's query types and reformulations in google search. Inf. Process. Manag. (IP&M) **54**(6), 1022–1041 (2018). https://www.sciencedirect.com/science/article/pii/S0306457317308889
15. Bonifacio, L.H., Abonizio, H.Q., Fadaee, M., Nogueira, R.F.: InPars: unsupervised dataset generation for information retrieval. In: Amigó, E., Castells, P., Gonzalo, J., Carterette, B., Culpepper, J.S., Kazai, G. (eds.) SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11–15, 2022, pp. 2387–2392. ACM, New York (2022). https://doi.org/10.1145/3477495.3531863
16. Bota, H.S., Zhou, K., Jose, J.M.: Playing your cards right: the effect of entity cards on search behaviour and workload. In: Kelly, D., Capra, R., Belkin, N.J., Teevan, J., Vakkari, P. (eds.) Proceedings of the 2016 ACM Conference on Human Information Interaction and Retrieval, CHIIR 2016, Carrboro, North Carolina, USA, March 13–17, 2016, pp. 131–140. ACM, New York (2016). https://doi.org/10.1145/2854946.2854967
17. Buckley, C., Dimmick, D., Soboroff, I., Voorhees, E.: Bias and the limits of pooling. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval', SIGIR '06, Association for Computing Machinery, New York, NY, USA, pp. 619–620 (2006). https://doi.org/10.1145/1148170.1148284
18. Büttcher, S., Clarke, C.L.A.: Efficiency vs. effectiveness in terabyte-scale information retrieval. In: Voorhees, E.M., Buckland, L.P. (eds.) Proceedings of the Fourteenth Text REtrieval Conference, TREC 2005, Gaithersburg, Maryland, USA, November 15–18, 2005, vol. 500-266. NIST Special Publication, National Institute of Standards and Technology (NIST), New York. http://trec.nist.gov/pubs/trec14/papers/uwaterloo-tera.pdf
19. Cao, H., Jiang, D., Pei, J., He, Q., Liao, Z., Chen, E., Li, H.: Context-aware query suggestion by mining click- through and session data. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '08, pp. 875–883. Association for Computing Machinery, New York, NY, USA (2008). https://doi.org/10.1145/1401890.1401995
20. Carterette, B., Clough, P.D., Hall, M.M., Kanoulas, E., Sanderson, M.: Evaluating retrieval over sessions: the TREC session track 2011-2014. In: Perego, R., Sebastiani, F., Aslam, J.A., Ruthven, I., Zobel, J. (eds.) Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17–21, 2016, pp. 685–688. ACM, New York (2016). https://doi.org/10.1145/2911451.2914675
21. Chakrabarti, D., Kumar, R., Punera, K.: Quicklink selection for navigational query results. In: Quemada, J., León, G., Maarek, Y.S., Nejdl, W. (eds.) Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain, April 20–24, 2009, pp. 391–400. ACM, New York (2009). https://doi.org/10.1145/1526709.1526762
22. Chapelle, O., Metlzer, D., Zhang, Y., Grinspan, P.: Expected reciprocal rank for graded relevance. In: Cheung, D.W., Song, I., Chu, W.W., Hu, X., Lin, J. (eds.) Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, Hong Kong, China, November 2–6, 2009, pp. 621–630. ACM, New York (2009). https://doi.org/10.1145/1645953.1646033
23. Chen, J., Zhang, R., Guo, J., de Rijke, M., Chen, W., Fan, Y., Cheng, X.: Continual learning for generative retrieval over dynamic corpora. In: Frommholz, I., Hopfgartner, F., Lee, M., Oakes, M., Lalmas, M., Zhang, M., Santos, R.L.T. (eds.) Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM 2023, Birmingham, United Kingdom, October 21–25, 2023, pp. 306–315, ACM, New York (2023). https://doi.org/10.1145/3583780.3614821

24. Chen, J., Zhang, R., Guo, J., de Rijke, M., Liu, Y., Fan, Y., Cheng, X.: A unified generative retriever for knowledge- intensive language tasks via prompt learning. In: Chen, H., Duh, W.E., Huang, H., Kato, M.P., Mothe, J., Poblete, B. (eds.) Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23–27, 2023, pp. 1448–1457. ACM, New York (2023). https://doi.org/10.1145/3539618.3591631
25. Chu, P., Komlodi, A., Rózsa, G.: Online search in English as a non-native language. In: Information Science with Impact: Research in and for the Community—Proceedings of the 78th ASIS&T Annual Meeting, ASIST 2015, St. Louis, Missouri, Missouri, USA, October 6–10, 2015, vol. 52, pp. 1–9. Wiley, New York (2015). https://doi.org/10.1002/pra2.2015.145052010040
26. Church, K., Smyth, B., Cotter, P., Bradley, K.: Mobile information access: A study of emerging search behavior on the mobile internet. ACM Trans. Web (TWEB) **1**(1), 4 (2007). https://doi.org/10.1145/1232722.1232726
27. Clarke, C.L.A., Agichtein, E., Dumais, S.T., White, R.W.: The influence of caption features on clickthrough patterns in web search. In: Kraaij, W., de Vries, A.P., Clarke, C.L.A., Fuhr, N., Kando, N. (eds.) Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23–27, 2007, pp. 135–142. ACM, New York (2007). https://doi.org/10.1145/1277741.1277767
28. Clarke, C.L.A., Culpepper, J.S., Moffat, A.: Assessing efficiency–effectiveness tradeoffs in multi-stage retrieval systems without using relevance judgments. Inf. Retr. **19**(4), 351–377 (2016). https://doi.org/10.1007/s10791-016-9279-1
29. Clarke, C.L., Kolla, M., Cormack, G.V., Vechtomova, O., Ashkan, A., Büttcher, S., MacKinnon, I.: Novelty and diversity in information retrieval evaluation. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 659–666 (2008)
30. Craswell, N., Zoeter, O., Taylor, M.J., Ramsey, B.: An experimental comparison of click position-bias models. In: Najork, M., Broder, A.Z., Chakrabarti, S. (eds.) Proceedings of the International Conference on Web Search and Web Data Mining, WSDM 2008, Palo Alto, California, USA, February 11–12, 2008, pp. 87–94. ACM, New York (2008). https://doi.org/10.1145/1341531.1341545
31. Croft, W.B., Harper, D.J.: Using probabilistic models of document retrieval without relevance information. J. Doc. **35**(4), 285–295 (1979). https://doi.org/10.1108/eb026683
32. Culpepper, J.S., Faggioli, G., Ferro, N., Kurland, O.: Topic difficulty: collection and query formulation effects. ACM Trans. Inf. Syst. (TOIS) **40**(1), 19:1–19:36 (2022). https://doi.org/10.1145/3470563
33. Deckers, N., Fröbe, M., Kiesel, J., Pandolfo, G., Schröder, C., Stein, B., Potthast, M.: The infinite index: Information retrieval on generative text-to-image models. In: Gwizdka, J., Rieh, S.Y. (eds.) Proceedings of the 2023 Conference on Human Information Interaction and Retrieval, CHIIR 2023, Austin, TX, USA, March 19–23, 2023, pp. 172–186. ACM, New York (2023). https://doi.org/10.1145/3576840.3578327
34. Engelmann, B., Breuer, T., Friese, J.I., Schaer, P., Fuhr, N.: Context-driven interactive query simulations based on generative large language models. In: Goharian, N., Tonellotto, N., He, Y., Lipani, A., McDonald, G., Macdonald, C., Ounis, I. (eds.) Advances in Information Retrieval—46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24–28, 2024, Proceedings, Part II. Lecture Notes in Computer Science, vol. 14609, pp. 173–188. Springer, Berlin (2024). https://doi.org/10.1007/978-3-031-56060-6_12
35. Faggioli, G., Dietz, L., Clarke, C.L.A., Demartini, G., Hagen, M., Hauff, C., Kando, N., Kanoulas, E., Potthast, M., Stein, B., Wachsmuth, H.: Perspectives on large language models for relevance judgment. In: Yoshioka, M., Kiseleva, J., Aliannejadi, M. (eds.) Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2023, Taipei, Taiwan, 23 July 2023, pp. 39–50. ACM, New York (2023). https://doi.org/10.1145/3578337.3605136

36. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Guo, Q., Wang, M., Wang, H.: Retrieval-augmented generation for large language models: a survey. CoRR abs/2312.10997 (2023). https://doi.org/10.48550/arXiv.2312.10997
37. Gienapp, L., Scells, H., Deckers, N., Bevendorff, J., Wang, S., Kiesel, J., Syed, S., Fröbe, M., Zuccon, G., Stein, B., Hagen, M., Potthast, M.: Evaluating generative ad hoc information retrieval. CoRR abs/2311.04694 (2023). https://doi.org/10.48550/arXiv.2311.04694
38. Glover, E.: The real world web search problem: bridging the gap between academic and commercial understanding of issues and methods. In: Steinberger, R., Fogelman-Soulié, F., Perrotta, D., Piskorski, J. (eds.) Mining Massive Data Sets for Security, pp. 115–129. IOS Press, Oxford (2007)
39. Haas, K., Mika, P., Tarjan, P., Blanco, R.: Enhanced results for web search. In: Ma, W., Nie, J., Baeza-Yates, R., Chua, T., Croft, W.B. (eds.) Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25–29, 2011, pp. 725–734. ACM, New York (2011). https://doi.org/10.1145/2009916.2010014
40. Hämäläinen, P., Tavast, M., Kunnari, A.: Evaluating large language models in generating synthetic HCI research data: a case study. In: Schmidt, A., Väänänen, K., Goyal, T., Kristensson, P.O., Peters, A., Mueller, S., Williamson, J.R., Wilson, M.L. (eds.) Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI 2023, Hamburg, Germany, April 23–28, 2023, pp. 433:1–433:19. ACM, New York (2023). https://doi.org/10.1145/3544548.3580688
41. Harman, D.: User-friendly systems instead of user-friendly front-ends. J. Am. Soc. Inf. Sci. (JASIST) **43**(2), 164–174 (1992). https://doi.org/10.1002/(SICI)1097-4571(199203)43:2<164::AID-ASI9>3.0.CO;2-W
42. Harvey, M., Pointon, M.: Searching on the go: the effects of fragmented attention on mobile web search tasks. In: Kando, N., Sakai, T., Joho, H., Li, H., de Vries, A.P., White, R.W. (eds.) Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7–11, 2017, pp. 155–164. ACM, New York (2017). https://doi.org/10.1145/3077136.3080770
43. Hearst, M.A., Pedersen, J.O.: Reexamining the cluster hypothesis: Scatter/gather on retrieval results. In: Frei, H., Harman, D., Schäuble, P., Wilkinson, R. (eds.) Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'96, August 18–22, 1996, Zurich, Switzerland (Special Issue of the SIGIR Forum), pp. 76–84. ACM, New York (1996). https://doi.org/10.1145/243199.243216
44. Jardine, N., van Rijsbergen, C.J.: The use of hierarchic clustering in information retrieval. Inf. Storage Retr. **7**(5), 217–240 (1971). https://doi.org/10.1016/0020-0271(71)90051-9
45. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Trans. Inf. Syst. (TOIS) **20**(4), 422–446 (2002). http://doi.acm.org/10.1145/582415.582418
46. Kurland, O.: The cluster hypothesis in information retrieval. In: de Rijke, M., Kenter, T., de Vries, A.P., Zhai, C., de Jong, F., Radinsky, K., Hofmann, K. (eds.) Advances in Information Retrieval—36th European Conference on IR Research, ECIR 2014, Amsterdam, The Netherlands, April 13–16, 2014. Proceedings. Lecture Notes in Computer Science, vol. 8416, pp. 823–826. Springer, Berlin (2014). https://doi.org/10.1007/978-3-319-06028-6_105
47. Mackenzie, J.M., Benham, R., Petri, M., Trippas, J.R., Culpepper, J.S. Moffat, A.: CC-News-En: a large English news corpus. In: d'Aquin, M., Dietze, S., Hauff, C., Curry, E., Cudré-Mauroux, P. (eds.) The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19–23, 2020, pp. 3077–3084. ACM, New York (2020). https://doi.org/10.1145/3340531.3412762
48. Mackie, I., Chatterjee, S., Dalton, J.: Generative relevance feedback with large language models. In: Chen, H., Duh, W.E., Huang, H., Kato, M.P., Mothe, J., Poblete, B. (eds.) Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23–27, 2023, pp. 2026–2031. ACM, New York (2023). https://doi.org/10.1145/3539618.3591992

49. Maxwell, D., Azzopardi, L.: Stuck in traffic: how temporal delays affect search behaviour. In: Elsweiler, D., Ludwig, B., Azzopardi, L., Wilson, M.L. (eds.) Fifth Information Interaction in Context Symposium, IIiX '14, Regensburg, Germany, August 26–29, 2014, pp. 155–164. ACM, New York (2014). https://doi.org/10.1145/2637002.2637021
50. Moffat, A., Mackenzie, J., Thomas, P., Azzopardi, L.: A flexible framework for offline effectiveness metrics. In: Amigó, E., Castells, P., Gonzalo, J., Carterette, B., Culpepper, J.S., Kazai, G. (eds.) The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11–15, 2022, pp. 578–587. ACM, New York (2022). https://doi.org/10.1145/3477495.3531924
51. Moffat, A., Zobel, J.: Rank-biased precision for measurement of retrieval effectiveness. ACM Trans. Inf. Syst. (TOIS) **27**(1), 2:1–2:27 (2008). https://doi.org/10.1145/1416950.1416952
52. Monchaux, S., Amadieu, F., Chevalier, A., Mariné, C.: Query strategies during information searching: effects of prior domain knowledge and complexity of the information problems to be solved. Inf. Process. Manag. (IP&M) **51**(5), 557–569 (2015). https://doi.org/10.1016/j.ipm.2015.05.004
53. Navalpakkam, V., Jentzsch, L., Sayres, R., Ravi, S., Ahmed, A., Smola, A.J.: Measurement and modeling of eye-mouse behavior in the presence of nonlinear page layouts. In: Schwabe, D., Almeida, V.A.F., Glaser, H., Baeza-Yates, R. Moon, S.B. (eds.) 22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013, International World Wide Web Conferences Steering Committee, pp. 953–964. ACM, New York (2013). https://doi.org/10.1145/2488388.2488471
54. Oard, D.W., Sakai, T., Kando, N.: Celebrating 20 years of NTCIR: the book. In: Ferro, N., Soboroff, I., Zhang, M. (eds.) Proceedings of the 9th International Workshop on Evaluating Information Access co-located with the 14th NTCIR Conference on the Evaluation of Information Access Technologies (NTCIR 2019), Tokyo, Japan, June 10, 2019 (2019). https://research.nii.ac.jp/ntcir/workshop/OnlineProceedings14/pdf/evia/01-EVIA2019-BOOK-OardD.pdf
55. Oliveira, B., Lopes, C.T.: The evolution of web search user interfaces—an archaeological analysis of google search engine result pages. In: Gwizdka, J., Rieh, S.Y. (eds.) Proceedings of the 2023 Conference on Human Information Interaction and Retrieval, CHIIR 2023, Austin, TX, USA, March 19–23, 2023, pp. 55–68 ACM (2023). https://doi.org/10.1145/3576840.3578320
56. Penha, G., Câmara, A., Hauff, C.: Evaluating the robustness of retrieval pipelines with query variation generators, In: Hagen, M., Verberne, S., Macdonald, C., Seifert, C., Balog, K., Nørvåg, K., Setty, V. (eds.), Advances in Information Retrieval—44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part I. Lecture Notes in Computer Science, vol. 13185, pp. 397–412. Springer, Berlin (2022). https://doi.org/10.1007/978-3-030-99736-6_27
57. Penha, G., Palumbo, E., Aziz, M., Wang, A., Bouchard, H.: Improving content retrievability in search with controllable query generation. In: Ding, Y., Tang, J., Sequeda, J.F., Aroyo, L., Castillo, C., Houben, G. (eds.) Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023–4 May 2023, pp. 3182–3192. ACM, New York (2023). https://doi.org/10.1145/3543507.3583261
58. Pera, M.S., Murgia, E., Landoni, M., Huibers, T., Aliannejadi, M.: Where a little change makes a big difference: a preliminary exploration of children's queries. In: Kamps, J., Goeuriot, L., Crestani, F., Maistro, M., Joho, H., Davis, B., Gurrin, C., Kruschwitz, U., Caputo, A. (eds.) Advances in Information Retrieval, pp. 522–533. Springer Nature Switzerland, Cham (2023)
59. Peters, C.: Information retrieval evaluation in a changing world lessons learned from 20 years of CLEF (2019)
60. Pradeep, R., Hui, K., Gupta, J., Lelkes, Á.D., Zhuang, H., Lin, J., Metzler, D., Tran, V.Q.: How does generative retrieval scale to millions of passages?. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6–10, 2023, pp. 1305–1321. Association for Computational Linguistics, California (2023). https://aclanthology.org/2023.emnlp-main.83

61. Rajput, S., Pavlu, V., Golbus, P.B., Aslam, J.A.: A nugget-based test collection construction paradigm, In: Macdonald, C., Ounis, I., Ruthven, I. (eds.) Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24–28, 2011, pp. 1945–1948. ACM, New York (2011). https://doi.org/10.1145/2063576.2063861
62. Rieh, S.Y., Xie, H.I.: Analysis of multiple query reformulations on the web: the interactive information retrieval context. Inf. Process. Manag. (IP&M) **42**(3), 751–768 (2006)
63. Robertson, S.: The probability ranking principle in IR. J. Doc. **33**, 294–304 (1977)
64. Sakai, T., Kato, M.P., Song, Y.-I.: Click the search button and be happy: evaluating direct and immediate information access. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management pp. 621–630 (2011)
65. Sanderson, M.: Test collection based evaluation of information retrieval systems. Found. Trends Inf. Retr. **4**(4), 247–375 (2010). https://doi.org/10.1561/1500000009
66. Sanderson, M., Scholer, F., Turpin, A.: Relatively relevant: assessor shift in document judgements. In: Australasian Document Computing Symposium (2010). https://api.semanticscholar.org/CorpusID:14426189
67. Scholer, F., Turpin, A., Sanderson, M.: Quantifying test collection quality based on the consistency of relevance judgements. In: Ma, W., Nie, J., Baeza-Yates, R., Chua, T., Croft, W.B. (eds.) Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25–29, 2011, pp. 1063–1072. ACM, New York (2011). https://doi.org/10.1145/2009916.2010057
68. Smucker, M.D., Clarke, C.L.A.: Time-based calibration of effectiveness measures. In: Hersh, W.R., Callan, J., Maarek, Y., Sanderson, M. (eds.) The 35th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 12, Portland, OR, USA, August 12–16, 2012, pp. 95–104. ACM, New York (2012). https://doi.org/10.1145/2348283.2348300
69. Spärck Jones, K., Van Rijsbergen, C.: Report on the need for and provision of an ideal information retrieval test collection, Technical Report 5266, British Library Research and Development Report (1975). https://cir.nii.ac.jp/crid/1570572699089480448
70. Sun, W., Yan, L., Chen, Z., Wang, S., Zhu, H., Ren, P., Chen, Z., Yin, D., de Rijke, M., Ren, Z.: Learning to tokenize for generative retrieval. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10–16, 2023 (2023). http://papers.nips.cc/paper_files/paper/2023/hash/91228b942a4528cdae031c1b68b127e8-Abstract-Conference.html
71. Teevan, J., Collins-Thompson, K., White, R.W., Dumais, S.: Slow search. Commun. ACM **57**(8), 36–38 (2014). https://doi.org/10.1145/2633041
72. Thomas, P., Spielman, S., Craswell, N., Mitra, B.: Large language models can accurately predict searcher preferences. CoRR abs/2309.10621 (2023). https://doi.org/10.48550/arXiv.2309.10621
73. Tombros, A., Sanderson, M.: Advantages of query biased summaries in information retrieval. In: Croft, W.B., Moffat, A., van Rijsbergen, C.J., Wilkinson, R., Zobel, J. (eds.) Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24–28 1998, Melbourne, Australia, pp. 2–10. ACM, New York (1998). https://doi.org/10.1145/290941.290947
74. Tonmoy, S.M.T.I., Zaman, S.M.M., Jain, V., Rani, A., Rawte, V., Chadha, A., Das, A.: A comprehensive survey of hallucination mitigation techniques in large language models. CoRR abs/2401.01313 (2024). https://doi.org/10.48550/arXiv.2401.01313
75. Torres, S.D., Hiemstra, D., Serdyukov, P.: Query log analysis in the context of information retrieval for children. In: Crestani, F., Marchand-Maillet, S., Chen, H., Efthimiadis, E.N., Savoy, J. (eds.) Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2010, Geneva, Switzerland, July 19–23, 2010, pp. 847–848 ACM. (2010). https://doi.org/10.1145/1835449.1835646

76. Turpin, A., Scholer, F.: User performance versus precision measures for simple search tasks. In: Efthimiadis, E.N., Dumais, S.T., Hawking, D., Järvelin, K. (eds.) Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA, August 6–11, 2006, pp. 11–18. ACM, New York (2006). https://doi.org/10.1145/1148170.1148176
77. van Rijsbergen, C.: Information Retrieval. Butterworth-Heinemann, New York (1979)
78. Vdorhees, E.M.: The cluster hypothesis revisited. SIGIR Forum **51**(2), 35–43 (2017). https://doi.org/10.1145/3130348.3130353
79. Voorhees, E., Harman, D., of Standards, N.I., (US), T.: TREC: experiment and evaluation in information retrieval, vol. 63. MIT Press, Cambridgeˆ eMA MA (2005)
80. Voorhees, E.M.: The evolution of Cranfield. In: Ferro, N., Peters, C. (eds.) Information Retrieval Evaluation in a Changing World—Lessons Learned from 20 Years of CLEF. The Information Retrieval Series, vol. 41, pp. 45–69. Springer, Berlin (2019). https://doi.org/10.1007/978-3-030-22948-1_2
81. Voorhees, E.M., Craswell, N., Lin, J.: Too many relevants: whither Cranfield test collections?. In: Amig, E.ó, Castells, P., Gonzalo, J., Carterette, B., Culpepper, J.S. Kazai, G. (eds.) The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11–15, 2022, pp. 2970–2980. ACM, New York (2022). https://doi.org/10.1145/3477495.3531728
82. Wang, L., Lin, J., Metzler, D.: A cascade ranking model for efficient ranked retrieval. In: Ma, W., Nie, J., Baeza-Yates, R., Chua, T., Croft, W.B. (eds.) Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25–29, 2011, pp. 105–114. ACM, New York (2011). https://doi.org/10.1145/2009916.2009934
83. White, R.W., Dumais, S.T., Teevan, J.: Characterizing the influence of domain expertise on web search behavior. In: Baeza-Yates, R., Boldi, P., Ribeiro-Neto, B.A., Cambazoglu, B.B. (eds.) Proceedings of the Second International Conference on Web Search and Web Data Mining, WSDM 2009, Barcelona, Spain, February 9–11, 2009, pp. 132–141. ACM, New York (2009). https://doi.org/10.1145/1498759.1498819
84. Wu, Z., Sanderson, M., Cambazoglu, B.B., Croft, W.B., Scholer, F.: Providing direct answers in search results: a study of user behavior. In: d'Aquin, M., Dietze, S., Hauff, C., Curry, E., Cudré-Mauroux, P. (eds.) The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020, pp. 1635–1644. ACM, New York (2020). https://doi.org/10.1145/3340531.3412017
85. Yang, T., Song, M., Zhang, Z., Huang, H., Deng, W., Sun, F., Zhang, Q.: Auto search indexer for end-to-end document retrieval. In: Bouamor, H., Pino, J., Bali, K. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6–10, 2023, pp. 6955–6970. Association for Computational Linguistics, California (2023). https://aclanthology.org/2023.findings-emnlp.464
86. Zhai, C., Cohen, W.W., Lafferty, J.D.: Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. SIGIR Forum **49**(1), 2–9 (2015). https://doi.org/10.1145/2795403.2795405
87. Zhang, Y., Hu, C., Liu, Y., Fang, H., Lin, J.: Learning to rank in the age of Muppets: effectiveness–efficiency tradeoffs in multi-stage ranking. In: Moosavi, N.S. Gurevych, I., Fan, A., Wolf, T., Hou, Y., Marasović, A., Ravi, S. (eds.) Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing Association for Computational Linguistics, Virtual, pp. 64–73 (2021). https://aclanthology.org/2021.sustainlp-1.8
88. Zobel, J.: How reliable are the results of large-scale information retrieval experiments? In: Croft, W.B., Moffat, A., van Rijsbergen, C.J., Wilkinson, R., Zobel, J. (eds.) Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24–28 1998, Melbourne, Australia, pp. 307–314. ACM, New York (1998). https://doi.org/10.1145/290941.291014

Chapter 7
Sociotechnical Implications of Generative Artificial Intelligence for Information Access

Bhaskar Mitra, Henriette Cramer, and Olya Gurevich

Abstract Robust access to trustworthy information is a critical need for society with implications for knowledge production, public health education, and promoting informed citizenry in democratic societies. Generative AI technologies may enable new ways to access information and improve effectiveness of existing information retrieval systems, but we are only starting to understand and grapple with their long-term social implications. In this chapter, we present an overview of some of the systemic consequences and risks of employing generative AI in the context of information access. We also provide recommendations for evaluation and mitigation and discuss challenges for future research.

7.1 Introduction

Robust access to trustworthy information is a critical need for society including implications for knowledge production, public health education, and promoting informed citizenry in democratic societies. Generative AI technologies such as Large Language Models (LLMs) may enable new ways to access information and improve effectiveness of existing Information Retrieval (IR) systems. More efficient basic task execution with the help of LLMs can also enable people to focus on the more challenging aspects of information retrieval–related tasks and research. However, the long-term social implications of deploying these technologies in the context of information access are not yet well understood. Existing research has focused on how these models may generate biased and harmful content [1, 16, 64, 74, 118, 153, 235] as well as the environmental costs [16, 25, 56, 161, 162, 241] of developing and deploying these models at scale. In the context of information

B. Mitra (✉)
Microsoft Research, Montréal, QC, Canada
e-mail: bmitra@microsoft.com

H. Cramer · O. Gurevich
PaperMoon AI, San Francisco, CA, USA

R. W. White, C. Shah (eds.), *Information Access in the Era of Generative AI*,
The Information Retrieval Series 51, https://doi.org/10.1007/978-3-031-73147-1_7

access, Shah and Bender [184] have argued that certain framings of LLMs as "search engines" lack the necessary theoretical underpinnings and may constitute as a category error.

In this current work, we present a broader perspective on the sociotechnical implications of generative AI for information access. Our perspective is informed by existing literature and aims to provide a summary of known challenges viewed through a systemic lens that we hope will serve as a useful resource for future critical research in this area. We present a summary of these implications next, followed by recommendations for evaluation and mitigation later in this chapter.

7.2 Implications of Generative AI for Information Access

We present a reflection on the potential sociotechnical implications of generative AI, with an emphasis on LLMs, for information access. Generative AI is still an emerging technology, and our understanding of its sociotechnical impact today, and how it may evolve over time, is fairly limited. Our treatment of this topic is therefore necessarily both incomplete and speculative. We are informed by several recent works [16, 197, 233, 234] that attempt to map the landscape of risks and harms from LLMs. What distinguishes our treatment of this topic relative to this previous literature is the specific focus on information access. There has also been work on the considerations for specific applications of LLMs in IR, such as for generating direct responses to users' expressed information needs [184], which is relevant to our current discussion. However, a thorough exploration of every potential application of LLMs in IR systems is beyond the scope of our current work. Instead, we explore the implications for information access through a broader lens that encompasses considerations for content creation, content retrieval, sociopolitical power dynamics, geopolitical inequities, crowd work, ecology, and future of IR research. We reference relevant previous taxonomies and studies throughout this section to both support our claims and to establish meaningful connections in an attempt to present a more complete and consistent view on this topic to the reader.

We adopt the Consequences-Mechanisms-Risks (CMR) framework proposed by Gausen et al. [70] to structure our presentation. Gausen et al. introduce the CMR framework to support designers and developers of AI (and in general any computational) systems to identify and understand: (i) The systemic consequences of developing and deploying the technology under study in the real world (ii) The mechanisms introduced by the said technology responsible for these consequences (iii) The corresponding risks to relevant stakeholders The framework intentionally explicates the higher-level consequences to motivate viewing the challenges through a more systemic lens. The mechanisms, in turn, focus on more low-level system behaviors and aspects of the technology development process that contribute to the consequences and risks and therefore represent sites for more actionable mitigation. These consequences and mechanisms are mapped to relevant potential risks. Through literature survey, in this work, we identify the consequences, mechanisms,

Table 7.1 Overview of potential negative consequences for information access from generative AI, the related mechanisms introduced by these AI technologies, and corresponding risks

Consequences	Mechanisms	Risks
Information ecosystem disruption (Sect. 7.2.1.1)	Content pollution (Sect. 7.2.1.1)	Risks to society: democracy, health and well-being, and global inequity (Sect. 7.2.2.1)
	The "Game of telephone" effect (Sect. 7.2.1.1)	
	Search engine manipulation (Sect. 7.2.1.1)	
	Degrading retrieval quality (Sect. 7.2.1.1)	
	Direct model access (Sect. 7.2.1.1)	
	The paradox of reuse (Sect. 7.2.1.1)	
Concentration of power (Sect. 7.2.1.2)	Compute and data moat (Sect. 7.2.1.2)	
	AI persuasion (Sect. 7.2.1.2)	
	AI alignment (Sect. 7.2.1.2)	
Marginalization (Sect. 7.2.1.3)	Appropriation of data labor (Sect. 7.2.1.3)	
	Bias amplification (Sect. 7.2.1.3)	
	AI exploitation and doxing (Sect. 7.2.1.3)	
Innovation decay (Sect. 7.2.1.4)	Industry capture (Sect. 7.2.1.4)	Risks to IR research (Sect. 7.2.2.2)
	Pollution of research artefacts (Sect. 7.2.1.4)	
Ecological impact (Sect. 7.2.1.5)	Resource demand and waste (Sect. 7.2.1.5)	Risks to environment (Sect. 7.2.2.3)
	Persuasive advertising (Sect. 7.2.1.5)	

and risks of generative AI in the context of information access and organize them according to the CMR framework as shown in Table 7.1. While we acknowledge that this list of consequences-mechanisms-risks is incomplete, we hope that it provides a summary of the sociotechnical concerns already identified in existing literature and provokes new questions for critical future research.

7.2.1 Consequences and Mechanisms

In the context of information access, we identify five potential categories of negative consequences of generative AI and corresponding mechanisms, which we discuss next.

7.2.1.1 Consequence: Information Ecosystem Disruption

To reflect on the implications of generative AI on information access, we must consider the information ecosystem as a whole, and not constrain our discussion only to the application of these emerging technologies directly in IR systems. This

ecosystem includes different actors and stakeholders such as information seekers, content producers, IR systems developers, advertisers, and other sociopolitical actors. While the information ecosystem is constantly evolving, generative AI holds the potential to significantly disrupt how each of these actors operate on their own and how they relate to other actors and stakeholders. This potential for disruption spans across how content is produced, consumed, monetized, and used toward specific ends. By no means do we want to imply that these plausible changes are inherently bad, but the scale of potential disruptions across the ecosystem should motivate careful and thoughtful considerations before these technologies are deployed at scale. We discuss next some the underlying mechanisms introduced by generative AI that may contribute to these disruptions. We encourage the reader to view these mechanisms not just in isolation but to also consider how they may interact with each other and how that may impact the ecosystem over time.

Mechanism: Content Pollution Generative AI enables low-cost generation of derivative low-quality content at an unprecedented scale. As a consequence, synthetic AI-generated content is rapidly and very widely appearing on the Web [98]. On Amazon,[1] AI-generated content includes scammy derivatives of existing publications [115, 129, 155] and fake travel guides [119]. On YouTube,[2] AI-generated video creators have targeted children [4, 97, 116]. We are also witnessing a proliferation of news Web sites almost entirely generated by AI [179], which are being surfaced in search results [46] and funded by online ads [27]. Even reputable publishers have reportedly published AI-generated articles under fake AI-generated author profiles [57]. Beyond news, other synthetic content such as AI-generated images is starting to pollute search results [5, 58]. According to another recent study [212], a "shocking" amount of content on the Web today is machine-translated text. The promise of machine translation is that it could make more content accessible to wider audiences. However, it also amplifies the influence of (sometimes questionable-quality) language technology choices. For example, Thompson et al. [212] found that more low-quality content—rather than high-quality content—was machine translated into lower-resource languages, likely with the goal of generating ad revenue. Concerns have also been raised about LLMs potentially serving as "Misinformation Superspreaders" [26, 157] as they make it trivially easy to inundate the Web with "firehoses of falsehoods."[3] Hoel [97] points out that AI pollution of our information ecosystems is a "tragedy of the commons" [92].

Pollution of our information ecosystem at such scale has critical implications for people and society. When authoring a document requires significant time and effort, then quality, style, and comprehensiveness are factors that readers may consider in deciding whether and how much to trust its content. However, when the cost

[1] https://www.amazon.com/

[2] https://www.youtube.com/

[3] https://en.wikipedia.org/wiki/Firehose_of_falsehood

of writing an extensive article approaches zero, it becomes significantly harder for the reader to make that decision. They may not be able to distinguish between an article created based on extensive research, fact-checking, and thoughtful writing practices and one generated instantly based on a short user prompt. Furthermore, the increasing adoption of these same AI authoring tools by reputable publishers and content producers may homogenize the language and style of content on the Web, making it even more difficult for readers to distinguish them from low-quality AI-generated content whose sole intent is to attract ad revenue or to mislead. Such Web pollution is also a concern for future AI models that require large Web-scale datasets to train on. Including AI-generated content in the training data for new AI models may have significant negative impact on model performance, what has been referred to as "Model collapse" [137, 189], "Model Autophagy Disorder" [6], and "Habsburg AI."[4]

Mechanism: The "Game of Telephone" Effect LLMs have recently been employed in conversational search interfaces. In systems such as Bing Copilot, the LLM has access to relevant Web search results from which it can draw information to produce appropriate responses for the information needs expressed by a user. In this scenario, the LLM performs a complex summarization task extracting relevant information from the retrieved documents to answer the search query. In doing so, the LLM now inserts itself between the user and the retrieved Web results. This shifts the responsibility of inspecting the information in the documents and assessing their relevance, trustworthiness, and surrounding context from the user to the LLM. Further, factual errors and inconsistencies may arise between what the LLM produces and what is in the retrieved documents. Seeing the model through an anthropomorphic lens, these errors are sometimes referred to as "hallucinations." A more technical view may see this as a noisy translation akin to the children's game of telephone.[5] Such errors, often subtle and hard to spot, may contribute to misinformation and reduce robustness of the information access system. While the LLM-generated responses may cite relevant documents, it is unlikely that users diligently click the provided links and verify the information in the response is indeed supported by said sources. Even if the LLM reproduces exact pieces of text from the source documents without error, taking these out of the context of the document may lead to unexpected negative consequences. Such examples have previously been reported [215] in context of extracted answers that search engines display on the Search Engine Result Pages (SERPs) as response to the user query. These issues may become more prevalent if conversational search interfaces become a popular way to access online information.

In a more radical proposal, Metzler et al. [141] have suggested that LLMs could directly replace retrieval systems and respond directly to the user based on information in their training data. LLMs are trained to produce statistically plausible

[4] https://twitter.com/jathansadowski/status/1625245803211272194

[5] https://en.wikipedia.org/wiki/Game_of_telephone

text sequences, and any semblance to an information retrieval system is likely an important mis-categorization of these models that we should be wary of [184]. The game of telephone effect is likely to be more intense when LLMs are expected to produce information from their training data and not just the in-context information in its input.

The interjection of the LLM between the user and the search results may have other long-term effects. These interfaces may disincentivize users from the practice of verifying information sources and make them less skilled over time at discerning online misinformation. If users get accustomed to information being presented neatly summarized and disconnected from original sources, the critical cognitive skills necessary to distinguish between trustworthy and untrustworthy information may atrophy.

Mechanism: Search Engine Manipulation New applications of LLMs to the IR stack have exposed new attack vectors. Prompt injection attacks [83, 131, 132] that try to blur the line between instructions and data have garnered specific interest. In these types of attacks, Web site owners may inject what looks like instructions to the LLM. When such documents are retrieved and included in the input of the LLM as augmentation, the LLM may mistake the injected prompt in the document content and be vulnerable to manipulation.

Recently, LLMs have also found application in relevance labeling for search [211]. It is not well understood yet whether this may make the search engine vulnerable to improper ranking manipulation by Web site owners and search engine optimization experts. For example, one may employ the same, or similar, LLMs to reproduce the labeling scheme externally and then adapt their Web site content and design to achieve undue high predicted relevance against queries to rank higher on SERPs.

Other attack vectors may include using LLMs to create effective content farms at low cost to manipulate the ranking of Web results or even use LLMs to artificially simulate users interacting with the search system to fake clicks and other user behavior signals, such as reformulations, which search engines depend on.

Mechanism: Degrading Retrieval Quality LLM usage can negatively impact search result quality in a number of (indirect) ways. LLMs can contribute to new attack vectors, but more worryingly, in some cases, the negative effect may be a result of the LLM behaving exactly as it is supposed to. For example, one potential consequence of using conversational search interfaces, is that the quality of feedback from user behavior signals on SERPs may significantly degrade. Historically, users of commercial Web search engines have given search systems noisy implicit feedback through clicks and other actions on SERPs. These actions are part of the key secret sauce of any modern search systems.

However, conversational interfaces may discourage direct user clicks on Web results and at best provide much weaker satisfaction signal that may be gleaned from the users' next utterance in the conversation. This over time may negatively impact the underlying retrieval quality. This makes it important to invest in methods that can infer user satisfaction with high certainty from the natural-language conversations.

However, methods for such signal interpretation are not yet at the level necessary to mitigate these impacts.

In conversational search interfaces and other applications, such as Microsoft Copilot for M365 [140, 231], the LLM may conduct the search on the user's behalf. In this process, the LLM generates search queries. If these queries differ from those that are likely to be submitted by users, then the underlying search system needs to optimize itself for both real user queries and LLM-generated queries. This may have consequences that are not yet well understood. Optimizing the search system directly to improve the LLMs natural language responses may also have unforeseen outcomes, especially in light of the fact that what makes for a good result set for retrieval-augmentation is not yet fully understood [51].

Mechanism: Direct Model Access Another important consideration is the implications of open foundation models [109]. While centralized systems have their own negative implications, as discussed in Sect. 7.2.1.2, open-access generative AI models without any access moderation also pose certain challenges. For example, there are many classes of harmful intents that systems should refuse to respond to. This may include search queries seeking information on methods to self-harm or cause harms to others or requests to generate harmful (and sometimes illegal) content such as Child Sex Abuse Material (CSAM) or Non-Consensual Intimate Information (NCII). Publicly accessible LLMs trained on large Web corpora may produce such irresponsible content in the absence of moderation. Even if a model is trained to not respond to certain classes of queries, it is likely that there will be leakage, and the safety alignment may also be compromised if the model is further finetuned [171]. Such leakage may also happen in the context of traditional search systems. However, in the latter case, all queries are typically logged, allowing for post hoc analysis and identification of critical gaps in the moderation system. Unfortunately, no such mitigation is possible once these generative AI models are released into the wild.

Mechanism: The Paradox of Reuse Content producers and information access technologies are critically inter-dependent [139, 225]. Web sites such as Wikipedia,[6] Stack Exchange,[7], and Reddit[8] produce critical content that is surfaced by information access platforms (e.g., Web search engines) and contribute to making these platforms significantly more useful to their users. In return, these platforms have historically sent traffic back to the Web sites that contribute to their increased readership, subscriptions, and monetization. However, when search platforms stop directing traffic back to Web sites—e.g., by instead surfacing relevant content directly on the search result pages (SERPs)—the relationship becomes less symbiotic toward the content producers, a phenomenon Taraborelli [206] termed the "paradox of reuse."

[6] https://www.wikipedia.org/

[7] https://stackexchange.com/

[8] https://www.reddit.com/

The application of LLMs as conversational information access interfaces is likely to significantly intensify this problem. For example, LLMs such as ChatGPT[9] and Google Gemini[10] may gobble up large quantities of content from Web sites as part of their training data and later regurgitate the same information without any attribution back to the sources. Even when models summarize information from multiple online sources with attribution, e.g., Bing Copilot,[11] they typically de-emphasize the references and reduce the likelihood of the searcher clicking through to the source Web sites as compared to the classic ten-blue-links interface. There is evidence [52] to suggest that this phenomenon is already happening at scale and is jeopardizing the "grand bargain at the heart of the web" [93].

7.2.1.2 Consequence: Concentration of Power

> We may have democracy, or we may have wealth concentrated in the hands of a few, but we can't have both.
> – Louis Brandeis
> As quoted by Lonergan [133]

Technology shapes and is shaped by the sociopolitical power structures within which it exists. The 2024 edition of the World Economic Forum's Global Risks Report [240] lists "technological power concentration" as one of the top global risks for the coming decade and as the biggest upward mover in their annual ranking of global risks compared to the previous year. Deliberation on the social consequences of any technology must therefore include critical consideration of how the technology, and general narratives about the said technology, shifts power and re-architects and codifies structures of hierarchy and control. In this context, the politics and values of those in power to oversee what and how technology is built or regulated, especially when they reinforce hierarchy and authoritarianism (e.g. [72]), (e.g. [59, 72, 120]), become important to consider.

A report [106] from the research institute AI Now[12] similarly asserts "the concentration of economic and political power in the hands of the tech industry—Big Tech in particular" as the core challenge posed by AI. They further note that not just the technologies but the narratives (both the hype and the fear-mongering) around them questionably bolster claims of "foundational" advancements and their unassailable equivalence with scientific progress. These concerns are complemented by discourses within the AI community, such as observations by Birhane et al. [21] that the prominent values expressed and operationalized in top-cited AI papers generally have implications in support of centralization of power. Even if platform owners act accountably to civil society, the concentration of power and control in

[9] https://chat.openai.com/
[10] https://gemini.google.com/app
[11] https://www.bing.com/chat
[12] https://ainowinstitute.org/

their hands makes them vulnerable to other actors, such as autocratic governments, and allows that power to be potentially abused for oppressive and harmful intents.

The popularization of generative AI can concentrate that power within large companies, since they emerge as some of the only institutions with the resources to develop and deploy these technologies [111]. The application of these technologies for information access may contribute to further concentration and growing inequities of wealth and power; we discuss three mechanisms in the context of generative AI that may contribute to concentration of power and control.

Mechanism: Compute and Data Moat The development of generative AI is heavily reliant on the availability of large swaths of training data and large-scale computing power for training and deployment. Only a handful of institutions, largely in the private sector, own and control these necessary resources while simultaneously evangelizing AI as crucial geopolitical leverage and critical social infrastructure [106]. Increased access to these models has sometimes been touted as potential paths to mitigation [194, 200], where access may range from being heavily restricted over Application Programming Interfaces (APIs) to "open weight" models [128]. The ability to download models with their learned parameters allows others to further adapt for their own applications and opens the door to more meaningful analysis and audit of these models. However, such "open access" also leads to severe limitations that we should recognize. The availability of the trained models does little to challenge the predominant visions put forth by large technology companies of what AI fundamentally should look like.

One potential direction would be to dismantle the data and compute moat by turning them over from private ownership into public infrastructure for independent researchers and developers and those affiliated with smaller institutions. This also illustrates the importance of existing institutions such as archives, libraries, and universities that have reliable, historical data. The availability of public computer infrastructure would allow a broader set of developers to participate in the reimagination and development of diverse approaches to AI and not merely being forced to be satisfied with critiquing and finetuning artefacts produced by other institutions. However, there is no guarantee that without careful planning and incentives, a proliferation of smaller projects will lead to transformative new or more sustainable results.

Democratizing the control over computational resources provides a mechanism of checks and balances on the future directions of AI systems and may allow for challenges to popular narratives and expectations about generative AI such as exponential growth in model size over time. Infrastructure is however also bound to the particular governing system and local underlying goals and processes. Larger investments in existing research institutes or new alternative companies or non-profits might in certain cases lead to faster results.

Similarly, the research community would benefit from easier access to industry models and APIs for critical studies and auditing. However, access to models or APIs alone is significantly limiting unless that access is also extended to the user-facing systems in which these technologies are deployed. The corresponding

instrumentation data would provide context on how these systems are used by people and potential consequences. This can lead to practical privacy and security questions for platform teams. Practical support for decision-making and, for example, the creation of standards to de-risk those concerns can help alleviate some of those concerns.

Mechanism: AI Persuasion There is an emerging recognition of the dangers of *AI persuasion* [30, 34, 62, 159], which Burtell and Woodside [30] define as "a process by which AI systems alter the beliefs of their users." AI systems may persuade users by appealing to their reason and argument or by using their cognitive biases and heuristics [62]. El-Sayed et al. [62] identify six mechanisms of generative AI persuasion—namely: (i) Trust and rapport (ii) Anthropomorphism (iii) Personalization (iv) Deception and lack of transparency (v) Manipulative strategies (vi) Alteration of choice environment—and corresponding model features that contribute to these mechanisms. In the context of information access and advertising, these capabilities of generative AI can be powerful tools to hyper-target users and steer their behaviors.

Modern online information access and communication platforms monetized with targeted advertising have been said to usher in an age of surveillance capitalism [247, 248]. Information access systems increasingly collect detailed user behavior data that allow them to build accurate user profiles for audience targeting. There is strong evidence that people are more likely to consume information that opposes their own personal views and beliefs when it employs language similar to their own political leanings [243]. So combining users' private preferences and behavioral data with the capabilities of generative AI to produce persuasive language could create worrying tools for mass behavioral manipulation. The impact of such pervasive *algorithmic nudging* [134] may be further pronounced over longer time periods from continuous interactions between the user and the system. Putting these capabilities in the hands of online platform owners, which typically tend to be large multinational for-profit institutions with largely hierarchical non-democratic internal governance structures, poses serious risks to functioning of democratic societies. At the same time, platforms must make decisions about what is acceptable on their platforms to avoid negative user experiences, spam, unwelcoming behavior, and other negative occurrences beyond those outlined in legal compliance alone. Platforms moderate content posted or accessible through the platform [77], and in doing so, they unavoidably impose implementations of values on their users or the values incentivized by, say, advertising needs or other business model–related motivations. For ads, this may mean an incentive to use generative AI to produce hyper-targeted highly personalized persuasive advertisements that convince users to make certain buying decisions. For content, when platforms optimize for increased user engagement, they may knowingly or unknowingly incentivize generative AI models to be producing highly charged content, such as "rage-bait" [101], because it tends to be more persuasive and engaging.

Mechanism: AI Alignment To prevent generative AI models from producing harmful and offensive content, recent research has focused on how to align

model outputs with "human values" [66, 67, 110, 177, 203]. Approaches such as Reinforcement Learning from Human Feedback (RLHF) [38, 246] have been effective in limiting certain types of problematic content from being produced. However, this approach presupposes some notions of desirable values and puts the burden of determining and enforcing them on the shoulders of platform/model developers. Any notions of universal values that might determine what type of content these models should generate—or not generate [214]—are highly contested [20, 105, 167, 170, 180]. Placing these decisions in the exclusive domain of the platform developers, especially in the absence of democratic and civil society oversight, further concentrates power and responsibility. This is not an argument against content moderation itself but against the centralization of control over it without civil oversight or broader societal participation. As a pragmatic example, platforms may not necessarily have the necessary knowledge in-house, making it imperative for them to make successful connections to outside expertise.

7.2.1.3 Consequence: Marginalization

Generative AI, both in its process of development and in its deployment in the context of information access, can marginalize groups and individuals by diminishing their value, power, and well-being. Next, we discuss some of the mechanisms that may contribute to this.

Mechanism: Appropriation of Data Labor Li et al. [123] define *data labor* as "activities that produce digital records useful for capital generation." The term encompasses both witting labor activities—as in the case of crowd work [7], peer production [207, 208], and content moderation [77]—and unwitting activities such as user behavior data and other data generated when users interact with and participate on the platforms. Data labor also encompasses the creation of artefacts by writers [40, 41], artists [220, 221], programmers [219], etc. outside of the AI development process that are nonetheless extracted from the Web and fed in as training data to generative AI models. Appropriation of data labor in this context includes both: (i) The uncompensated appropriation of works by writers, authors, programmers, and peer production communities like Wikipedia [10, 28, 29, 36, 37, 40, 41, 76, 136, 188, 218, 219, 221, 223, 224] (ii) Under-compensated crowd work for data labeling that has been instrumental in the development of these technologies [7, 90, 91, 165, 204, 239, 242]

It is particularly harmful when technology developed on appropriated labor is then employed to displace and automate the jobs of those whose labor was appropriated [8, 47, 223]. Introduction of such automation may involve vicious cycles of perceived skill transfer from people to AI models whereby professional jobs are replaced by corresponding lesser-paid gigified equivalent as auditing and editing of model outputs only [82]. Proprietary AI model capabilities may then continue to improve by learning from workers' inputs, while workers progressively

lose their economic value and power or are even relegated into the role of *moral crumple zones* [63].

This is a critical challenge in the context of information access because: (i) The devaluation of writers and artists have direct implications for the quality of content on the Web (ii) These automated content generation tools are starting to get incorporated directly in information access platforms [166] Similar concerns of commodification and appropriation have also been raised in other information and knowledge access contexts such as in the enterprise [70].

AI for Me, Data Labor for Thee Another pernicious aspect of AI data labor dynamics discussed in the literature is how they can mirror and reify racial capitalism and coloniality, employ global labor exploitation and extractive practices, and reinforce the global north and south divide [19, 45, 88, 114, 149, 154, 202]. While worldwide jobs might be created in certain cases, the workers are typically low paid and deprived of any share of the profit made from technologies built with their labor. These dynamics encompass accruing the benefits of generative AI to privileged populations, while data labor is relegated to already marginalized populations, for example, in the global south. Communities that significantly contribute to AI data labor may even find their own linguistic styles being labeled AI-ese [95] and being forced to repeatedly prove their own humanity [53, 138]. Attempts to bridge the global north-south data gap also in turn may further intensify data extractive practices in the global south [39].

Mechanism: Bias Amplification LLMs and other generative models reproduce and amplify harmful biases and stereotypes from their training datasets [1, 16, 23, 24, 31, 79], which can lead to allocative and representational harms [49]. Harms may also materialize from *demographic blindness* [70] when the model (or the system it is embedded in) treats different individuals and groups as alike when, in fact, it is unwarranted. Examples may include the handling of certain languages as one homogeneous entity without regard for sociolects or dialects [22] or holding different perspectives as equally valid without considerations for historical context or structural dynamics of power. These biases are concerning in the context of information access systems that are responsible for supporting informed citizenry and functioning democracies, health literacy, and knowledge production among other societal needs.

Mechanism: AI Exploitation and Doxing *"AI doxing"* can describe the act of leaking people's private information by an AI system. Weidinger et al. [233] note that this may be caused by models leaking private information (e.g., address and telephone number) present in their training data [33] or when these models are employed to predict people's sensitive attributes (e.g., political and sexual identities) based on what is known about them publicly [117, 158, 172, 244]. Private information in the training data is a challenge even if datasets have been sourced from the public Web because models may continue to regurgitate that information after it has been removed from the Web or bypass safety measures that would prevent such information from surfacing through Web search—e.g., the information

may be protected by robots.txt that blocks popular search crawlers but misses crawler bots that specifically collect data for AI model training. In many contexts, applications of these models to predict people's private information may be based on shaky scientific grounds [2, 217], to put it mildly. However, such applications may still contribute to serious harms and discrimination regardless of their accuracy as long as some people are convinced of their predictive power and employ them to marginalize others. AI doxing may also take other forms such as reverse-image-search [14], a functionality supported by some search engines that may be abused for stalking and harassment. In turn, exploitative materials produced with GenAI (such as deepfake revenge porn or CSAM) might be amplified.

7.2.1.4 Consequence: Innovation Decay

Generative AI may find innovative new applications in information access. However, the excitement around these technologies and the significant investments from industry, government, and academia on corresponding research and development have broader implications for IR research. Next, we discuss some of the mechanisms associated with the research and development of generative AI that may potentially throttle innovation in information access technologies.

Mechanism: Industry Capture The compute and data moat that concentrates power in the hands of big tech, as discussed earlier in Sect. 7.2.1.2, also creates significant barriers to entry for academic research. These barriers limit academic AI research to a handful of institutions that have the necessary means and connections to industry who provide access to compute and data resources to incentivize research in areas of their economic interests. Academics who want to contribute to research on large-scale AI systems or critique their sociotechnical impacts are pressured to play well with institutions holding monopolistic control over compute, data, and systems [150]. Access to "open-access" models—without the compute and data necessary to build them from scratch—allows academic researchers to invest in finding more effective applications of these technologies that serve industry interests, but not to reimagine/rearchitect them to in radically different ways. Students and other academics who may someday want to work in industry are shepherded into integrating themselves into this homogenized research agenda.

Such "industry capture" [237] allows for inordinate influence of the sociotechnical imaginaries[13] of profit-driven corporations over, for example, academic researchers [146]. This can thwart research that may not be immediately monetizable or challenges the status quo of power concentration and complements the "regulatory capture" by bigger tech companies [13, 130, 182]. As Mitra [146] asks:

[13] Jasanoff and Kim [103] define *sociotechnical imaginaries* as "collectively held, institutionally stabilized, and publicly performed visions of desirable futures, animated by shared understandings of forms of social life and social order attainable through, and supportive of, advances in science and technology."

"*Whose sociotechnical imaginaries are granted normative status and what myriad of radically alternative futures are we overlooking?*" Narratives of the inevitability of these technologies that are hyped up to be both transformative forces for society and simultaneously posing existential risks for humanity (often purported by the same actors) only bolster their imagined importance to accumulate increasing global investments, including from governments. Researchers who care about sociotechnical impact and ecological sustainability are busy with enumerating the harms of rapidly emerging new AI technologies and chasing potential mitigations instead of having the full means to imagine and develop systems for social good. While industry practitioners can contribute to both identifying new research challenges grounded in real-world systems and practical methods to mitigate some of the risks of emerging technologies, it is imperative that we create avenues for increasing independent research while preserving the benefits of various modes of industry-academia collaborations.

Even as the grounded risks from these technologies (such as those discussed here) gather consensus from academic communities and civil society, it can be difficult to create space for alternative ways of development that are perceived as "slowing down." Critical research on sociotechnical harms of AI is also under risk when attempts are made to shift attention from concerns about real harms to marginalized people today to unsubstantiated imagined future concerns [71, 72]. Calls for regulations to address these imagined future harms [73] further detract from real progress and contribute to reinforcement of monopolistic powers of those who have already added these technologies to their arsenals. This has led some sociotechnical researchers in AI to explicitly draw attention to how these systems shift power (e.g., [23, 70, 107, 142]) and to prioritize research guided by alternative visions for sociotechnical futures grounded in universal emancipation and social justice [146]. It is thus important that access to investments to enable development is also available to those trying to not only mitigate existing systems' harms but also develop new avenues, including work on social good and new business models.

As generative AI starts to accumulate the lion's share of research investments, it may starve out other areas of information access research. Generative AI has had exciting but limited deployments in information access systems today. There are significant open challenges to making these models broadly useful, including but not limited to concerns of potential sociotechnical harms. There is a risk that if these challenges are not mitigated in spite of the extensive resources already invested on them at present, there may be calls for even larger investments in future prompted by the sunk cost fallacy.[14] It would be astute for the IR community to consciously continue to invest in research on systems and applications that societies need beyond what existing AI technologies make plausible [146, 184].

Mechanism: Pollution of Research Artefacts Risks to academic research from generative AI may also emerge through the applications of generative AI models

[14] https://en.wikipedia.org/wiki/Sunk_cost#Fallacy_effect

in IR scholarship—e.g., for authoring scientific papers and peer reviewing. There is evidence that researchers in computational sciences are already leveraging these tools [127], sometimes with hilariously terrible outcomes [163]. While the use of language models for light editing may (eventually) fall within the norms of socially acceptable behavior in research, their application in scholarship does raise concerns of plagiarism and scientific inaccuracies. This is an area that currently has more questions than answers, and the IR community would benefit from proactively considering potential implications of this trend on future IR research.

7.2.1.5 Consequence: Ecological Impact

Another important consequence of generative AI is its impact on the environment. In this context, it is important for us to consider the direct environmental cost of developing and deploying generative AI systems at scale as well as the potential impact of these technologies on the climate change discourse online.

Mechanism: Resource Demand and Waste The ecological cost of deep learning models has been a subject of much concern and debate in the AI community [16, 17, 25, 56, 108, 161, 162, 199, 241]. Similar concerns have also been raised within the IR community with respect to the application of these models for information access [181, 249]. By some estimates, the computing power being utilized for deep learning research has been doubling every 3.4 months since 2012 [32]. In the USA, data centers consumed more than 4% of the total national electricity in 2022, and that number is projected to grow to 6% by 2026 [87]. Another study [15] estimates that by 2040, the information and communications technology industry on the whole will account for 14% of global emissions. Beyond emissions, data centers' water consumption is also raising alarm bells [50, 81, 84, 86, 89, 124, 152, 173]. By 2027, global AI demand may be responsible for withdrawal of 1.1–1.7 trillion gallons of fresh water annually [89, 124]. Serious concerns also revolve around the rising levels of electronic waste [112]. Even as we make progress in reducing the ecological cost of training and deploying the current AI models, we risk encouraging the development of even larger models and their wider deployment worsening the overall ecological impact (i.e., Jevons paradox).[15]

Mechanism: Persuasive Advertising Generative AI may not only negatively impact the environment through increasing demand for natural resources and increasing generation of waste but may also supercharge climate change disinformation [43, 55, 68, 174, 175, 195]. For example, the fossil-fuel industry may attempt to sway public opinion through advertising that leverages generative AI's persuasion capabilities discussed in Sect. 7.2.1.2. Persuasive advertising may also be employed by other environment-unfriendly business models like fast fashion [42].

[15] https://en.wikipedia.org/wiki/Jevons_paradox

While the direct ecological cost of generative AI justifiably garners lots of attention, its potential impact on related online discourse also deserves scrutiny.

7.2.2 Risks

We categorize the risks of generative AI broadly to our society, to IR research, and to the environment. We map the first three consequences discussed earlier in this section—i.e.: (i) Information ecosystem disruption (Sect. 7.2.1.1) (ii) Concentration of power (Sect. 7.2.1.2) (iii) Marginalization (Sect. 7.2.1.3)—and their corresponding mechanisms as potentially contributing to the risks to society. We further map the last two consequences—i.e.: (iv) Innovation decay (Sect. 7.2.1.4) (v) Ecological impact (Sect. 7.2.1.5)—to the risks to IR research and the environment, respectively.

7.2.2.1 Risks to Society

Information access is a critical need of any democratic society and a necessary ingredient for social transformation [44, 78, 80, 96, 168]. It is also a social determinant of economic progress [151, 245] and health [147]. Disruptions to the information ecosystem bear potentially grave risks to most aspects of our social lives. A confluence of the pandemic [35, 183, 209], rising global conflicts [210, 213], and escalating climate catastrophes [102, 160, 169] is pushing the world toward precarious instability. Our information ecosystems are already struggling under the weight of misinformation and disinformation that in this critical moment is eroding public trust in online platforms, institutions, and each other. It is imperative that researchers and developers of information access systems prioritize safeguarding social interests and be vigilant in considering potential risks of disruption and ecosystem collapse when integrating generative AI technologies in the IR stack. This includes identifying the necessary conditions under which these technologies can be safely deployed and developing practical safeguards and alternatives.

Risks to society are not just from potential disruptions of the information ecosystem but also from how these technologies simultaneously concentrate power away from those at the margins of society. As institutions that develop and operate these technologies are themselves beneficiaries of this concentration, we need democratic oversights. If technologies further exacerbate already worsening wealth and power inequities, this additionally may pose severe threats to democratic institutions and human rights. There is an opportunity cost of not re-imagining information access in light of sociotechnical ambitions of human emancipation, culture, and knowledge production, instead of being constrained solely by what these emerging technologies make plausible and the homogenized visions put forth by institutions who wield these technologies [146].

7.2.2.2 Risks to IR Research

IR research can suffer from a confluence of different factors including the distancing of academic researchers from the data and compute they need to do their work and how narratives about the inevitability of AI technologies shapes what computational research gets funded. The concentration of access to the networks around these technologies in a subset of institutions shapes what is considered "foundational" or even "AI." Research on generative AI should not be performed only in the context of corporate economic interests while academia is hollowed out and prevented from exploring radical new methods that challenge the status quo. This risk of homogenization of academic research agendas and the opportunity cost of not exploring more diverse approaches to online information access can have material consequences. Instead, the IR community must be empowered with both the space and the resources necessary to explore a diversity of these visions and critique dominant narratives. IR research should have a plurality of work, which includes work with access to industry to change current practices. However, we especially also need to ensure that not all IR research is simply an extension of industrial system development and risk the demise of fundamental research on alternative avenues.

7.2.2.3 Risks to the Environment

Information access provides one of the large-scale application settings for generative AI. However, the impact of such wide-scale deployment of these technologies on the impending climate crisis should be a critical consideration. Climate costs pose substantial existential risks for ecosystems and people, in more direct ways than some other "existential risks" that lack adequate scientific basis but have nonetheless been popular discourse in some parts of the AI community. This means both choosing what to deploy and investment in methods to mitigate negative impacts that build on existing environmental work. As we discussed in Sect. 7.2.1.5, these concerns include not just the ecological cost of developing and deploying generative AI technologies but also their impact on online discourse on societal priorities.

7.3 Methods to Evaluate Risks and Impact

7.3.1 Evaluating the Impact of Generative IR Applications

Evaluating the impact of generative IR applications requires methods, as do data-informed interventions to steer that impact. Creating an LLM-based demo has become exceedingly easy. Understanding the impact of a system when it gets used in real-life contexts, and getting to a high-quality experience for a wide

variety of users, is much harder. Standards for impact assessment have not kept up a similar pace as tech developments. Khlaaf points out the need to carefully consider the differences in value alignment of the goals of a system and safety considerations, harms, and risks [113]. A wide range of online, offline, and human-assisted evaluations are possible—and necessary—to get a full sense of the impact of a system.

There are a number of frameworks that can provide helpful starting points for evaluating the impact of generative IR applications and potential quality or safety improvements. Not surprisingly however, they can measure quite different aspects of a system and its underlying models. Distinctions have to be made between evaluating a model, a system, or a technology as a whole. For example, standards for foundation model evaluations might not take into account the impact of a system that uses such a model (or a combination of models) in a specific application context.

Measurement and interventions are possible at every stage of the development life cycle of products and their underlying models and data. In this regard, general insights around, for example, harm mitigation interventions being possible throughout the machine learning life cycle [201] also apply to generative IR. To improve quality and safety, we need to be able to operationalize and measure the impact of potential interventions. This includes evaluations on aspects of that might be both system performance issues but also of societal importance, e.g., harmful/toxic output, hallucination, and differing model performance across languages/demographics.

7.3.2 Threat Identification, Assessment, and Modeling

When the emergence of a new technology or application becomes apparent, the assessment of whether this poses risks or opportunities within specific domains poses a challenge. Before development of a system, threats and opportunities can be identified. As Kapoor et al. [109] point out, it is crucial not to evaluate the risks and impact of new systems in isolation but rather in comparison with existing technologies. For example, the impact of usage of foundation models in *search* should be compared to existing Web *search*. For this purpose, Kapoor et al. present an evaluation framework that focuses on marginal risks, applied to Open Foundation Models. Their framework is based on threat identification work from cybersecurity and consists of six steps necessary to demonstrate such marginal risk. These steps are (1) threat identification, (2) evaluating existing risk absent open foundation models, (3) considering existing defenses absent open foundation models, (4) evidence of marginal risk of open foundation models, (5) ease of defending against new risks, and (6) outlining uncertainty and assumptions. Note that this framework does not set exact assessment criteria but rather defines the steps to get to such evaluations.

In practical settings, this might mean having to select standards for the development process (e.g., emerging standards from organizations such as the National

Institute of Standards and Technology (NIST) [3] or International Standards Organization (ISO) [12], company-specific standards such as Microsoft's Responsible AI Standard v2 General Requirements [143], or following new (local) legal requirements). However, mapping out potential consequences and identifying mechanisms that introduce risks in the specific context of a system needs to go much further. How to disrupt potential negative mechanisms in order to mitigate those risks requires gauging a wide range of consumer-side impacts [61] but also wider societal impacts. That includes frameworks focused on worker consequences [70] or practical methods focused on reducing the (legal) risks of using certain types of copyrighted or restricted training data vs. expected performance gains [144].

7.3.3 Evaluation During Model Development

7.3.3.1 Model Benchmarks vs. Actual System Context

LLM benchmarks are widely used to compare the quality and safety progress made by new *model* releases, resulting in model leaderboards on different scenarios. The Stanford HELM [198] leaderboard, for example, shows the performance of different LLM models on benchmarks, and these benchmarks include societal impact and bias-related measures. Their HELM ("holistic framework for evaluating foundation models") framework [126] uses scenarios and measures seven metrics. Those are accuracy, calibration, robustness, efficiency, and also more social impact-oriented fairness, bias, and toxicity. Each scenario focuses on one use case and consists of a dataset of instances, such as the LegalBench set of legal reasoning tasks [85] or medical board exam problem sets [104]. The larger BIG-bench ("Beyond the Imitation Game benchmark") [18] consists of 200+ tasks, contributed by hundreds of authors at a variety of institutes. More specific benchmarks for trustworthiness such as DecodingTrust, in turn, focus on subsets such as toxicity, stereotyping, adversarial and out-of-distribution robustness, privacy, machine ethics, and fairness [230], while, for example, the much more specific recurring TREC Fair Ranking track competitively evaluates systems according to how fairly they rank documents on a specific test task [60].

Paradoxically, while these benchmarks include aspects of societal impacts such as bias and toxicity, they do not necessarily cover the aspects that matter most in a specific application context in practice. Benchmarks are generally geared toward structured comparisons between models, *not* toward evaluating end-user applications in practice. This means that they may not be particularly suitable for a specific application and the people involved in its usage. In addition, using such large benchmarks can be quite resource-intensive, making "lite" versions necessary that are less comprehensive. Both HELM and BIG-bench are also implemented as Lite versions. However, the evaluation differences that arise from specific, lighter implementations of benchmarks can significantly impact model comparison results [196]. This makes it necessary to go beyond these benchmarks and ensure suitable

evaluations for the application at hand to avoid deriving conclusions about safety or responsibility devoid from actual application concerns.

7.3.3.2 Combining IR and Generative AI Evaluation Metrics

It is challenging that standards for measuring societal impact, including bias, fairness, etc., are yet scarce in IR *product* settings. For example, Smith et al. [193] provide an overview of different metrics available for evaluating bias and fairness in recommendation systems and the challenges practitioners face when choosing between them. In some cases, it may be more appropriate to, for example, focus on "traditional" performance and accuracy metrics but study the performance and subsequent quality of experiences for different groups of people by segmenting/slicing results by group. This approach assumes the ability to define relevant groups or relies on more advanced methods to find clusters that may—or may not—have significant differences in performance or quality.

Specific methods might also be necessary to match new techniques. For example, Retrieval-Augmented Generation (RAG) might be used to include more reliable information in a specific domain and reduce hallucinations in an LLM setting. However, RAG does not necessarily fully solve every hallucination-related issue. Specific frameworks that fit an application context are still necessary to evaluate these techniques and their actual impact on aspects such as factuality within that context. One example is Saad-Falcon et al. [178], who present an evaluation framework, ARES, for RAG-assisted question-and-answering settings. This framework uses three evaluation scores: context relevance of the retrieved information, answer faithfulness (the answer's grounding in the retrieved context), and answer relevance to the question asked. These are similar to IR evaluations but might need adjustment to the setting at hand, and datasets used need to reflect actual needs in current circumstances.

7.3.3.3 LLMs to Evaluate LLM

Beyond specific metrics, ongoing research is investigating the efficacy of LLMs to evaluate LLMs (*LLM-as-judge*) [187, 232]. For example, [232] et al. use an LLM to rate the factuality of a long-form response to prompts while also using Google Search. While promising, such more complex evaluation constellations also lead to additional complexity in understanding what is being evaluated and changes therein as the evaluator LLM changes. This leads to having to validate the validation in itself [187]. While a human-and-LLM agent collaboration can help in this validation (as in, e.g., [187]'s EvalGen approach), the evaluation criteria cannot be fully separated from observation of model outputs, resulting in a feedback loop from output to adjusted evaluation criteria.

7.3.4 Evaluation Pre-/Post-system Release

7.3.4.1 Online Evaluation Using Actual User Behavior vs. Offline Evaluation

Whether evaluations are done online or offline can deeply impact results. Offline evaluations—even when using thoughtful standards—might not reflect what actual end users do in real-life settings or system performance over time. Online evaluations similarly are limited to which metrics have been instrumented and how actual user interactions are captured. It involves field testing, getting an IR system online and out to actual users, and analyzing their interactions with the system. It can include methods such as controlled experiments or extended A/B testing and analysis of interactions. Hoffmann provides an overview of the most common techniques used in IR settings [100].

7.3.4.2 Stress Testing, Red Teaming, and Qualitative End-User Evaluations

Beyond metrics and quantitative analysis–oriented methods, it is crucial to apply a combination of safety/security-inspired methods, user design, and User Experience (UX) research methods to understand the actual reactions of users.

The logistics around red teaming can provide a good glimpse into the importance of appropriate combinations of methods. Red teaming is a common way to test LLM applications for undesirable system responses [135, 145]. Red teaming can be automated using, for example, sets of (generated) prompts or done in full by human red teamers, including both the general public and invited experts. Using LLMs as red teamers [164] by generating risky prompts at scale, or using large-scale human red teaming efforts with thousands of participants who need access points, might yield different results. Human red team approaches in which "a group of people authorized and organized to emulate a potential adversary's attack or exploitation capabilities against an enterprise's security posture" (if we follow the definition from NIST) also lead to questions about tooling, recruiting, and operational process design. Markov et al. [135], for example, provide a helpful discussion of practical data challenges in content moderation use cases. In turn, model characteristics might have consequences on red teaming results. Ganguli et al. [69], for instance, find that RLHF models are increasingly difficult to red-team as they scale, while they do not find similar challenges for other models. Interestingly, this means that techniques such as RLHF that are explicitly meant to help align agents with human preferences could also result in challenges in evaluating the systems that use them.

This means that like any evaluation method, red teaming has to be combined with other types of stress testing, assessment of security issues, as well as evaluation of experiences of actual users. Khlaaf points out the need for carefully considering

what methods and terminology are appropriate for evaluations that probe for vulnerabilities of a specific system toward the outside world [113].

7.3.5 Societal Impact of a System Beyond Its Direct Implementation and Use

The impact of a system can reach much beyond its direct usage context. For example, the increasing demand for data and compute power of LLMs has environmental impact. However, such indirect impact can be hard to calculate without deep expertise. It is crucial to spend the time to evaluate evaluations methods for their suitability. Methods have been developed in both the IR and LLM communities around reducing environmental harm [181], and sustainability industry teams exist to ensure more energy-efficient data centers for both environmental and monetary reasons. Others in turn try to assess whether LLMs could help in generating more green code and develop metrics to assess the code's "green capacity" based on earlier sustainability metrics [216].

Similarly, a plethora of work points out the potential of amplifying and entrenching power structures through the usage of generative AI methods or changing market conditions through releasing new models for free [176], de facto changing standards to the model that gets used most in practice. However, IR and Machine Learning (ML) evaluation methods are not generally suitable for the analysis of such impact that a particular technique or system might have. Methods from political analysis and behavioral economics might be more suitable but are generally not shared in IR or ML venues. Challenging in the evaluation of systems is a deeper understanding of the long-term incentives that are created and the resulting "rational" use of LLMs in undesirable ways. A compounding challenge is that new incentives are also necessary to ensure that interventions from actual practice can be shared. Trust and safety teams might be doing scenario planning or prepare for incidents and crises.

7.3.6 Sharing Evaluation Methods

From the above selection of methods, which is by no means comprehensive, it is clear that practitioners have to carefully pick and choose which methods work for them. However, different organizations come from different evaluation traditions.

Incentives to share methods and results might not align with practical product team incentives and pressures. Metrics and standards for evaluations from actual practice are often not shared in scientific literature. Security community-style (external) red and (internal) blue teams, trust and safety incident monitoring approaches, IR communities' existing offline and online user feedback methods, or UX product testing approaches might be more (or less) top of mind, depending

Table 7.2 Different types of existing evaluation frameworks relevant for generative IR impact and safety. Note that this is not an exhaustive overview but rather a quick peek at the variety of methods evaluators can (and have to) choose from

Evaluation focus	Examples
Marginal system impact, e.g., release decisions in comparison with existing technology	Kapoor et al., risk framework based on cybersecurity [109]
Comparison benchmarks between LLM models that include fairness, bias, toxicity-type aspects	Benchmarks used in leaderboards, e.g., HELM [198], BIG-bench [18], or trustworthiness benchmarks [230]
Online or offline IR metrics, including accuracy or quality across groups	Online IR-evaluation methods [99], impact/fairness/bias metrics in recommendation systems [192]
Evaluation metrics using automated evaluation for specific LLM techniques or risks	LLMs as agents evaluating factuality of other LLMs' statements [187, 232]
Qualitative evaluation including human adversarial testing	Red teaming[69, 164] and UX evaluation

on the organization and prior expertise. This means there is a gap in the generative IR literature in terms of shared understanding of actual practices and efficacy of methods [48]. If we as a community are to properly address the social risks as outlined in Sect. 7.2.2.1, it is imperative we find fast and effective ways to share these methods and align them with practical needs, especially with the increasing speed of the field, the variety of fields involved, and volume of new techniques (Table 7.2).

7.4 Actors, Incentives, and Ways of Getting Organized

7.4.1 Incentives Toward Misuse of AI

Emerging AI capabilities and their consequences (good or bad) are a hot topic of discussion. But it is just as important to talk about incentives or why individuals or organizations might choose to use AI in certain ways.

Below are some examples of types of actors and their possible incentives that can lead to harmful uses of AI, along with ways in which some of them can be shifted in a more positive direction. AI can be transformative for human experience and quality of life, but only if incentives (both short term and long term) for its use are aligned with the benefits to humanity.

Actor State actors and ideological groups.

Incentive Geopolitical influence in favor or against something. This includes the use of extra-persuasive [238], micro-targeted content and deepfakes to sow malicious narratives [191], undermine support and trust in democratic institutions [148], weaken social cohesion, etc.

Modification The most effective way to modify this behavior is by making it prohibitively expensive or inconvenient to use AI for these purposes, through harsh legal consequences, content moderation, or counter-speech. The burden of implementing countermeasures falls on governments, content platforms, and community organizations.

Actor Criminal or unscrupulous organizations.

Incentive Financial gains from scams, ad-monetized Web site traffic, or product sales. This includes more legit-looking phishing content [229] and "Nigerian prince" letters or gaming search engines via AI-generated SEO-friendly content [156].

Modification The incentives for financial gain are always going to exist and be exploited; protection against them can take the form of better (AI-enhanced) cybersecurity and anti-spam tools, implemented and deployed by most consumer-facing Web surfaces.

Actor Commercial enterprises.

Incentive Economic competitive advantage and increased shareholder value. Taken to its worst extreme, this incentive can lead to deceptive or discriminatory business practices, hasty deployment of cheaply developed AI to customers [65], premature restructuring of teams [121], etc. In the case of social media platforms, the high engagement on polarizing or sensationalist content can lead the platforms to tolerate, encourage, and algorithmically amplify it.

Modification The same drive for competitive advantage can also be a force for good, particularly when it is aligned with public opinion or customer sentiment. The best-case scenario is when trustworthy and safe AI makes products more usable, attracting more customers (akin to Apple's "It just works" aesthetic that has no shortage of fans despite being more expensive than the competition). Government-led compliance requirements can also create positive incentives, like for food or car safety, and in some cases, a punitive legal strategy also works, like in the suing of tobacco companies or opiate producers, creating incentives for surviving companies to behave better.

Actor Individuals.

Incentives Faster completion of work tasks, improved social status, revenge against perceived slights, or exploitation of the vulnerable. At worst, these can lead to cheating, misrepresentation of one's identity of accomplishments, slander, deepfake pornography, or AI-enhanced grooming.

Modifications While some of these behaviors are illegal or fundamentally antisocial (and should be prosecuted as dictated by law), the urge to improve one's work performance or social status can be a good thing. If AI tools are designed to enhance human productivity while rewarding our creative impulses, and feel fun, joyful, and satisfying to use, people will be more likely to employ them to good ends.

7.4.2 Who Can Shift Incentives and How

In the broadest sense, it will take a whole-of-society approach to ensure that technological advances will align with the best interests of humans impacted by them (see Fig. 7.1). Technology builders (company and individual), governments, academia, and civil society all bear responsibility for ensuring that technological advances in information access align with societal interests. The rest of this section focuses on what can be done at the intersection of these groups or actors, since inter-group coordination is most often where things go awry.

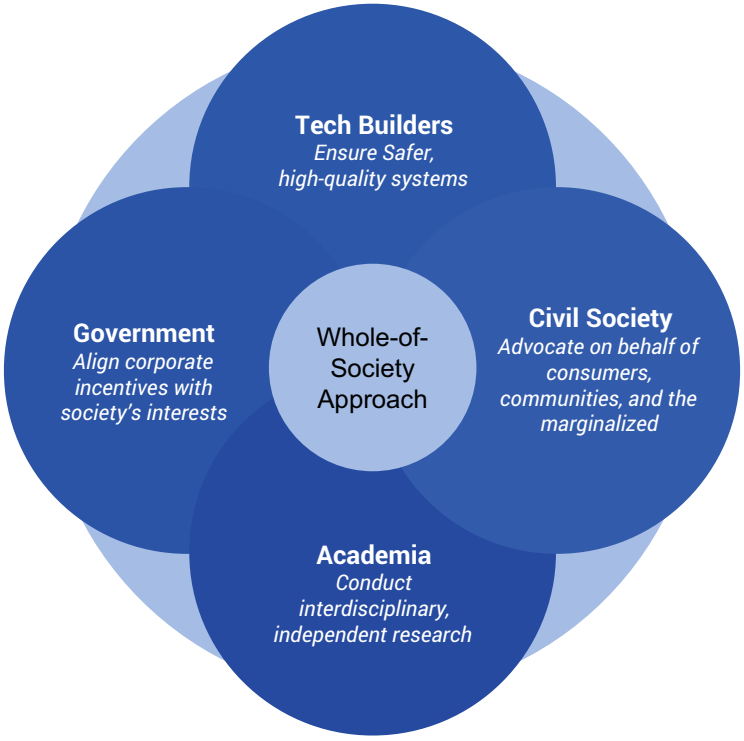

Fig. 7.1 Primary actors responsible for aligning technology with societal interests

7.4.2.1 Organizational Factors

While most of the literature and education in computer science by definition focuses on technical approaches, the impact of generative IR techniques can be influenced in other ways as well.

Changing work processes *within* organizations can have a direct impact on the expectations set on teams. This includes policies, explicit Go/No-Go procedures, roles and responsibilities to monitor systems, algorithmic impact assessments and model cards, or other types of documentation. In different organizations, the responsibility for different measurement and mitigation might look very different. In one organization, a machine learning team may be expected to look at the energy consumption of their system design choices, whereas other organizations might have a technical sustainability team. In another organization, a trust and safety or integrity team might deliver evaluations of system output toxicity, whereas in another organization, a separate data science team or product teams themselves might have to do this work. In any case, if this responsibility is unclear, it is much harder to get this work done.

External engagement can help address internal deficiencies. Especially for audiences working on generative IR systems, some of these might not necessarily be familiar routes. Examples include:

- **External advice and safety boards.** These are increasingly created by companies to provide external advice for more complex safety or content moderation questions. These include Facebook's Oversight Board,[16] which provides independent rulings on content moderation questions, parent company Meta's Safety Advisory Council,[17] or Spotify's Safety Advisory Board.[18] These do not necessarily have decision-making power but provide a more formalized way to advise external organizations and researchers.
- **Regulatory advisory groups and expert consultations.** Organizations such as the UN, EU, various regions, and countries working on future AI policy have all formed advisory boards (e.g., the UN AI advisory board,[19] the Nordic AI advisory board). Apart from such official avenues, individual lawmakers and legal firms often consult experts. While regulatory capture is a very real concern [236], this also allows for actually implementable regulation. This means however that considering the potential overlap between advisory boards, as well as perhaps a lack of overlap with more specific AI experts, not all relevant expertise will be represented.
- **Professional organizations.** Organizations such as Association for Computing Machinery (ACM), Institute for Electrical and Electronics Engineers (IEEE),

[16] https://www.oversightboard.com

[17] https://www.facebook.com/help/222332597793306/

[18] https://newsroom.spotify.com/2022-06-13/introducing-the-spotify-safety-advisory-council/

[19] https://www.un.org/en/ai-advisory-body

Association for the Advancement of Artificial Intelligence (AAAI), or the Trust & Safety Professional Association allow for formal and informal exchange of best practices. A major challenge is ensuring that best practices in fast-moving areas are also gathered and exchanged *between* organizations and to the public at large.

For the above arrangements, getting to collections of concrete examples of what has worked in the past is increasingly important. AI developments are speeding up, and increasingly diverse professional communities are both being impacted and getting involved. This makes efficient and effective coordination even more important. For policymakers, governmental agencies, and journalists, it may be hard to get an overview of which professional communities can provide actionable advice—especially with new AI developments being "louder" than, for example, long-standing IR communities. Inside of companies, in order to benefit from external advice or research, tech teams still have to navigate how to best work with external organizations. Researchers and non-governmental organizations in turn have to know where to invest their time and expertise most effectively and how to offer actionable advice to appropriate individuals or teams in tech companies. This includes big-picture scenario planning of where to best invest and how to create incentives that truly will have a positive impact. Implicit hierarchies of the value of different types of produced knowledge (e.g., "being the first" or "more technically complex"), but also a simple lack of knowledge about how certain processes work, can stand in the way of sharing of paved paths toward desired results and of sharing these in accessible ways. It can also involve very pragmatic on-the-ground work, such as knowing how to set up contractual arrangements that work for all parties (not a skill commonly taught in IR or AI-related programs).

7.4.2.2 Data-Focused Methods

While a complete overview of all different mechanisms to positively affect AI development is outside the scope of this chapter, one area does provide ample inspiration. Extensive literature exists on data labor and the need to understand how to effectively advocate for that labor's value [11, 75, 123, 218, 223]. Especially in the realm of training data concerns, multiple practical routes already exist, including:

- **Business and partnership model development,** including developing new types of licensing and new types of business partnerships [9, 190], along with ways to get funding to data creators. There is also research on the efficacy of suggested mechanisms, such as data dividends that are suggested as a means of AI profit sharing [227].
- **Collective action**. When new business models do not work out, coordinated action is imperative. These can be focused on data through data strikes [226], as well as large-scale labor organizing and strikes focused on treatment of data workers. More recently, the Hollywood strikes illustrated how those particularly impacted by the ways their work and likeness can be used as data can effectively

organize, lay out clear demands, and succeed through both technical and organizational competence. This included understanding what incentives are at play and what leverage data producers have [222]. Methods include data strikes to withhold data [205, 226] and data poisoning [54] techniques such as NightShade [94, 186], Glaze [185] and Mist [125]. Ways to empower end users and the wider public in their relationship with tech companies are important [228], as is understanding their potential leverage and means for protest through adjusted usage [122].

For effective research-informed mitigations, however, it is crucial that generative IR researchers have access to ways to learn how to effectively organize and navigate organizational and political structures or how to communicate their results to others. Implicit hierarchies in what knowledge is appreciated in generative IR circles can become a hurdle in effectively identifying and addressing the risks outlined in earlier sections, Sects. 7.2.2.1, 7.2.2.2, and 7.2.2.3. A critical factor is knowing which concrete situations matter, what to ask for in those situations, and how to assess whether impacts and risks are successfully steered.

7.5 Conclusion

In this chapter, we have presented a discussion on the sociotechnical implications of generative AI for information access. These deliberations are grounded in how these emerging technologies are currently being applied in IR applications as well as their future applications as being envisioned by practitioners and researchers. It is important to recognize that sociotechnical visions of what information access should look like in the future are not just shaped by what emerging technologies like generative AI make plausible but that visions for the future of information access in turn shape AI technologies themselves. Mitra [146] proposed the hierarchy of IR stakeholder needs shown in Fig. 7.2 and argued that IR research and system development require a fundamental shift toward re-centering societal needs and that we should reimagine information access as a vehicle for alternative futures. When contemplating the implications of emerging technologies, we risk of falling in the trap of limiting ourselves to how the technology (and its process of development) is today, rather than how it can be or *should be* in the future. Neither generative AI nor its application in the context of information access is predetermined. So while it is important that we consider potential harms of contemporary applications of generative AI in the context of information access, we close with some open question for the reader: *If not this status quo, then what—and especially how? What is the future of information access that we want to imagine for our collective well-being, and how can generative AI be another tool in the toolbox toward that transformation?*

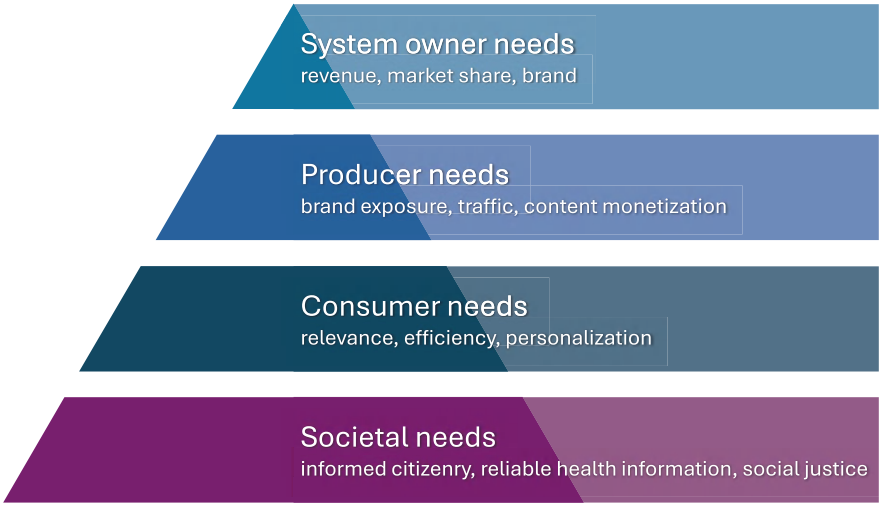

Fig. 7.2 Mitra's [146] hierarchy of IR stakeholder needs. More critical needs are at the bottom of the pyramid. This figure has been reproduced from the original paper with permission

References

1. Abid, A., Farooqi, M., Zou, J.: Persistent anti-Muslim bias in large language models. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 298–306 (2021)
2. Agüera y Arcas, B., Mitchell, M., Todorov, A.: Physiognomy's new clothes (2017). https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a
3. AI standards, NIST (2024). https://www.nist.gov/artificial-intelligence/ai-standards
4. Ai used to target kids with disinformation (2023). https://www.bbc.co.uk/newsround/66796495
5. Al-Sibai, N.: The top google image for Israel Kamakawiwo'ole is AI-generated. Futurism (2023)
6. Alemohammad, S., Casco-Rodriguez, J., Luzi, L., Humayun, A.I., Babaei, H., LeJeune, D., Siahkoohi, A., Baraniuk, R.G.: Self-consuming generative models go mad (2023). arXiv preprint arXiv:2307.01850
7. Altenried, M.: The platform as factory: Crowdwork and the hidden labour behind artificial intelligence. Capital Class **44**(2), 145–158 (2020)
8. Anguiano, D., Beckett, L.: How Hollywood Writers Triumphed Over AI–and Why It Matters. The Guardian (2023)
9. AP, OpenAI agree to share select news content and technology in new collaboration, 2023. https://www.ap.org/media-center/press-releases/2023/ap-open-ai-agree-to-share-select-news-content-and-technology-in-new-collaboration/
10. Appel, G., Neelbauer, J., Schweidel, D.A.: Generative AI has an intellectual property problem. Harvard Bus. Rev. **7** (2023)
11. Arrieta-Ibarra, I., Goff, L., Jiménez-Hernández, D., Lanier, J., Weyl, E.G.: Should we treat data as labor? moving beyond "free". In: AEA Papers and Proceedings, vol. 108, pp. 38–42. American Economic Association, Nashville (2018)
12. Artificial intelligence (AI) standards. https://www.iso.org/sectors/it-technologies/ai

13. Asokan, A.: UK government warned of AI regulatory capture by big tech. BankInfoSecurity (2024)
14. Baio, A.: 'most disturbing AI site on Internet' can find every picture of you that exists. Indy100 (2024)
15. Belkhir, L., Elmeligi, A.: Assessing ICT global emissions footprint: Trends to 2040 & recommendations. J. Clea. Prod. **177**, 448–463 (2018)
16. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: can language models be too big?🦜. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (2021)
17. Berreby, D.: As use of A.I. soars, so does the energy and water it requires. Yale Environment 360 (2024)
18. BIG-bench authors: Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. Trans. Mach. Learn. Res. (2023). ISSN 2835-8856. https://openreview.net/forum?id=uyTL5Bvosj
19. Birhane, A.: Algorithmic colonization of Africa. SCRIPTed **17**, 389 (2020)
20. Birhane, A., Cummins, F.: Algorithmic injustices: towards a relational ethics (2019). arXiv preprint arXiv:1912.07376
21. Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., Bao, M.: The values encoded in machine learning research. In: 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 173–184 (2022)
22. Blodgett, S.L., Green, L., O'Connor, B.: Demographic dialectal variation in social media: a case study of African-American English (2016). arXiv preprint arXiv:1608.08868
23. Blodgett, S.L., Barocas, S., Daumé III, H., Wallach, H.: Language (technology) is power: A critical survey of" bias" in nlp (2020). arXiv preprint arXiv:2005.14050
24. Bolukbasi, T., Chang, K.W., Zou, J.Y., Saligrama, V., Kalai, A.T.: Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: Advances in Neural Information Processing Systems, vol. 29 (2016)
25. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al.: On the opportunities and risks of foundation models (2021). arXiv preprint arXiv:2108.07258
26. Brewster, J., Arvanitis, L., Sadeghi, M.: Funding the next generation of content farms: some of the world's largest blue chip brands unintentionally support the spread of unreliable AI-generated news websites (2023). NewsGuard
27. Brewster, J., Fishman, Z., Xu, E.: Funding the next generation of content farms: some of the world's largest blue chip brands unintentionally support the spread of unreliable AI-generated news websites (2023). NewsGuard
28. Burke, K.: 'biggest act of copyright theft in history': thousands of Australian books allegedly used to train AI model. The Guardian (2023)
29. Burke, K.: Generative AI is a marvel. is it also built on theft? The Economist (2024)
30. Burtell, M., Woodside, T.: Artificial influence: an analysis of AI-driven persuasion (2023). arXiv preprint arXiv:2303.08721
31. Caliskan, A., Bryson, J.J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. Science **356**(6334), 183–186 (2017)
32. Cantrell, T.: The True Cost of AI Innovation. Scientific Computing World
33. Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., et al.: Extracting training data from large language models. In: 30th USENIX Security Symposium (USENIX Security 21), pp. 2633–2650 (2021)
34. Carroll, M., Chan, A., Ashton, H., Krueger, D.: Characterizing manipulation from AI systems. In: Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, pp. 1–13 (2023)
35. Centers for Disease Control and Prevention et al.: CDC museum COVID-19 timeline. 2022, 2022
36. Chayka, K.: Is A.I. Art Stealing from Artists? The New Yorker (2023)

37. Chesterman, S.: Good models borrow, great models steal: intellectual property rights and generative AI. Policy Soc. puae006 (2024)
38. Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S., Amodei, D.: Deep reinforcement learning from human preferences. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
39. Coffey, D.: Māori are trying to save their language from big tech. Wired UK (2021)
40. Cohan, W.D.: Ai is learning from stolen intellectual property. It needs to stop. The Washington Post (2023)
41. Coldewey, D.: Thousands of authors sign letter urging AI makers to stop stealing books. TechCrunch (2023)
42. Coleman, J.: AI's climate impact goes beyond its emissions. Scientific American (2023)
43. Corbett, J.: Report warns generative AI could turbocharge climate disinformation. Common Dreams (2024)
44. Correia, A.M.R.: Information literacy for an active and effective citizenship. In: White Paper prepared for UNESCO, the US National Commission on Libraries and Information Science, and the National Forum on Information Literacy, for use at the Information Literacy Meeting of Experts, Prague, The Czech Republic (2002)
45. Couldry, N., Mejias, U.A.: Data colonialism: Rethinking big data's relation to the contemporary subject. Television New Media **20**(4), 336–349 (2019)
46. Cox, J.: Google news is boosting garbage AI-generated articles. 404 Media (2024)
47. Coyle, J.: In Hollywood writers' battle against AI, humans win (for now) (2023)
48. Cramer, H.: Practical routes in the UX of AI, or sharing more beaten paths. Interactions **29**(5), 89–91 (2022). ISSN 1072-5520. https://doi.org/10.1145/3555834
49. Crawford, K.: The trouble with bias. In: Conference on Neural Information Processing Systems, Invited Speaker (2017)
50. Criddle, C., Bryan, K.: Ai boom sparks concern over big tech's water consumption. The Conversation (2024)
51. Cuconasu, F., Trappolini, G., Siciliano, F., Filice, S., Campagnano, C., Maarek, Y., Tonellotto, N., Silvestri, F.: The power of noise: redefining retrieval for rag systems (2024). arXiv preprint arXiv:2401.14887
52. del Rio-Chanona, M., Laurentsyeva, N., Wachs, J.: Are large language models a threat to digital public goods? Evidence from activity on stack overflow (2023). arXiv preprint arXiv:2307.07367
53. Dhawan, S.: Universities leveraging AI detectors: international students fear they may be wrongly accused of cheating. Financial Express (2023)
54. Dickson, B.: What is machine learning data poisoning? The Verge (2020)
55. Disinformation, Climate Action Against: Artificial intelligence threats to climate change (2024)
56. Dodge, J., Prewitt, T., Tachet des Combes, R., Odmark, E., Schwartz, R., Strubell, E., Luccioni, A.S., Smith, N.A., DeCario, N., Buchanan, W.: Measuring the carbon intensity of AI in cloud instances. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 1877–1894 (2022)
57. Dupré', M.H.: Sports Illustrated published articles by fake, AI-generated writers. Futurism (2023)
58. Dupré', M.H.: Top Google result for "Edward Hopper" an AI-generated fake. Futurism (2023).
59. Duran, G.: The tech baron seeking to "ethnically cleanse" San Francisco. The New Republic (2024)
60. Ekstrand, M.D., McDonald, G., Raj, A., Johnson, I.: Overview of the trec 2021 fair ranking track. In: The Thirtieth Text REtrieval Conference (TREC 2021) Proceedings (2022)

61. Ekstrand, M.D., Beattie, L., Pera, M.S., Cramer, H.: Not just algorithms: strategically addressing consumer impacts in information retrieval. In: Advances in Information Retrieval: 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24–28, 2024, Proceedings, Part IV, pp. 314–335. Springer, Berlin, Heidelberg (2024). ISBN 978-3-031-56065-1. https://doi.org/10.1007/978-3-031-56066-8_25
62. El-Sayed, S., Akbulut, C., McCroskery, A., Keeling, G., Kenton, Z., Jalan, Z., Marchal, N., Manzini, A., Shevlane, T., Vallor, S., et al.: A mechanism-based approach to mitigating harms from persuasive generative AI (2024). arXiv preprint arXiv:2404.15058
63. Elish, M.C.: Moral crumple zones: cautionary tales in human-robot interaction (pre-print). Engaging Science, Technology, and Society (pre-print) (2019)
64. Ferrara, E.: Should ChatGPT be biased? Challenges and risks of bias in large language models (2023). arXiv preprint arXiv:2304.03738
65. Fowler, J.A.: Turbotax and h&r block now use AI for tax advice. It's awful. The Washington Post (2024). https://www.washingtonpost.com/technology/2024/03/04/ai-taxes-turbotax-hrblock-chatbot/
66. Gabriel, I.: Artificial intelligence, values, and alignment. Minds Mach. **30**(3), 411–437 (2020)
67. Gabriel, I., Ghazavi, V.: The challenge of value alignment: from fairer algorithms to AI safety (2021). arXiv preprint arXiv:2101.06060
68. Galaz, V., Metzler, H., Daume, S., Olsson, A., Lindström, B., Marklund, A.: AI could create a perfect storm of climate misinformation (2023). arXiv preprint arXiv:2306.12807
69. Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K., Jones, A., Bowman, S., Chen, A., Conerly, T., DasSarma, N., Drain, D., Elhage, N., El-Showk, S., Fort, S., Hatfield-Dodds, Z., Henighan, T., Hernandez, D., Hume, T., Jacobson, J., Johnston, S., Kravec, S., Olsson, C., Ringer, S., Tran-Johnson, E., Amodei, D., Brown, T., Joseph, N., McCandlish, S., Olah, C., Kaplan, J., Clark, J.: Red teaming language models to reduce harms: methods, scaling behaviors, and lessons learned (2022)
70. Gausen, A., Mitra, B., Lindley, S.: A framework for exploring the consequences of AI-mediated enterprise knowledge access and identifying risks to workers. In: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (2024)
71. Gebru, T.: Effective altruism is pushing a dangerous brand of 'AI safety' (2022)
72. Gebru, T., Torres, É.P.: Eugenics and the promise of utopia through artificial general intelligence. First Monday (2023)
73. Gebru, T., Bender, E.M., McMillan-Major, A., Mitchell, M.: Statement from the listed authors of stochastic parrots on the "AI pause" letter (2023). https://www.dair-institute.org/blog/letter-statement-March2023/
74. Gehman, S., Gururangan, S., Sap, M., Choi, Y., Smith, N.A.: Realtoxicityprompts: evaluating neural toxic degeneration in language models (2020). arXiv preprint arXiv:2009.11462
75. Gershgorn, D.: GitHub's automatic coding tool rests on untested legal ground. The Verge (2021)
76. Gertner, J.: Wikipedia's moment of truth. The New York Times Magazine (2023)
77. Gillespie, T.: Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media. Yale University Press (2018)
78. Goldstein, S.: Informed Societies. Facet Publishing (2020)
79. Gonen, H., Goldberg, Y.: Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In: Proceedings of the NAACL, pp. 609–614 (2019)
80. González, M.: A better-informed society is a freer society (2021). https://www.unesco.org/en/articles/better-informed-society-freer-society
81. Gordon, C.: AI is accelerating the loss of our scarcest natural resource: water. Forbes (2024)
82. Gordon, A.D., Negreanu, C., Cambronero, J., Chakravarthy, R., Drosos, I., Fang, H., Mitra, B., Richardson, H., Sarkar, A., Simmons, S., et al.: Co-audit: tools to help humans double-check AI-generated content (2023). arXiv preprint arXiv:2310.01297

83. Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., Fritz, M.: Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. In: Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security, pp. 79–90 (2023)
84. Guerrini, F.: AI's unsustainable water use: how tech giants contribute to global water shortages. Forbes (2023)
85. Guha, N., Nyarko, J., Ho, D.E., Ré, C., Chilton, A., Narayana, A., Chohlas-Wood, A., Peters, A., Waldon, B., Rockmore, D.N., Zambrano, D., Talisman, D., Hoque, E., Surani, F., Fagan, F., Sarfaty, G., Dickinson, G.M., Porat, H., Hegland, J., Wu, J., Nudell, J., Niklaus, J., Nay, J., Choi, J.H., Tobia, K., Hagan, M., Ma, M., Livermore, M., Rasumov-Rahe, N., Holzenberger, N., Kolt, N., Henderson, P., Rehaag, S., Goel, S., Gao, S., Williams, S., Gandhi, S., Zur, T., Iyer, V., Li, Z.: LegalBench: a collaboratively built benchmark for measuring legal reasoning in large language models (2023)
86. Gupta, J., Bosch, H., Vliet, L.V.: AI's excessive water consumption threatens to drown out its environmental contributions. The Conversation (2024)
87. Halper, E.: Amid explosive demand, America is running out of power. The Washington Post (2024)
88. Hao, K.: Artificial intelligence is creating a new colonial world order. MIT Technology Review (2022)
89. Hao, K.: AI is taking water from the desert. The Atlantic (2024)
90. Hao, K., Hernández, A.P.: How the AI industry profits from catastrophe. MIT Technology Review (2022)
91. Hao, K., Seetharaman, D.: Cleaning up ChatGPT takes heavy toll on human workers. The Wall Street Journal **24** (2023)
92. Hardin, G.: The tragedy of the commons. In: Classic Papers in Natural Resource Economics Revisited, pp. 145–156. Routledge (2018)
93. Hays, K., Barr, A.: AI is killing the grand bargain at the heart of the web. 'we're in a different world.'. Business Insider (2024)
94. Heikkila, M.: This new data poisoning tool lets artists fight back against generative AI. MIT Technology Review (2023)
95. Hern, A.: TechScape: How cheap, outsourced labour in Africa is shaping AI English (2024). https://www.theguardian.com/technology/2024/apr/16/techscape-ai-gadgest-humane-ai-pin-chatgpt
96. Higgins, S., Gregory, L.: Information Literacy and Social Justice: Radical Professional Praxis. Library Juice Press (2013)
97. Hoel, E.: Here lies the Internet, murdered by generative AI. The Intrinsic Perspective (2024). https://www.theintrinsicperspective.com/p/here-lies-the-internet-murdered-by
98. Hoel, E.: A.I.-generated garbage is polluting our culture. The New York Times (2024)
99. Hofmann, K., Li, L., Radlinski, F.: Online evaluation for information retrieval. Found. Trends Inf. Retr. **10**(1), 1–117 (2016). ISSN 1554-0669. https://doi.org/10.1561/1500000051
100. Hofmann, K., Li, L., Radlinski, F.: Online evaluation for information retrieval. Found. Trends®Inform. Retrieval **10**(1), 1–117 (2016). ISSN 1554-0669. https://doi.org/10.1561/1500000051
101. Hom, K.-L.: Rage baiting. Westside Seattle (2015)
102. IPCC, Climate Change et al.: The physical science basis, the working group i contribution to the UN IPCC's Fifth Assessment Report (wg1 ar5) (2013)
103. Jasanoff, S., Kim, S.-H.: Dreamscapes of Modernity: Sociotechnical Imaginaries and the Fabrication Of Power. University of Chicago Press (2015)
104. Jin, D., Pan, E., Oufattole, N., Weng, W.H., Fang, H., Szolovits, P.: What disease does this patient have? A large-scale open domain question answering dataset from medical exams. Appl. Sci. **11**(14) (2021). ISSN 2076-3417. https://doi.org/10.3390/app11146421
105. Jobin, A., Ienca, M., Vayena, E.: The global landscape of AI ethics guidelines. Nat. Mach. Intell. **1**(9), 389–399 (2019)

106. Kak, A., West, S.M.: AI Now 2023 landscape: Confronting tech power (2023). https://ainowinstitute.org/2023-landscape
107. Kalluri, P., et al.: Don't ask if artificial intelligence is good or fair, ask how it shifts power. Nature **583**(7815), 169–169 (2020)
108. Kanungo, A.: The green dilemma: can AI fulfil its potential without harming the environment? Earth.Org (2023)
109. Kapoor, S., Bommasani, R., Klyman, K., Longpre, S., Ramaswami, A., Cihon, P., Hopkins, A., Bankston, K., Biderman, S., Bogen, M., et al.: On the societal impact of open foundation models (2024)
110. Kasirzadeh, A., Gabriel, I.: In conversation with artificial intelligence: aligning language models with human values. Philos. Technol. **36**(2), 1–24 (2023)
111. Khanal, S., Zhang, H., Taeihagh, A.: Why and how is the power of big tech increasing in the policy process? The case of generative AI. Policy Soc. puae012 (2024)
112. Khattak, R.: The environmental impact of e-waste. Earth.Org (2023)
113. Khlaaf, H.: Toward comprehensive risk assessments and assurance of AI-based systems. Trail of Bits (2023)
114. Klein, N.: AI machines aren't 'hallucinating'. but their makers are. The Guardian (2023)
115. Knibbs, K.: Scammy AI-generated book rewrites are flooding amazon. Wired (2024). https://www.wired.com/story/scammy-ai-generated-books-flooding-amazon/
116. Knibbs, K.: Your kid may already be watching AI-generated videos on YouTube. Wired (2024) https://www.wired.com/story/your-kid-may-be-watching-ai-generated-videos-on-youtube/
117. Kosinski, M., Stillwell, D., Graepel, T.: Private traits and attributes are predictable from digital records of human behavior. Proc. Natl. Acad. Sci. **110**(15), 5802–5805 (2013)
118. Kotek, H., Dockum, R., Sun, D.: Gender bias and stereotypes in large language models. In: Proceedings of The ACM Collective Intelligence Conference, pp. 12–24 (2023)
119. Kugel, S., Hiltner, S.: A new frontier for travel scammers: A.I.-generated guidebooks. The New York Times (2023). https://www.nytimes.com/2023/08/05/travel/amazon-guidebooks-artificial-intelligence.html
120. LaFrance, A.: The rise of techno-authoritarianism. The Atlantic (2024)
121. Landymore, F.: Sports Illustrated lays off journalists after announcing pivot to AI content. Futurism (2023). https://futurism.com/the-byte/sports-illustrated-lays-off-journalists-ai-content
122. Li, H., Vincent, N., Tsai, J., Kaye, J., Hecht, B.: How do people change their technology use in protest? Understanding. Proc. ACM Hum.-Comput. Interact. **3**(CSCW), 1–22 (2019)
123. Li, H., Vincent, N., Chancellor, S., Hecht, B.: The dimensions of data labor: A road map for researchers, activists, and policymakers to empower data producers. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, pp. 1151–1161 (2023)
124. Li, P., Yang, J., Islam, M.A., Ren, S.: Making AI less" thirsty": uncovering and addressing the secret water footprint of AI models (2023). arXiv preprint arXiv:2304.03271
125. Liang, C., Wu, X., Hua, Y., Zhang, J., Xue, Y., Song, T., Xue, Z., Ma, R., Guan, H.: Adversarial example does good: preventing painting imitation from diffusion models via adversarial examples. arXiv preprint arXiv:2302.04578 (2023)
126. Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C.D., Ré, C., Acosta-Navas, D., Hudson, D.A., Zelikman, E., Durmus, E., Ladhak, F., Rong, F., Ren, H., Yao, H., Wang, J., Santhanam, K., Orr, L., Zheng, L., Yuksekgonul, M., Suzgun, M., Kim, N., Guha, N., Chatterji, N., Khattab, O., Henderson, P., Huang, Q., Chi, R., Xie, S.M., Santurkar, S., Ganguli, S., Hashimoto, T., Icard, T., Zhang, T., Chaudhary, V., Wang, W., Li, X., Mai, Y., Zhang, Y., Koreeda, Y.: Holistic evaluation of language models (2023)

127. Liang, W., Izzo, Z., Zhang, Y., Lepp, H., Cao, H., Zhao, X., Chen, L., Ye, H., Liu, S., Huang, Z., et al.: Monitoring AI-modified content at scale: a case study on the impact of ChatGPT on AI conference peer reviews (2024). arXiv preprint arXiv:2403.07183
128. Liesenfeld, A., Dingemanse, M.: Rethinking open source generative AI: open washing and the EU AI Act. In: The 2024 ACM Conference on Fairness, Accountability, and Transparency, pp. 1774–1787 (2024)
129. Limbong, A.: Authors push back on the growing number of AI 'scam' books on Amazon. National Public Radio (2024). https://www.npr.org/2024/03/13/1237888126/growing-number-ai-scam-books-amazon
130. Liu, L.: Letter: setting rules for AI must avoid regulatory capture by big tech. Financial Times (2023)
131. Liu, Y., Deng, G., Li, Y., Wang, K., Zhang, T., Liu, Y., Wang, H., Zheng, Y., Liu, Y.: Prompt injection attack against LLM-integrated applications (2023). arXiv preprint arXiv:2306.05499
132. Liu, X., Yu, Z., Zhang, Y., Zhang, N., Xiao, C.: Automatic and universal prompt injection attacks against large language models (2024). arXiv preprint arXiv:2403.04957
133. Lonergan, R.: Mr. Justice Brandeis, great American. Mr. Justice Brandeis, Great American: Press Opinion and Public Appraisal (1941)
134. Möhlmann, M.: Algorithmic nudges don't have to be unethical. Harvard Business Review **22** (2021)
135. Markov, T., Zhang, C., Agarwal, S., Eloundou Nekoul, F., Lee, T., Adler, S., Jiang, A., Weng, L.: A holistic approach to undesired content detection in the real world. Proc. AAAI Conf. Artif. Intell. **37**(12), 15009–15018 (2023). https://doi.org/10.1609/aaai.v37i12.26752
136. Marr, B.: Is generative AI stealing from artists? Forbes (2023)
137. Martínez, G., Watson, L., Reviriego, P., Hernández, J.A., Juarez, M., Sarkar, R.: Towards understanding the interplay of generative artificial intelligence and the Internet (2023). arXiv preprint arXiv:2306.06130
138. Mathewson, T.: AI detection tools falsely accuse international students of cheating. The Markup (2023)
139. McMahon, C., Johnson, I., Hecht, B.: The substantial interdependence of Wikipedia and Google: a case study on the relationship between peer production communities and information technologies. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 11, pp. 142–151 (2017)
140. Mehdi, Y.: Bringing the full power of copilot to more people and businesses (2024). https://blogs.microsoft.com/blog/2024/01/15/bringing-the-full-power-of-copilot-to-more-people-and-businesses/
141. Metzler, D., Tay, Y., Bahri, D., Najork, M.: Rethinking search: making domain experts out of dilettantes. In: ACM SIGIR Forum, vol. 55, pp. 1–27. ACM, New York (2021)
142. Miceli, M., Posada, J., Yang, T.: Studying up machine learning data: why talk about bias when we mean power? Proc. ACM on Hum.-Comput. Interact. **6**(GROUP), 1–14 (2022)
143. Microsoft responsible AI standard, v2, 2022. https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE5cmFl
144. Min, S., Gururangan, S., Wallace, E., Hajishirzi, H., Smith, N.A., Zettlemoyer, L.: Silo language models: Isolating legal risk in a nonparametric datastore (2023). arXiv preprint arXiv:2308.04430
145. Mishkin, P., Ahmad, L., Brundage, M., Krueger, G., Sastry, G.: Dalle-2-preview/system-card.md at main · openai/dalle-2-preview. https://github.com/openai/dalle-2-preview/blob/main/system-card.md
146. Mitra, B.: Search and society: Reimagining information access for radical futures (2024). arXiv preprint arXiv:2403.17901
147. Moretti, F.A., Oliveira, V.E.d., Silva, E.M.K.d.: Access to health information on the Internet: a public health issue? Revista da Associação Médica Brasileira **58**, 650–658 (2012)
148. Mularczyk, K.: Row over deepfake of Polish PM in opposition-party broadcast. Brussels Signal (2023). https://brusselssignal.eu/2023/08/row-over-deepfake-of-polish-pm-in-opposition-party-broadcast/

149. Muldoon, J., Wu, B.A.: Artificial intelligence in the colonial matrix of power. Philos. Technol. **36**(4), 80 (2023)
150. Murgia, M.: AI academics under pressure to do commercial research. Financial Times **13** (2019)
151. Mutula, S.M.: Digital divide and economic development: case study of sub-Saharan Africa. Electron. Library **26**(4), 468–489 (2008)
152. Naughton, J.: AI's craving for data is matched only by a runaway thirst for water and energy. The Guardian (2024)
153. Navigli, R., Conia, S., Ross, B.: Biases in large language models: origins, inventory, and discussion. ACM J. Data Inform. Quality **15**(2), 1–21 (2023)
154. O'Gorman, M.: At the heart of artificial intelligence is racism and colonialism that we must excise. The Globe and Mail Web Edition (2023)
155. Oremus, W.: He wrote a book on a rare subject. then a ChatGPT replica appeared on Amazon. The Washington Post (2023) https://www.washingtonpost.com/technology/2023/05/05/ai-spam-websites-books-chatgpt/
156. Orland, K.: Lazy use of AI leads to Amazon products called "i cannot fulfill that request". Ars Technica (2024). https://arstechnica.com/ai/2024/01/lazy-use-of-ai-leads-to-amazon-products-called-i-cannot-fulfill-that-request/
157. Pan, Y., Pan, L., Chen, W., Nakov, P., Kan, M.Y., Wang, W.Y.: On the risk of misinformation pollution with large language models (2023). arXiv preprint arXiv:2305.13661
158. Park, G., Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Kosinski, M., Stillwell, D.J., Ungar, L.H., Seligman, M.E.: Automatic personality assessment through social media language. J. Pers. Soc. Psychol. **108**(6), 934 (2015)
159. Park, P.S., Goldstein, S., O'Gara, A., Chen, M., Hendrycks, D.: AI deception: a survey of examples, risks, and potential solutions (2023). arXiv preprint arXiv:2308.14752
160. Parmesan, C., Morecroft, M.D., Trisurat, Y.: Climate change 2022: impacts, adaptation and vulnerability (2022)
161. Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.M., Rothchild, D., So, D., Texier, M., Dean, J.: Carbon emissions and large neural network training (2021)
162. Patterson, D., Gonzalez, J., Hölzle, U., Le, Q., Liang, C., Munguia, L.M., Rothchild, D., So, D.R., Texier, M., Dean, J.: The carbon footprint of machine learning training will plateau, then shrink. Computer **55**(7), 18–28 (2022)
163. Pearson, J.: Scientific journal publishes AI-generated rat with gigantic penis in worrying incident. Vice (2024)
164. Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., Irving, G.: Red teaming language models with language models (2022)
165. Perrigo, B.: Exclusive: OpenAI used Kenyan workers on less than $2 per hour to make ChatGPT less toxic (2023). https://time.com/6247678/openai-chatgpt-kenya-workers/. Accessed 19 2023
166. Pierce, D.: You can now use the Dall-E 3 AI image generator inside Bing Chat. The Verge (2023)
167. Png, M.-T.: At the tensions of South and North: Critical roles of global South stakeholders in AI governance. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 1434–1445 (2022)
168. Polizzi, G.: Information literacy in the digital age: why critical digital literacy matters for democracy. In: Informed Societies: Why Information Literacy Matters for Citizenship, Participation and Democracy, pp. 1–23 (2020)
169. Poynting, M., Rivault, E.: 2023 confirmed as world's hottest year on record (2024)
170. Prabhakaran, V., Mitchell, M., Gebru, T., Gabriel, I.: A human rights-based approach to responsible AI (2022). arXiv preprint arXiv:2210.02667
171. Qi, X., Zeng, Y., Xie, T., Chen, P.Y., Jia, R., Mittal, P., Henderson, P.: Fine-tuning aligned language models compromises safety, even when users do not intend to! arXiv preprint arXiv:2310.03693 (2023)

172. Quercia, D., Kosinski, M., Stillwell, D., Crowcroft, J.: Our Twitter profiles, our selves: predicting personality with Twitter. In: 2011 IEEE third international Conference on Privacy, Security, Risk and Trust and 2011 IEEE third international Conference on Social Computing, pp. 180–185. IEEE (2011)
173. Ren, S.: How much water does AI consume? The public deserves to know it (2023)
174. Report: Ai fueling climate change, energy usage and disinformation (2024). https://sustainablebrands.com/read/product-service-design-innovation/ai-fueling-climate-change-energy-disinformation
175. Report: Artificial intelligence a threat to climate change, energy usage and disinformation (2024). https://foe.org/news/ai-threat-report/
176. Robins-Early, N.: New GPT-4o AI model is faster and free for all users, OpenAI announces. The Guardian (2024)
177. Russell, S., Dewey, D., Tegmark, M.: Research priorities for robust and beneficial artificial intelligence. AI Mag. **36**(4), 105–114 (2015)
178. Saad-Falcon, J., Khattab, O., Potts, C., Zaharia, M.: Ares: an automated evaluation framework for retrieval-augmented generation systems (2024)
179. Sadeghi, M., Arvanitis, L.: Rise of the newsbots: Ai-generated news websites proliferating online. NewsGuard (2023)
180. Sambasivan, N., Arnesen, E., Hutchinson, B., Doshi, T., Prabhakaran, V.: Re-imagining algorithmic fairness in India and beyond. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 315–328 (2021)
181. Scells, H., Zhuang, S., Zuccon, G.: Reduce, reuse, recycle: green information retrieval research. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2825–2837 (2022)
182. Schaake, M.: Big tech calls for 'regulation' but is fuzzy on the details. Financial Times (2021)
183. Scientist, N.: COVID-19: the story of a pandemic. New Scientist **10** (2021)
184. Shah, C., Bender, E.M.: Situating search. In: Proceedings of the 2022 Conference on Human Information Interaction and Retrieval, pp. 221–232 (2022)
185. Shan, S., Cryan, J., Wenger, E., Zheng, H., Hanocka, R., Zhao, B.Y.: Glaze: Glaze: Protecting artists from style mimicry by {Text-to-Image} models. In: 32nd USENIX Security Symposium (USENIX Security 23), pp. 2187–2204 (2023)
186. Shan, S., Ding, W., Passananti, J., Zheng, H., Zhao, B.Y.: Prompt-specific poisoning attacks on text-to-image generative models (2023). arXiv preprint arXiv:2310.13828
187. Shankar, S., Zamfirescu-Pereira, J.D., Hartmann, B., Parameswaran, A.G., Arawjo, I.: Who validates the validators? Aligning LLM-assisted evaluation of LLM outputs with human preferences (2024)
188. Shrivastava, R.: OpenAI and Microsoft sued by nonfiction writers for alleged 'rampant theft' of authors' works. Forbes (2023)
189. Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., Anderson, R.: The curse of recursion: training on generated data makes models forget (2023). arXiv preprint arXiv:2305.17493
190. Shutterstock expands partnership with OpenAI, signs new six-year agreement to provide high-quality training data (2023). https://investor.shutterstock.com/news-releases/news-release-details/shutterstock-expands-partnership-openai-signs-new-six-year
191. Simchon, A., Edwards, M., Lewandowsky, S.: The persuasive effects of political microtargeting in the age of generative artificial intelligence. PNAS Nexus **3**(2), pgae035 (2024). ISSN 2752-6542. https://doi.org/10.1093/pnasnexus/pgae035
192. Smith, J.J., Beattie, L.: RecSys fairness metrics: many to use but which one to choose? arXiv preprint arXiv:2209.04011 (2022)
193. Smith, J.J., Beattie, L., Cramer, H.: Scoping fairness objectives and identifying fairness metrics for recommender systems: the practitioners' perspective. In: Proceedings of the ACM Web Conference 2023, WWW '23, pp. 3648–3659. Association for Computing Machinery, New York (2023). ISBN 9781450394161. https://doi.org/10.1145/3543507.3583204.

194. Solaiman, I.: The gradient of generative AI release: Methods and considerations. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, pp. 111–122 (2023)
195. Speare-Cole, R.: Generative AI could 'supercharge' climate disinformation, report warns. Independent (2024)
196. Srivastava, A., Rastogi, A., Rao, A., et al.: Beyond the imitation game: quantifying and extrapolating the capabilities of language models (2023)
197. Stahl, B.C., Eke, D.: The ethics of ChatGPT–exploring the ethical issues of an emerging technology. Int. J. Inform. Manag. **74**, 102700 (2024)
198. Stanford: Stanford HELM (2024). https://crfm.stanford.edu/helm/
199. Strubell, E., Ganesh, A., McCallum, A.: Energy and policy considerations for deep learning in NLP (2019). arXiv preprint arXiv:1906.02243
200. Supporting open source and open science in the EU AI Act (2023). https://huggingface.co/blog/assets/eu_ai_act_oss/supporting_OS_in_the_AIAct.pdf
201. Suresh, H., Guttag, J.: A framework for understanding sources of harm throughout the machine learning life cycle. In: Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, EAAMO '21. Association for Computing Machinery, New York (2021) ISBN 9781450385534. https://doi.org/10.1145/3465416.3483305.
202. Tacheva, J., Ramasubramanian, S.: Ai empire: Unraveling the interlocking systems of oppression in generative AI's global order. Big Data Soc. **10**(2), 20539517231219241 (2023)
203. Tamkin, A., Brundage, M., Clark, J., Ganguli, D.: Understanding the capabilities, limitations, and societal impact of large language models (2021). arXiv preprint arXiv:2102.02503
204. Tan, R., Cabato, R.: Behind the AI boom, an army of overseas workers in 'digital sweatshops'. The Washington Post (2023)
205. Tani, M.: New York Times drops out of AI coalition. Semafor (2023)
206. Taraborelli, D.: The sum of all human knowledge in the age of machines: a new research agenda for Wikimedia. In: ICWSM-15 Workshop on Wikipedia (2015)
207. Tarkowski, A.: How Wikipedia can shape the future of AI. Open Future (2023)
208. Tarkowski, A.: Stewarding the sum of all knowledge in the age of AI. Open Future (2023)
209. Taylor, L.: COVID-19: True global death toll from pandemic is almost 15 million, says WHO. Br. Med. J. **377**, o1144 (2022)
210. Taylor, A.: A historic rise in global conflict deaths suggests a violent new era (2023). https://www.washingtonpost.com/world/2023/06/29/conflict-war-deaths-global-peace-rise-casualty/
211. Thomas, P., Spielman, S., Craswell, N., Mitra, B.: Large language models can accurately predict searcher preferences (2023). arXiv preprint arXiv:2309.10621
212. Thompson, B., Dhaliwal, M.P., Frisch, P., Domhan, T., Federico, M.: A shocking amount of the web is machine translated: insights from multi-way parallelism (2024). arXiv preprint arXiv:2401.05749
213. United Nations Meetings Coverage and Press Releases: With highest number of violent conflicts since Second World War, United Nations must rethink efforts to achieve, sustain peace, speakers tell security council (2023). https://press.un.org/en/2023/sc15184.doc.htm
214. Urman, A., Makhortykh, M.: The silence of the LLMs: cross-lingual analysis of political bias and false information prevalence in ChatGPT, Google Bard, and Bing Chat (2023)
215. Varghese, S.: How a Google search could end up endangering a life. iTWire (2021)
216. Vartziotis, T., Dellatolas, I., Dasoulas, G., Schmidt, M., Schneider, F., Hoffmann, T., Kotsopoulos, S., Keckeisen, M.: Learn to code sustainably: an empirical study on LLM-based green code generation (2024)
217. Vincent, J.: The invention of AI 'gaydar' could be the start of something much worse. The Verge **21** (2017)
218. Vincent, N.: Don't give OpenAI all the credit for GPT-3: You might have helped create the latest "astonishing" advance in AI too, 2020. https://www.psagroup.org/blogposts/62
219. Vincent, J.: The lawsuit that could rewrite the rules of AI copyright. The Verge **22** (2022)

220. Vincent, J.: Shutterstock will start selling AI-generated stock imagery with help from OpenAI. The Verge **25** (2022).
221. Vincent, J.: AI art tools stable diffusion and midjourney targeted with copyright lawsuit. The Verge (2023)
222. Vincent, N.: The WGA strike is a canary in the coal mine for AI labor concerns (2023). https://dataleverage.substack.com/p/the-wga-strike-is-a-canary-in-the
223. Vincent, N., Li, H.: GitHub Copilot and the exploitation of "data labor": A wake-up call for the tech industry (2021). https://www.psagroup.org/blogposts/62
224. Vincent, N., Li, H.: ChatGPT stole your work. so what are you going to do? (2023).
225. Vincent, N., Johnson, I., Hecht, B.: Examining Wikipedia with a broader lens: quantifying the value of Wikipedia's relationships with other large-scale online communities. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pp. 1–13 (2018)
226. Vincent, N., Hecht, B., Sen, S.: "data strikes": evaluating the effectiveness of a new form of collective action against technology companies. In: The World Wide Web Conference, pp. 1931–1943 (2019)
227. Vincent, N., Li, Y., Zha, R., Hecht, B.: Mapping the potential and pitfalls of" data dividends" as a means of sharing the profits of artificial intelligence (2019). arXiv preprint arXiv:1912.00757
228. Vincent, N., Li, H., Tilly, N., Chancellor, S., Hecht, B.: Data leverage: a framework for empowering the public in its relationship with technology companies. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 215–227 (2021)
229. Violino, B.: AI tools such as ChatGPT are generating a mammoth increase in malicious phishing emails. CNBC (2023). https://www.cnbc.com/2023/11/28/ai-like-chatgpt-is-creating-huge-increase-in-malicious-phishing-email.html
230. Wang, B., Chen, W., Pei, H., Xie, C., Kang, M., Zhang, C., Xu, C., Xiong, Z., Dutta, R., Schaeffer, R., Truong, S.T., Arora, S., Mazeika, M., Hendrycks, D., Lin, Z., Cheng, Y., Koyejo, S., Song, D., Li, B.: DecodingTrust: a comprehensive assessment of trustworthiness in GPT models (2024)
231. Warren, T.: Microsoft's new Copilot Pro brings AI-powered Office features to the rest of us, 2024. https://www.theverge.com/2024/1/15/24038711/microsoft-copilot-pro-office-ai-apps
232. Wei, J., Yang, C., Song, X., Lu, Y., Hu, N., Huang, J., Tran, D., Peng, D., Liu, R., Huang, D., Du, C., Le, Q.V.: Long-form factuality in large language models (2024)
233. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., et al.: Ethical and social risks of harm from language models (2021). arXiv preprint arXiv:2112.04359
234. Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., et al.: Taxonomy of risks posed by language models. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 214–229 (2022)
235. Welbl, J., Glaese, A., Uesato, J., Dathathri, S., Mellor, J., Hendricks, L.A., Anderson, K., Kohli, P., Coppin, B., Huang, P.S.: Challenges in detoxifying language models (2021). arXiv preprint arXiv:2109.07445
236. Whittaker, M.: The steep cost of capture. Interactions **28**(6), 50–55 (2021). ISSN 1072-5520. https://doi.org/10.1145/3488666
237. Whittaker, M.: The steep cost of capture. Interactions **28**(6), 50–55 (2021)
238. Williams, R.: Humans may be more likely to believe disinformation generated by AI. MIT Technology Review (2023). https://www.technologyreview.com/2023/06/28/1075683/humans-may-be-more-likely-to-believe-disinformation-generated-by-ai/
239. Williams, A., Miceli, M., Gebru, T.: The exploited labor behind artificial intelligence. Noema Mag. **13** (2022). https://www.noemamag.com/the-exploited-labor-behind-artificial-intelligence/
240. World economic forum global risks report 2024 (2024) https://www.weforum.org/publications/global-risks-report-2024/

241. Wu, C.J., Raghavendra, R., Gupta, U., Acun, B., Ardalani, N., Maeng, K., Chang, G., Aga, F., Huang, J., Bai, C., et al.: Sustainable AI: environmental implications, challenges and opportunities. Proc. Mach. Learn. Syst. **4**, 795–813 (2022)
242. Xiang, C.: OpenAI used Kenyan workers making $2 an hour to filter traumatic content from ChatGPT. VICE (2023)
243. Yom-Tov, E., Dumais, S., Guo, Q.: Promoting civil discourse through search engine diversity. Soc. Sci. Comput. Rev. **32**(2), 145–154 (2014)
244. Youyou, W., Kosinski, M., Stillwell, D.: Computer-based personality judgments are more accurate than those made by humans. Proc. Natl. Acad. Sci. 112(4), 1036–1040 (2015)
245. Yu, P.K.: Bridging the digital divide: Equality in the information age. Cardozo Arts Ent. LJ **20**, 1 (2002)
246. Ziegler, D.M., Stiennon, N., Wu, J., Brown, T.B., Radford, A., Amodei, D., Christiano, P., Irving, G.: Fine-tuning language models from human preferences (2019). arXiv preprint arXiv:1909.08593
247. Zuboff, S.: The age of surveillance capitalism. In: Social Theory Re-Wired, pp. 203–213. Routledge (2023)
248. Zuboff, S., Möllers, N., Wood, D.M., Lyon, D.: Surveillance capitalism: an interview with Shoshana Zuboff. Surveill. Soc. **17**(1/2), 257–266 (2019)
249. Zuccon, G., Scells, H., Zhuang, S.: Beyond CO_2 emissions: the overlooked impact of water consumption of information retrieval models. In: Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval, pp. 283–289 (2023)

Chapter 8
Recommendation in the Era of Generative Artificial Intelligence

Wenjie Wang, Yongfeng Zhang, and Tat-Seng Chua

Abstract The landscape of recommendation systems has undergone significant transformation, driven by advancements in generative AI. This section explores how generative AI, particularly Large Language Models (LLMs), can revolutionize traditional recommendation systems. By leveraging their powerful capabilities in language comprehension, reasoning, planning, and generation, recommendation systems can facilitate more intelligent user-system interactions, enhance personalized content generation, improve data representation, achieve generative item recall and ranking, and contribute to evaluation processes. These advancements promise to enhance user experience and system performance but also present challenges such as ensuring trustworthiness in AI-generated content and managing high computational costs. We discuss these developments and identify open problems and future research directions for integrating generative AI into recommendation systems.

8.1 Introduction

The landscape of recommendation systems has evolved dramatically over the past few decades. Generally speaking, recommendation systems aim to infer user preference from behaviors and provide personalized recommendations by various algorithms such as collaborative filtering and content-based approaches [61, 62]. As digital data explodes and computational power surges, recommendation systems advance significantly, incorporating powerful machine learning and deep learning models [21, 22]. With the maturing of technology, recommendation systems have grown into a critical infrastructure for information dissemination across various

W. Wang (✉) · T.-S. Chua
National University of Singapore, Singapore, Singapore
e-mail: dcscts@nus.edu.sg

Y. Zhang
Rutgers University, Piscataway, NJ, USA
e-mail: yongfeng.zhang@rutgers.edu

R. W. White, C. Shah (eds.), *Information Access in the Era of Generative AI*,
The Information Retrieval Series 51, https://doi.org/10.1007/978-3-031-73147-1_8

Internet platforms, ranging from social media and content streaming services to e-commerce [38].

In recent years, generative AI has experienced explosive growth, affecting information generation and exchange on the Internet. On the one hand, generative AI showcases remarkable content generation capabilities across various modalities including text, images, videos, and audio, fundamentally altering the landscape of information generation [60, 63, 64, 82, 91, 101]. In contrast to generating content by humans, generative AI can swiftly produce vast amounts of personalized content and collaborate with humans for editing, significantly enhancing the personalization and efficiency of content generation. On the other hand, LLMs as epitomized by ChatGPT [8, 55] exhibit powerful capabilities in reasoning, planning, and language comprehension and generation, thereby revolutionizing the way of information exchange. LLMs facilitate more intelligent interactions with users, allowing users to freely express information needs while enabling LLMs to better understand user intents and preferences to deliver personalized content. As such, by revolutionizing the ways of information generation and exchange, generative AI has the potential to reshape the paradigm of information dissemination in traditional recommendation systems.

We outline how generative AI can enhance traditional recommendation systems from five different perspectives.

(1) **Intelligent interactions**. LLMs enable more intelligent interactions between the recommendation systems and users, facilitating seamless information exchange. On the user side, individuals can articulate their information needs more actively and efficiently through natural language and proactively correct any unsuitable recommendations, thus exerting better control over the recommended items (controllability) [37]. On the recommendation system side, the system gains a deeper understanding of user intent and preferences; meanwhile, the system may provide superior explanations for recommendations (explainability) [95]. More importantly, recommendation systems can optimize long-term objectives based on LLMs' reasoning and planning abilities, shifting toward some long-term recommendation tasks such as proactive recommendation [6].

(2) **Personalized content generation**. Generative AI can collaborate with content producers and users to generate or edit more diverse and personalized content across various domains such as news [16, 34, 102], short videos [77], movies [104], music [13], and clothing design [90]. In these domains, generative AI can combine implicit user feedback (e.g., click and favorite), explicit user instructions, and context information to model user preference and then personalize existing content or generate entirely new content.

(3) **Data augmentation and representation**. Generative models are able to enrich recommendation data for traditional recommendation models. First, its powerful generative capabilities can be harnessed for data augmentation, thereby enhancing data heterogeneity and model generalization for traditional rec-

ommendation systems [48, 85]. Second, it can enhance feature encoding by providing superior feature representation ability [35, 58].

(4) **Generative recall and ranking**. Generative item recall and ranking represent another crucial research direction. Classic generative models, such as Variational Autoencoders (VAEs), have shown promising performance in item ranking tasks [40, 78]. Nowadays, LLMs and diffusion models are also being applied to item recall and ranking [26, 29, 79]. For example, given users' features and interaction history in natural language, LLMs can generate the item identifier (e.g., ID and title) as the recalled item [27]. Diffusion models can also generate the next item to interact with or directly generate the ranking scores for existing items based on user interaction history [39, 79].

(5) **Evaluation**. Generative AI also contributes to the evaluation of recommendations from the perspective of content auditing and user simulation. For instance, intelligent LLMs can inspect recommended content from the angles of biases, authenticity, legality, and more. Additionally, LLM-based agents can be optimized to simulate user behaviors, and then such agents can interact with the recommendation systems to interactively evaluate the recommendations.

Based on these five dimensions, we anticipate several open problems and future directions for leveraging generative AI to empower recommendation systems. First, optimizing LLMs to efficiently achieve more natural and seamless interactions with users for recommendations remains a significant challenge. As agent technology matures, the interaction between agents and users for information recommendations will become a trend in the future. Second, generative AI aiding content generation needs to integrate domain-specific knowledge, user instructions, and feedback for personalized content generation. Additionally, ensuring trustworthiness in AI-generated content (AIGC) for recommendation faces various challenges, particularly due to issues about copyright, bias, privacy, hallucination/authenticity, and safety. Third, integrating multimodal information into multimodal LLMs for more accurate recommendations across multiple scenarios and domains for open-world recommendation is a promising development direction. Nevertheless, the high costs of using LLMs for recommendation in industry remain a significant challenge. Finally, exploring content auditing and user simulation via LLMs presents numerous research challenges.

8.2 Literature Review

In this section, we critically examine the evolution of recommendation systems, tracing the progression from traditional machine learning techniques to deep learning approaches in a generative manner. And then we envision the advancements toward generative recommendation.

8.2.1 Machine Learning-Based Recommendation

Recommendation systems typically rely on Collaborative Filtering (CF) for personalized recommendations [23], which assumes users with similar interactions (e.g., ratings or clicks) share similar preferences on items. To achieve CF, early effort has been made to develop memory-based methods, which predict user interactions by memorizing similar user's or item's ratings [7, 24, 44, 65]. In the same period, Singular Value Decomposition (SVD) [65] has also been applied to reduce the dimensionality of the user-item interaction data for more efficient CF modeling. Later on, popularized by the Netflix Prize, Matrix Factorization (MF) [62] has become one of the most representative CF approaches. It decomposes the user-item interactions into user and item latent factors stored in two matrices and then leverages the inner product of the user and item latent factors to predict the user-item interaction. In addition to CF-based methods, an orthogonal line of research focuses on content-based techniques, which attempt to encode user/item features for interaction prediction. Factorization machine [61] stands as a notable content-based method that represents user/item features as latent factors and models the high-order features to predict user-item interactions.

8.2.2 Deep Learning-Based Recommendation

As deep neural networks have demonstrated exceptional learning capabilities across various domains, there emerges a trend in leveraging deep learning methodologies to tackle complex user-item interaction patterns in recommendation systems [10, 32]. Notably, Neural Collaborative Filtering (NCF) [23] has been developed to effectively model noisy implicit feedback data with the use of multi-layer perceptrons (MLP), thereby enhancing recommendation performance. Additionally, Bert4Rec [70] leverages deep bidirectional self-attention mechanisms to analyze user behavior sequences for sequential recommendation tasks. Caser [71] introduces a convolutional sequence model that employs both horizontal and vertical convolutional filters to identify sequential patterns. Subsequently, motivated by the graphical nature of user-item interactions, researchers have explored the application of graph neural networks for recommendation tasks. This research line has facilitated the exploitation of high-order neighbor information to improve the representation of users and items, as exemplified by Neural Graph Collaborative Filtering (NGCF) [80] and LightGCN [22].

8.2.3 *Toward Generative Recommendation*

Most recommendation systems utilize discriminative models to predict user-item interactions. Nevertheless, many efforts have showcased the promising performance of using generative models for recommendation.

Early studies employ Variational Autoencoder (VAE)-based methods to generate user-item interactions in parallel [40]. Subsequently, inspired by the powerful generative capability of diffusion models, researchers have explored utilizing diffusion models for interaction prediction in various recommendation tasks [39, 79, 100]. In addition to interaction generation, the strong content generation ability of generative AI allows for the repurposing of existing items or creation of new items based on user preference [77]. For instance, a recent work DiFashion [90] harnesses exceptional diffusion models to simultaneously generate multiple fashion images for personalized outfit generation, showing promising results.

Moreover, we have also witnessed a rapidly growing interest in leveraging LLMs for various recommendation tasks, e.g., conversational recommendations [15, 41], sequential recommendation [18, 42], multimodal recommendation [19, 47], and explainable recommendation [36, 49]. Consequently, the research on employing generative models for recommendation emerges as a particularly promising avenue.

8.3 Generative Recommendation

As shown in Fig. 8.1, we overview the benefits of using generative models for recommendation systems and detail the five perspectives in the following sections.

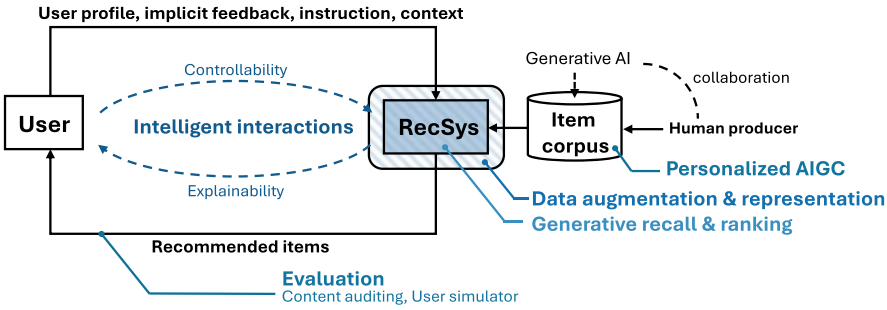

Fig. 8.1 Overview of using generative models for recommendation systems from five perspectives

8.3.1 Intelligent Interactions

Traditional recommendation systems usually infer user preference from implicit user feedback such as clicks. Surpassing the traditional recommendation systems, the powerful LLMs provide intelligent interactions through natural language between users and systems more transparently and efficiently [15]. These intelligent interactions offer several benefits to recommendation systems, for example, controllability, explainability, and long-term planning, demonstrating benefits from both the user and system perspectives. Beyond these, it shows great potential to envision more intelligent interactions between agents and users. For example, an agent can leverage tools or collaborate with more agents to serve users' information needs [96].

- **Controllability**. From the user side, users can explicitly express their information needs or feedback for the recommended items, guiding the system to generate, adjust, or correct the recommendations according to the users' personalized information needs [14]. The integration of such explicit instructions can enhance the users' controllability over the system, improving the user's experience and satisfaction.
- **Explainability**. From the system side, the recommendation system can better capture users' intent and preferences through both implicit feedback and user instructions. Besides, the recommendation systems can also actively ask users questions to clarify the users' information needs [41]. Thereafter, the system might provide explanations with the recommended items to facilitate users' understanding and interactions [17].
- **Long-term optimization**. LLMs might help recommendation systems optimize some long-term objectives through their planning and reasoning abilities, such as user retention rate and proactive recommendation. For instance, proactive recommendation aims to guide users' preferences to escape from filter bubbles through multiple rounds of recommendations in a period [67]. LLMs can help estimate the effect of recommended items on users' preferences and steer users to interact with some items by planning a sequence of recommended items [103].

8.3.2 Personalized Content Generation

In recent years, there have been remarkable advancements in generative AI, showcasing unprecedented abilities to produce high-quality content across various modalities, including image (e.g., Stable Diffusion [63], DALL·E [59]), video (e.g., Make-A-Video [69]), text (ChatGPT [8, 55]), and audio (WaveNet [54]). Leveraging the powerful generative AI in item production can significantly sup-

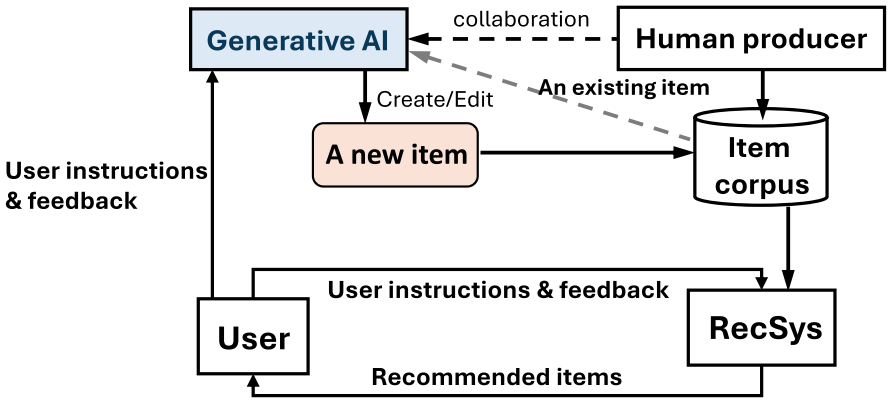

Fig. 8.2 Illustration of generative AI for content generation for recommendation

plement human-generated items in traditional recommendation systems and create more personalized content [77].

To elaborate, AIGC enriches the traditional item corpus in two primary ways [77]: (1) *Content repurposing*, which edits the existing items based on the user preference inferred from implicit feedback or instructions. As shown in Fig. 8.2, generative AI may take the user information and an existing item as input and then perform conditional generation [77] to edit an existing item to meet user-specific preference. (2) *Content creation*, which creates new items tailored to satisfy user-specific intent and preferences. In this way, generative AI can directly generate a new item without retrieving any existing item in the item corpus. In these two ways, generative AI can either collaborate with item producers and regular users for content generation or edit/create content by itself in some scenarios. By integrating generative AI, we can significantly enhance the personalization and efficiency of content generation in recommendation systems [77, 90].

8.3.3 Data Augmentation and Representation

- **Data augmentation**. Data sparsity is a long-standing problem in recommendation systems, where insufficient features and user-item interactions hinder the modeling of collaborative information, thus hurting the accuracy, robustness, and generalization of recommendation systems. Fortunately, generative models can alleviate this issue by augmenting the training data, including user profiles [99], item features [46], and user-item interactions [84] based on rich prior knowledge in generative models.
- **Feature representation**. In addition to data augmentation, leveraging generative models as the feature encoder is another prevailing approach to bolster the representations of user preference and item characteristics. Traditional

recommendation models typically encode the features through neural networks trained from scratch [86] or some medium-sized pre-trained models [35]. In contrast, LLMs, pre-trained over significantly larger datasets, possess rich world knowledge and excel in capturing intricate patterns and nuances in user behaviors and item characteristics. Therefore, harnessing LLMs to encode the user and item features offers significant benefits to downstream recommendation models.

8.3.4 *Generative Recall and Ranking*

Variational Autoencoders and Diffusion Models Compared to traditional recommendation systems that predict user-item interactions via discriminative models, VAE- and diffusion-based recommenders [40, 51, 79] predict the preference over all items in a generative manner. In particular, VAE-based models can learn the underlying probability distribution from the user's historical interactions and then predict the interactions over items simultaneously through neural models.

Another promising line is to utilize diffusion models to replace VAE-based models for interaction modeling with enhanced generative and representation abilities [33]. The diffusion-based recommendation models work by gradually adding noises into the user interactions in diffusing steps and predicting future interactions during the reverse denoising steps [79]. As shown in Fig. 8.3a, during training, the user's interactions are gradually diffused into random noise through the forward process. And then the diffusion model is optimized to recover the

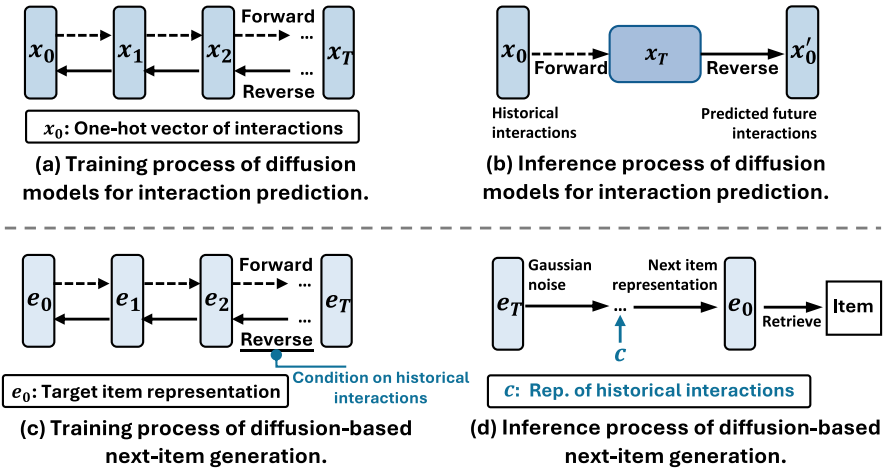

Fig. 8.3 Illustration of diffusion-based recommendation models. (**a**)–(**b**) Training/inference process of interaction prediction. (**c**)–(**d**) Training/inference process of next-item generation

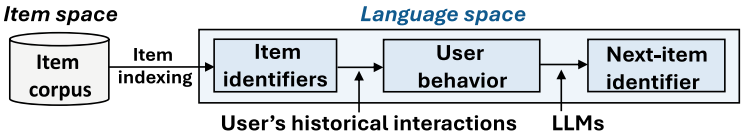

Fig. 8.4 Illustration of using LLMs for generative recall

interactions step by step through the reverse process, i.e., denoising [25]. As for inference (Fig. 8.3b), given the user's historical interactions, the diffusion model will corrupt the historical interactions and predict the future interactions via reverse steps.

Besides using diffusion models for interaction prediction, many efforts also shed light on directly generating the next-interacted item as the recommendation, such as generating item features or embeddings [39, 92]. As illustrated in Fig. 8.3c, diffusion models are trained to reconstruct the target item representation from the corrupted counterpart. Notably, to achieve personalized recommendations, the item representation is reconstructed guided by users' historical interactions [92]. During inference as shown in Fig. 8.3d, given a user's historical interactions, it generates the next-item representation conditioned on the representation of historical interactions [39]. Based on the generated item representation, we can obtain the recommended items via rounding strategies such as KNN [92].

Large Language Models More recently, in the wake of LLMs, extensive efforts have tried to explore leveraging LLMs to reformulate the sequential recommendation task into a language modeling task and produce recommendations in a generative manner [18, 53]. There are mainly three crucial steps as shown in Fig. 8.4: (1) *item indexing*, which assigns each item an identifier, i.e., a token sequence, transiting the items from the item space into the language space [37, 42]; (2) *user behavior formulation*, which typically converts the user's historical interactions in natural language based on the item identifiers, as the input for LLMs; and (3) next-item generation, which autoregressively generates the item identifier via beam search [5] based on the user's historical interactions in natural language. In this way, LLM-based recommendation models generate recommendations without calculating each candidate's ranking score. They implicitly enumerate all candidates for next-item generation in the language space, thereby drastically reducing the computational costs for discriminative interaction prediction and memory usage for storing all item embeddings in large-scale item recommendation scenarios. Meanwhile, beyond next item generation, there is another research line that takes the user's and item's features as LLMs' input to generate whether the user will interact with the item [3], utilizing richer features to pursue more fine-grained item ranking.

Moreover, recent work also harnesses the strong generalization and reasoning ability of LLMs for cross-domain recommendation, cold-start recommenda-

tion [28], as well as explainable recommendation [49]. Moreover, LLMs show promising results on multimodal recommendation [19, 47], where items' visual features are mapped to the semantic space of LLMs for next-item generation.

8.3.5 Evaluation

Traditional recommendation evaluation typically relies on ground-truth items through accuracy metrics, e.g., Recall and Normalized Discounted Cumulative Gain (NDCG). However, they may fall short in assessing generative recommendations, especially in evaluating user satisfaction with AI-generated items under the paradigm of interactive recommendation. As such, we expect new and novel evaluation approaches for recommendation in the era of generative AI to take a step beyond traditional evaluation strategies.

- **Evaluation protocol and metrics.** To assess user satisfaction with intelligent interactive recommendations, it is crucial to devise novel evaluation protocols for flexible and dynamic user-system interactions. For example, [81] proposes an evaluation framework, which utilizes user simulators to evaluate interactive recommendations using both objective and subjective metrics. Specifically, the user simulator will first interact with recommendation systems, comprising user requests for content or the system's proactive solicitation of user preferences. Subsequently, the recommendation systems are evaluated based on two metrics: an objective metric (i.e., Recall), and a persuasiveness score obtained through LLM-based scorers [11], reflecting whether the user is persuaded to accept recommendations.
- **Agent for user simulation.** Considering the high costs of online evaluation with real users, the key to interactive evaluation under the interactive recommendation paradigm lies in building high-quality user simulators. Recently, LLM-based agents have shown good potential to simulate users, bridging the gap between offline and online evaluation. The environment between recommendation systems and users is quite complex due to various interference factors [74]. Nevertheless, due to their rich world knowledge, powerful memorization, and reasoning abilities, LLM-based agents possess the potential to collect users' information and simulate their various actions, such as conversations with the recommendation systems and click behaviors on the items [74]. In particular, as illustrated in Fig. 8.5, the agent for user simulation is typically equipped with a profile module to store the user feature, a memory module to record the user behaviors such as interacted items [98] or inferred user preference [94], and an action module to interact with the environment (e.g., items or other agents). By simulating user behaviors, LLM-based agents can advance traditional evaluation paradigms to interactive evaluation, thereby estimating the influence of the recommendation policies and recommended content on users and the community.

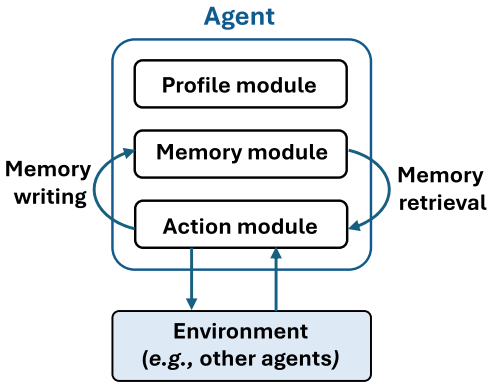

Fig. 8.5 Illustration of agent-based user simulator

- **LLMs for item auditing.** As for the item side, a key aspect lies in the evaluation of the AI-generated items, especially the trustworthiness checks, such as quality, misinformation, and adherence to legal and ethical standards [77]. To facilitate the automatic content auditing of AI-generated items, LLMs offer promising solutions with their competent capabilities, e.g., strong multimodal reasoning ability and deep understanding of the common knowledge [1]. For example, LLMs can be utilized to detect errors and correct answers through self-evaluation [89], and external knowledge can be retrieved by LLMs to detect misinformation [57].

8.4 Open Problems and Future Directions

8.4.1 Intelligent Interactions

LLMs may facilitate intelligent interactions between users and recommendation systems, thereby enhancing the user experience and ensuring long-term ecological stability within these systems. Nevertheless, it is crucial to consider specific challenges, particularly in terms of controllability, explainability, and long-term optimization.

- **Controllability.** Controllability in intelligent interactions allows users to tailor recommendation results based on their control instructions [76]. However, the implementation of controllability faces several challenges: (1) **The range of controllability.** Users can control various aspects of their recommendations, such as diversity, topicality, and item attributes, to suit their preferences and information needs. It is thus essential for LLMs to determine the proper control scope in an appropriate context for users to take charge of their recommendation outcomes. (2) **The inspiration for controllability.** Encouraging users to actively utilize control mechanisms is vital. LLMs need to know when and

how to give users instructions and hints to control the recommendation result. (3) **The execution of controllability.** Adjusting recommendations on the fly in response to user instructions is crucial. The challenge lies in ensuring how these modifications can accurately reflect the user's desires without being affected by irrelevant factors, such as previously out-of-date interactions [76].

- **Explainability.** Explainability plays a pivotal role in intelligent interactions to foster user trust. When users grasp the rationale behind specific recommendations, they are more likely to trust the system's suggestions [2]. However, the implementation of explainability presents several challenges: (1) **The generation of explanation.** Crafting personalized explanations that align with user preferences and recommendation outcomes in a specific context remains a complex issue. (2) **The update of explanation mechanism.** Regularly refining the explanation mechanisms based on user feedback and directives is essential. Determining effective methods for integrating such feedback and instructions to update the strategy of explanation generation is a promising area for future research.
- **Long-term optimization.** Traditional recommendation algorithms have predominantly focused on optimizing short-term accuracy, often overlooking the critical need for long-term optimization, which significantly impacts user retention rates and the overall ecological stability of the system [76]. It remains an open question to leverage the advanced planning capabilities of LLMs for strategic macro-control of recommendation results to enhance long-term performance metrics. Moreover, the journey toward long-term optimization is fraught with challenges, including shifts in user interests and the disruption caused by external environmental factors. It deserves to consider how to adjust the planning strategy in time to regulate the recommendation result in the long term.

8.4.2 Personalized Content Generation

Given that items created by humans may not meet the diverse and dynamic informational needs of users, generative AI offers the capability to generate personalized content tailored to individual specifications [77, 90]. However, there are three key challenges for content generation: (1) **Align with user preferences.** It is crucial to ensure the generated content precisely captures and interprets users' evolving preferences and information requirements. This requires the generative models to be capable of dynamically learning from user interactions, feedback, and consumption habits to tailor content personalization effectively. (2) **Quality and trustworthiness.** The assurance of the quality and trustworthiness of generated content emerges as a critical concern. This aspect requires the content to not only adhere to superior quality standards but also integrate ethical considerations, such as privacy and fairness from the trustworthy side. (3) **Copyright.** The legal and ethical implications of copyright in the realm of AI-Generated Content (AIGC)

present a complex landscape for exploration. Developing an approach that honors intellectual property rights while leveraging the creative capabilities of generative AI is essential. There is an urgent need to develop policies and technologies that manage the generation, identification, distribution, consumption, and destruction of AIGC, to ensure that users and creators navigate the generated content in compliance with the copyright laws and ethical guidelines.

8.4.3 Trustworthiness in Generative Recommendation

In addition to the ranking accuracy of generative models for recommendation, it is crucial to consider the trustworthiness of recommended content, which is highly related to ethical, social, and legal implications. Ensuring the trustworthiness of generative content is essential for maintaining user trust and satisfaction, supporting unbiased and fair decision-making, ensuring privacy, legality, and security, etc. However, numerous facets of trustworthiness still pose unresolved challenges.

- **Fairness and bias.** AIGC faces unique fairness challenges [97] that significantly impact user satisfaction and the long-term diversity of the ecosystem. The fairness issues can be categorized into two distinct types: (1) **User-side fairness.** There is a tendency for generative models to offer differing recommended content to users based on sensitive attributes, such as gender and race [38]. Addressing how to provide equitable recommendations across diverse user demographics when using generation for recommendations remains a critical, yet unresolved issue. (2) **Item-side fairness.** Generative models for recommendation may exhibit varying degrees of exposure for different item groups, where groups can be divided by various aspects such as uploaders and political bias [31, 97]. It remains a challenging task to ensure item-side fairness when using generative models for recommendation.
- **Hallucination/Authenticity.** Generative models, such as GPTs, are prone to generating text that may contain incorrect facts, fabricated details, or inconsistent information, a phenomenon commonly referred to as "hallucination" [12]. This issue might stem from limitations in a lack of related knowledge in generative models, potentially leading to diminished user trust and information pollution. Ensuring the authenticity of generated content remains a significant unresolved challenge. Some promising directions lie in self-evaluation [88] and verifying the accuracy of facts, statistics, and claims in generated content against external reliable sources [20]. Moreover, investigating data-centric strategies, such as data cleaning and data augmentation, to mitigate hallucinations is also critical [87].
- **Privacy.** The training of generative recommendation models demands substantial user data, highlighting the critical need for protecting data privacy. The effectiveness of representative methods of data privacy protection needs exploration such as privacy-preserving data cleansing [66], encrypted data

aggregation [73], and local differential privacy [75]. In addition, distributed learning methods, such as federated learning, also offer a promising solution by enabling model training and inference on client devices [50, 56]. Nonetheless, within the context of some large generative recommendations such as LLMs, the large model size incurs significant communication and computational costs in a federated learning framework. The challenge of reducing these costs while maintaining privacy in LLM-based recommendation systems remains an unresolved challenge.

- **Safety.** The content generated by AI must not pose any risks of harm to users, including risks of physical and psychological harm [30, 83]. For instance, the generated micro-video for teenagers should not contain any unhealthy content. However, the safety side of generated content has not been well explored. It is still an open problem to devise some pre-processing strategies like data cleansing, in-processing strategies such as building robust models, and post-processing methods such as rigorous output scrutiny in personalized content generation for recommendation.

8.4.4 Multimodal Large Language Models for Open-World Recommendation

With advancements in multimodal LLMs, exploring multimodal LLMs for open-world recommendation across various domains presents a promising avenue [47]. This might contribute to intelligent recommendation models to understand user behaviors across domains and yield recommendations in the open domain. However, this line of research is challenging:

- **Encoding and decoding of multimodal content and user behaviors.** A critical aspect of leveraging multimodal LLMs for recommendation systems involves the efficient encoding of user interaction behaviors and multimodal user and item content into the semantic space of multimodal LLMs, followed by an effective decoding process that maps back to the item space to generate recommendations. This requires methodologies capable of translating users' and items' semantics and interaction behaviors between their native modalities and the semantic space of LLMs, ensuring seamless understanding, reasoning, and recommendation [72].
- **Multimodal alignment.** The alignment of multimodal LLMs for recommendation is another key issue, including (1) the alignment across different modalities to pursue accurate content retrieval and generation [9] and (2) the alignment of multimodal LLMs to reduce hallucination [45] and meet social norms [68], leading to trustworthy recommendations. Despite promising results on the alignment of LLMs in the general domain, aligning multimodal LLMs specifically for recommendation remains challenging. Specifically, it is crucial to consider personalization for recommendation, and thus the alignment

involves understanding user preference accurately and aligning fine-grained multimodal content such as object-level content alignment with user preference. In addition, for the open-world recommendation, recommended content may exhibit hallucinations and generate toxic arguments, potentially resulting in serious consequences in high-stakes scenarios [93]. However, it is still non-trivial to pursue the alignment of multimodal LLMs to adhere to social norms for trustworthy recommendations.

- **Noises in multimodal content.** There are many noisy features in multimodal user/item content rather than useful information for recommendation, leading to a low signal-to-noise ratio in multimodal content and posing a significant challenge for interaction prediction. Strategies must be devised to navigate through multimodal content, extracting valuable features that closely align with user behaviors, thus enhancing the recommendation quality.
- **Utilization of cross-domain data.** Another pivotal area of exploration is the utilization of multimodal data across different tasks and domains to enhance open-world recommendations. The core challenge is devising models that can effectively learn invariant, robust, and effective user/item representations from heterogeneous data, thereby enabling a synthetic enhancement of recommendation quality in various scenarios.

8.4.5 Efficiency

In LLM-based recommendation, fine-tuning LLMs with user behaviors is crucial [43]. However, fine-tuning LLMs on large-scale recommendation data demands substantial computational resources and time costs, thereby diminishing the practicality of LLM-based recommendation systems in real-world applications. As such, it is essential to enhance the fine-tuning efficiency of LLM-based recommendation systems. Some valuable directions in enhancing efficiency include: (1) **Data selection.** Identifying and extracting the most typical and representative samples from the fine-tuning dataset is critical. Utilizing this selectively curated data for fine-tuning can significantly enhance efficiency [43]. (2) **Model architecture.** It is vital to explore strategies such as model pruning and distillation during the training phase to reduce the model size, thereby enhancing efficiency [52].

8.4.6 Agent for User Simulation

Given the advanced contextual understanding and reasoning capabilities of LLMs, it becomes feasible to develop user simulators via LLM-based agents that can accurately mimic user preferences, instructions, behaviors, and responses [4]. Such user simulators could play a pivotal role in assessing the effectiveness of specific recommendation algorithms or augmenting datasets. However, this area encounters

significant challenges: (1) **Reality of user simulators.** Achieving a high degree of accuracy in simulating user behaviors to closely replicate real user actions is quite challenging. This endeavor requires LLMs to accurately predict users' decisions in diverse contexts. Attaining a high precision presents considerable difficulties, as user decisions are subject to a wide array of influences, such as personal preferences, situational contexts, and psychological states. (2) **Adaptive upgrade of user simulators.** It is essential to update user simulators that align with dynamic user preferences and behaviors, since user preferences are constantly changing due to various internal and external factors. It remains an open question to explore how to flexibly and adaptively upgrade user simulators based on their latest interaction histories and external influences, such as social relationships. (3) **Simulation of user groups.** Another important challenge is how to use a user simulator to simulate user groups, including modeling the social influence within the user group, accurately predicting users' social behaviors, and capturing the dynamic changes in group behaviors.

8.4.7 Evaluation and Benchmarks

Redefining evaluation protocol for recommendation in the era of generative AI is also crucial [77]. Traditional ranking metrics such as Recall and NDCG do not fully encompass the multifaceted nature of user satisfaction that generative models aim to fulfill. Intuitively, persisting with metrics like ranking accuracy would constrain our understanding of a generative model's capability, particularly its generation quality, and aptitude for delivering personalized content. It is promising to consider designing new metrics that can evaluate the quality and relevance of generated items to user preference. Furthermore, it is also imperative to develop new benchmark datasets and evaluation paradigms for interactive evaluation.

8.5 Conclusion

Generative AI has experienced explosive prosperity recently, revolutionizing ways of information generation and exchange in recommendation systems. In this chapter, we outlined how generative AI can enhance traditional recommendation systems from five aspects: (1) LLMs enable more intelligent interactions between the recommendation systems and users, facilitating natural information exchange with the advantages of controllability, explainability, and long-term optimization. (2) Generative AI can supplement human-generated content with AIGC, revolutionizing information generation paradigms. (3) Generative models may enrich the recommendation data via superior data representation and data augmentation. (4) Generative models such as diffusion models and LLMs have been widely applied for item recall and ranking tasks, showing promising performance. (5) LLMs can

promote recommendation evaluation from the perspectives of LLM-based auditing and user simulation.

Generative AI holds immense potential for reshaping the landscape of recommendation systems. Despite many research challenges that lie ahead, including ensuring trustworthiness in AI-generated content, integrating multimodal information, efficiency, and evaluation challenges, addressing these challenges will pave the way for intelligent next-generation recommendation systems, significantly elevating the capabilities of personalized information dissemination in the era of generative AI.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv:2303.08774 (2023)
2. Ahmadian, M., Mahmood, A., Ahmadian, S.: A reliable deep representation learning to improve trust-aware recommendation systems. Expert Syst. Appl. **197**(2022), 116697 (2022)
3. Bao, K., Zhang, J., Zhang, Y., Wang, W., Feng, F., He, X.: TALLRec: an effective and efficient tuning framework to align large language model with recommendation. In: RecSys, pp. 1007–1014. ACM, New York (2023)
4. Bernard, N.: Leveraging user simulation to develop and evaluate conversational information access agents. In: WSDM, pp. 1136–1138 (2024)
5. Bevilacqua, M., Ottaviano, G., Lewis, P., Yih, S., Riedel, S., Petroni, F.: Autoregressive search engines: generating substrings as document identifiers. NeurIPS **35**, 31668–31683 (2022)
6. Bi, S., Wang, W., Pan, H., Feng, F., He, X.: Proactive recommendation with iterative preference guidance. In: arXiv:2403.07571 (2024)
7. Breese, J.S., Heckerman, D., Kadie, C.: Empirical analysis of predictive algorithms for collaborative filtering. In: UAI, pp. 43–52 (1998)
8. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. In: NeurIPS, pp. 1877–1901. Curran Associates, Inc., New York (2020)
9. Cao, M., Li, S., Li, J., Nie, L., Zhang, M.: Image-text retrieval: A survey on recent research and development. In: IJCAI. ijcai.org, pp. 5410–5417 2022
10. Chen, X., Zhang, Y., Qin, Z.: Dynamic explainable recommendation based on neural attentive models. In: AAAI, vol. 33, pp. 53–60 (2019)
11. Chen, X., Zhang, Y., Wen, J.-R.: Measuring“ why” in recommender systems: A comprehensive survey on the evaluation of explainable recommendation. arXiv:2202.06466 (2022)
12. Chen, Y., Fu, Q., Yuan, Y., Wen, Z., Fan, G., Liu, D., Zhang, D., Li, Z., Xiao, Y.: Hallucination detection: Robustly discerning reliable answers in large language models. In: CIKM, pp. 245–255 2023
13. Dai, S., Ma, X., Wang, Y., Dannenberg, R.B.: Personalised popular music generation using imitation and structure. J. New Music Res. **51**(1), 69–85 (2022)
14. Dong, Z., Chen, B., Liu, X., Polak, P., Zhang, P.: MuseChat: A conversational music recommendation system for videos. arXiv:2310.06282 (2023)
15. Friedman, L., Ahuja, S., Allen, D., Tan, T., Sidahmed, H., Long, C., Xie, J., Schubiner, G., Patel, A., Lara, H., et al.: Leveraging Large Language Models in Conversational Recommender Systems. arXiv:2305.07961 (2023)
16. Gao, S., Fang, J., Tu, Q., Yao, Z., Chen, Z., Ren, P., Ren, Z.: Generative News Recommendation. In: WWW. ACM, New York (2024)

17. Gao, Y., Sheng, T., Xiang, Y., Xiong, Y., Wang, H., Zhang, J.: Chat-rec: towards interactive and explainable LLMs-augmented recommender system. arXiv:2303.14524 (2023)
18. Geng, S., Liu, S., Fu, Z., Ge, Y., Zhang, Y.: Recommendation as language processing (RLP): a unified pretrain, personalized prompt & predict paradigm (p5). In: RecSys, pp. 299–315 (2022)
19. Geng, S., Tan, J., Liu, S., Fu, Z., Zhang, Y.: Vip5: towards multimodal foundation models for recommendation. In: EMNLP, ACL (2023)
20. Guo, Z., Michael, S., Vlachos, A.: A survey on automated fact-checking. Trans. Assoc. Comput. Linguistics **10**(2022), 178–206 (2022)
21. He, X., Chua, T.-S.: Neural factorization machines for sparse predictive analytics. In: SIGIR, pp. 355–364. ACM, New York (2017)
22. He, X., Deng, K., Wang, X., Li, Y., Zhang, Y., Wang, M.: LightGCN: simplifying and powering graph convolution network for recommendation. In: SIGIR, pp. 639–648. ACM, New York 2020
23. He, X., Liao, L., Zhang, H., Nie, L., Hu, X., Chua, T.-S.: Neural collaborative filtering. In: *WWW*, pp. 173–182. ACM, New York (2017)
24. Herlocker, J.L., Konstan, J.A., Borchers, A., Riedl, J.: An algorithmic framework for performing collaborative filtering. In: SIGIR, pp. 230–237. ACM, New York (1999)
25. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS, pp. 6840–6851. Curran Associates, Inc., New York (2020)
26. Hua, W., Ge, Y., Xu, S., Ji, J., Zhang, Y.: Up5: unbiased foundation model for fairness-aware recommendation. In: EACL, ACL (2024)
27. Hua, W., Xu, S., Ge, Y., Zhang, Y..: How to index item ids for recommendation foundation models. In: SIGIR-AP, pp. 195–204. ACM, New York (2023)
28. Huang, F., Yang, Z., Jiang, J., Bei, Y., Zhang, Y., Chen, H.: Large Language Model Interaction Simulator for Cold-Start Item Recommendation. arXiv:2402.09176 (2024)
29. Ji, J., Li, Z., Xu, S., Hua, W., Ge, Y., Tan, J., Zhang, Y.: GenRec: Large language model for generative recommendation. arXiv e-prints (2023)
30. Ji, J., Liu, M., Dai, J., Pan, X., Zhang, C., Bian, C., Chen, B., Sun, R., Wang, Y., Yang, Y.: Beavertails: towards improved safety alignment of LLM via a human-preference dataset. In: NeurIPS, vol. 36 (2024)
31. Jiang, M., Bao, K., Zhang, J., Wang, W., Yang, Z., Feng, F., He, X.: Item-side Fairness of Large Language Model-based Recommendation System. In: WWW. ACM, New York (2024)
32. Kang, W.-C., McAuley, J.: Self-attentive sequential recommendation. In: ICDM, pp. 197–206. IEEE, New York (2018)
33. Kingma, D.P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., Welling, M.: Improved variational inference with inverse autoregressive flow. In: NeurIPS, pp. 4743–4751. Curran Associates, Inc., New York (2016)
34. Kreps, S., McCain, R.M., Brundage, M.: All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. J. Exp. Political Sc. **9**(1), 104–117 (2022)
35. Li, J., Wang, M., Li, J., Fu, J., Shen, X., Shang, J., McAuley, J.: Text is all you need: learning language representations for sequential recommendation. In: KDD, pp. 1258–1267 (2023b)
36. Li, L., Zhang, Y., Chen, L.: Personalized prompt learning for explainable recommendation. TOIS **41**(4), 1–26 (2023c)
37. Li, L., Zhang, Y., Liu, D., Chen, L.: Large language models for generative recommendation: A survey and visionary discussions. In: LREC-COLING (2024)
38. Li, Y., Chen, H., Fu, Z., Ge, Y., Zhang, Y.: User-oriented fairness in recommendation. In: WWW, pp. 624–632. ACM, New York (2021)
39. Li, Z., Sun, A., Li, C.: DiffuRec: a diffusion model for sequential recommendation. TOIS **42**(3), pp. 1–28 (2023a)
40. Liang, D., Krishnan, R.G., Hoffman, M.D., Jebara, T.: Variational autoencoders for collaborative filtering. In: WWW, pp. 689–698 (2018)
41. Lin, G., Zhang, Y.: Sparks of Artificial General Recommender (AGR): experiments with ChatGPT. Algorithms **16**(9), 432 (2023)

42. Lin, X., Wang, W., Li, Y., Feng, F., See-Kiong, N., Tat-Seng, C.: Bridging Items and Language: A Transition Paradigm for Large Language Model-Based Recommendation. In: KDD. ACM, New York (2024)
43. Lin, X., Wang, W., Li, Y., Yang, S., Feng, F., Wei, Y., Tat-Seng, C.: Data-efficient Fine-tuning for LLM-based Recommendation. In: SIGIR. ACM, New York (2024)
44. Linden, G., Smith, B., York, J.: Amazon. com recommendations: Item-to-item collaborative filtering. IEEE Internet Comput. **7**(1), 76–80 (2003)
45. Liu, H., Xue, W., Chen, Y., Chen, D., Zhao, X., Wang, K., Hou, L., Li, R., Peng, W.: A survey on hallucination in large vision-language models. arXiv:2402.00253 (2024).
46. Liu, Q., Chen, N., Sakai, T., Wu, X.-M.: A First Look at LLM-Powered Generative News Recommendation. arXiv:2305.06566 (2023)
47. Liu, Y., Wang, Y., Sun, L., Yu, P.S.: Rec-GPT4V: Multimodal Recommendation with Large Vision-Language Models. arXiv:2402.08670 (2024a)
48. Luo, S., Yao, Y., He, B., Huang, Y., Zhou, A., Zhang, X., Xiao, Y., Zhan, M., Song, L.: Integrating Large Language Models into Recommendation via Mutual Augmentation and Adaptive Aggregation. arXiv:2401.13870 (2024)
49. Luo, Y., Cheng, M., Zhang, H., Lu, J., Chen, E.: Unlocking the potential of large language models for explainable recommendations. arXiv:2312.15661 (2023)
50. Ma, C., Ren, X., Xu, G., He, B.: FedGR: Federated graph neural network for recommendation systems. Axioms **12**(2), 170 (2023b)
51. Ma, J., Zhou, C., Cui, P., Yang, H., Zhu, W.: Learning disentangled representations for recommendation. In: NeurIPS, vol. 32 (2019)
52. Ma, X., Fang, G., Wang, X.: LLM-pruner: On the structural pruning of large language models. In: NeurIPS, vol. 36, pp. 21702–21720 (2023)
53. Mei, K., Zhang, Y.: LightLM: A Lightweight Deep and Narrow Language Model for Generative Recommendation. arXiv:2310.17488 (2023)
54. van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K.: WaveNet: A generative model for raw audio. arXiv:1609.03499 (2016)
55. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training Language Models to Follow Instructions with Human Feedback. In: NeurIPS. Curran Associates Inc, New York (2022)
56. Pei, Y., Mao, R., Liu, Y., Chen, C., Xu, S., Qiang, F., Tech, B.E.: Decentralized federated graph neural networks. In: IJCAI (2021)
57. Qian, H., Zhu, Y., Dou, Z., Gu, H., Zhang, X., Liu, Z., Lai, R., Cao, Z., Nie, J.-Y., Wen, J.-R.: WebBrain: Learning to Generate Factually Correct Articles for Queries by Grounding on Large Web Corpus. arXiv:2304.04358 (2023)
58. Qiu, Z., Wu, X., Gao, J., Fan, W.: U-BERT: Pre-training user representations for improved recommendation. In: AAAI, vol. 35, pp. 4320–4327 (2021)
59. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv:2204.06125 1, 2, 3 (2022)
60. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: ICML, pp. 8821–8831. PMLR, New York (2021)
61. Rendle, S.: Factorization Machines. In: ICDM, pp. 995–1000. IEEE, New York (2010)
62. Rendle, S., Freudenthaler, C., Gantner, Z., Schmidt-Thieme, L.: BPR: Bayesian personalized ranking from implicit feedback. In: UAI, pp. 452–461. AUAI Press, New York (2009)
63. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR, pp. 10684–10695 2022
64. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Lopes, R.G., Ayan, B.K., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. In: NeurIPS, vol. 35, pp. 36479–36494. Curran Associates, Inc., New York (2022)
65. Sarwar, B., Karypis, G., Konstan, J., Riedl, J.: Item-based collaborative filtering recommendation algorithms. In: WWW, pp. 285–295. ACM, New York (2001)

66. Shearer, C.: The CRISP-DM model: the new blueprint for data mining. J. Data Warehousing **5**(4), 13–22 (2000)
67. Shi, W., He, X., Zhang, Y., Gao, C., Li, X., Zhang, J., Wang, Q., Feng, F.: Large Language Models are Learnable Planners for Long-Term Recommendation. In: SIGIR. ACM, New York (2024)
68. Shi, Z., Wang, Z., Fan, H., Zhang, Z., Li, L., Zhang, Y., Yin, Z., Sheng, L., Qiao, Y., Shao, J.: Assessment of Multimodal Large Language Models in Alignment with Human Values. arXiv:2403.17830 (2024b)
69. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-A-Video: Text-to-Video Generation without Text-Video Data. In: *ICLR* (2022)
70. Sun, F., Liu, J., Wu, J., Pei, C., Lin, X., Ou, W., Jiang, P.: BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In: *CIKM*, pp. 1441–1450. ACM, New York (2019)
71. Tang, J., Wang, K.: Personalized top-n sequential recommendation via convolutional sequence embedding. In: WSDM, pp. 565–573. ACM, New York (2018)
72. Tao, Z., Wei, Y., Wang, X., He, X., Huang, X., Chua, T.-S.: MGAT: Multimodal graph attention network for recommendation. Inf. Process. Manag. **57**(5), 102277 (2020)
73. Toch, E., Wang, Y., Cranor, L.F.: Personalization and privacy: a survey of privacy risks and remedies in personalization-based systems. User Model. User-Adap. Inter. **22**(2012), 203–220 (2012)
74. Wang, L., Zhang, J., Yang, H., Chen, Z., Tang, J., Zhang, Z., Chen, X., Lin, Y., Song, R., Wayne X. Zhao, Xu, J., Dou, Z., Wang, J., Wen, J.-R.: User Behavior Simulation with Large Language Model based Agents. arXiv:2306.02552 (2024)
75. Wang, T., Zhao, J., Hu, Z., Yang, X., Ren, X., Kwok-Yan, L.: Local differential privacy for data collection and analysis. Neurocomputing **426**(2021), 114–133 (2021)
76. Wang, W., Feng, F., Nie, L., Chua, T.-S.: User-controllable recommendation against filter bubbles. In: SIGIR, pp. 1251–1261 (2022a)
77. Wang, W., Lin, X., Feng, F., He, X., Chua, T.-S.: Generative recommendation: Towards next-generation recommender paradigm. arXiv:2304.03516 (2023b)
78. Wang, W., Lin, X., Feng, F., He, X., Lin, M., Chua, T.-S.: Causal Representation Learning for Out-of-Distribution Recommendation. In: WWW, pp. 3562–3571. ACM, New York (2022b)
79. Wang, W., Xu, Y., Feng, F., Lin, X., He, X., Chua, T.-S.: Diffusion Recommender Model. In: SIGIR, pp. 832–841. ACM, New York (2023d)
80. Wang, X., He, X., Wang, M., Feng, F., Chua, T.-S.: Neural Graph Collaborative Filtering. In: SIGIR, pp. 165–174. ACM, New York (2019)
81. Wang, X., Tang, X., Zhao, W.X., Wang, J., Wen, J.-R.: Rethinking the evaluation for conversational recommendation in the era of large language models. In: EMNLP (2023c)
82. Wang, Y., Chen, X., Ma, X., Zhou, S., Huang, Z., Wang, Y., Yang, C., He, Y., Yu, J., Yang, P., et al.: Lavie: High-quality video generation with cascaded latent diffusion models. arXiv:2309.15103 (2023a)
83. Wei, A., Haghtalab, N., Steinhardt, J.: Jailbroken: How does LLM safety training fail?. In: NeurIPS, vol. 36 (2024a)
84. Wei, W., Ren, X., Tang, J., Wang, Q., Su, L., Cheng, S., Wang, J., Yin, D., Huang, C.: LLMRec: Large Language Models with Graph Augmentation for Recommendation. In: WSDM. ACM, New York (2023)
85. Wei, W., Ren, X., Tang, J., Wang, Q., Su, L., Cheng, S., Wang, J., Yin, D., Huang, C.: LLMRec: Large language models with graph augmentation for recommendation. In: WSDM, pp. 806–815. ACM, New York (2024b)
86. Wei, Y., Wang, X., Li, Q., Nie, L., Li, Y., Li, X., Chua, T.-S.: Contrastive learning for cold-start recommendation. In: MM, pp. 5382–5390. ACM. New York (2021)
87. Xie, S.M., Santurkar, S., Ma, T., Liang, P.S.: Data selection for language models via importance resampling. In: NeurIPS, vol. 36 (2024c)

88. Xie, Y., Kawaguchi, K., Zhao, Y., Zhao, J.X., Kan, M.-Y., He, J., Xie, M.: Self-evaluation guided beam search for reasoning. In: NeurIPS, vol. 36 (2024a)
89. Xie, Y., Kawaguchi, K., Zhao, Y., Zhao, X., Kan, M.-Y., He, J., Xie, Q.: Decomposition enhances reasoning via self-evaluation guided decoding. In: NeurIPS (2024b)
90. Xu, Y., Wang, W., Feng, F., Ma, Y., Zhang, J., He, X.: DiFashion: Towards Personalized Outfit Generation. In: SIGIR. ACM, New York (2024)
91. Yang, P., Zhou, S., Tao, Q., Loy, C.C.: PGDiff: Guiding Diffusion Models for Versatile Face Restoration via Partial Guidance. In: NeurIPS, vol. 36. Curran Associates, Inc., New York (2024b)
92. Yang, Z., Wu, J., Wang, Z., Wang, X., Yuan, Y., He, X.: Generate What You Prefer: Reshaping Sequential Recommendation via Guided Diffusion. In: NeurIPS, vol. 36. Curran Associates, Inc., New York (2024a)
93. Yu, T., Yao, Y., Zhang, H., He, T., Han, Y., Cui, G., Hu, J., Liu, Z., Zheng, H.-T., Sun, M., et al.: RLHF-V: Towards trustworthy MLLMs via behavior alignment from fine-grained correctional human feedback. arXiv:2312.00849 (2023)
94. Zhang, A., Sheng, L., Chen, Y., Li, H., Deng, Y., Wang, X., Chua, T.-S.: On generative agents in recommendation. In: SIGIR. ACM, New York (2024)
95. Zhang, G.: User-Centric Conversational Recommendation: Adapting the Need of User with Large Language Models. In: RecSys, pp. 1349–1354 (2023)
96. Zhang, J., Bao, K., Wang, W., Zhang, Y., Shi, W., Xu, W., Feng, F., Chua, T.-S.: Prospect Personalized Recommendation on Large Language Model-based Agent Platform. arXiv:2402.18240 (2024a)
97. Zhang, J., Bao, K., Zhang, Y., Wang, W., Feng, F., He, X.: Is ChatGPT fair for recommendation? evaluating fairness in large language model recommendation. In: RecSys, pp. 993–999 (2023a)
98. Zhang, J., Hou, Y., Xie, R., Sun, W., McAuley, J., Zhao, W.X., Lin, L., Wen, J.-R.: AgentCF: Collaborative learning with autonomous language agents for recommender systems. In: WWW. ACM, New York (2024)
99. Zhang, J., Xie, R., Hou, Y., Zhao, W.X., Lin, L., Wen, J.-R.: Recommendation as instruction following: A large language model empowered recommendation approach. arXiv:2305.07001 (2023b)
100. Zhao, J., Wang, W., Xu, Y., Sun, T., Feng, F.: Denoising Diffusion Recommender Model. In: SIGIR. ACM, New York (2024)
101. Zhou, S., Yang, P., Wang, J., Luo, Y., Loy, C.C.: Upscale-A-Video: Temporal-Consistent Diffusion Model for Real-World Video Super-Resolution. In: CVPR. IEEE, New York (2024)
102. Zhou, X., Zafarani, R.: A survey of fake news: fundamental theories, detection methods, and opportunities. CSUR **53**(5), 1–40 (2020)
103. Zhu, H., Ge, H., Gu, X., Zhao, P., Lee, D.L.: Influential Recommender System. In: ICDE, pp. 1406–1419. IEEE, New York (2023a)
104. Zhu, J., Yang, H., He, H., Wang, W., Tuo, Z., Cheng, W.-H., Gao, L., Song, J., Fu, J.: MovieFactory: Automatic movie creation from text using large generative models for language and images. In MM, pp. 9313–9319. ACM, New York (2023b)

Chapter 9
Designing for the Future of Information Access with Generative Information Retrieval

Vanessa Murdock, Chia-Jung Lee, and William Hersh

Abstract Generative Artificial Intelligence (AI) offers powerful tools that fundamentally change the design of information access systems; however, it is unclear how to use them to best serve the needs of people. At present, Large Language Models (LLMs) process natural language (and multi-modal) input and present credible-appearing but often completely untrue multi-modal output. This opens the door to research into how to produce true, complete, relevant information, where and how to design retrieval augmentation to personalize and ground the system, and how to evaluate beyond relevance for truth, completeness, utility, and satisfaction. The applications of generative AI for information-seeking tasks are broad. In this chapter, we present recent developments in four domains that have been well studied in the information retrieval community (education, biomedical, legal, and finance). We follow with a discussion of new challenges (agentic systems) and research areas that are common to most applications of generative AI to information seeking tasks (credibility and veracity, new paradigms for evaluation, and synthetic data generation). The field of Information Retrieval (IR) is at the leading edge of a transformation in how people access information and accomplish tasks. We have the rare opportunity to design and build the future we want to live in.

Vanessa Murdock and Chia-Jung Lee are employed by Amazon Web Services, Inc. The work described in this chapter is not related to their roles in AWS.

V. Murdock (✉) · C.-J. Lee
Amazon, Seattle, WA, USA
e-mail: vmurdock@amazon.com; cjlee@amazon.com

W. Hersh
Oregon Health & Science University, Portland, OR, USA
e-mail: hersh@ohsu.edu

R. W. White, C. Shah (eds.), *Information Access in the Era of Generative AI*,
The Information Retrieval Series 51, https://doi.org/10.1007/978-3-031-73147-1_9

The release of ChatGPT and the subsequent proliferation of LLMs—large and small, open access, and proprietary—caused a collective panic in the research and industrial communities.[1,2] Projects centered on question answering, summarization, translation, information extraction, and recommendation were so greatly simplified it appeared everything could be built on a license to a high-quality LLM. Teams were redirected to figuring out how to use LLMs for their tasks, as companies pivoted to the new paradigm.[3] People asked whether LLMs would replace software engineers and applied scientists [88] and how science would be disrupted [14]. The concern that LLMs would accelerate learning, empowering bad actors to quickly gain new skills in nuclear, biological, or chemical weapons, putting the future of humanity at risk, prompted specific mention in a 2023 White House Executive Order[4] followed by studies by policy organizations [98].

Having a tool as powerful as an LLM makes easy what was previously hard (such as open-domain question answering and summarization) and opens new opportunities previously unreachable. Still, a nagging question remains: what do people really need from an LLM? The initial application (more a demonstration than a viable product) was a chatbot designed to interact in a natural language conversation by text [106]. It is unclear whether or when people want to interact with a system by giving a full-text prompt and then possibly clarifying with another full-text input. People have deftly found ways to reduce the amount of typing in both search (expressing even complex needs in 2–4 keywords) [66] and text conversations (using emojis and abbreviations, dropping punctuation and capitalization) [126]. It is not unreasonable to predict people will continue with abbreviated expressions of their needs and expect the new technology to do a better job of inferring what they want, possibly taking an action on their behalf without being specifically prompted. Fully multi-modal systems may become the norm, enabling people to interact as they wish with gesture, image, video, sound, speech, or text. It is reasonable to expect the system to adapt to the user's preferred mode of interaction and respond naturally, in any modality.

The advancement in LLMs over the last 10 years has been dizzying, but it is unclear how long this pace of innovation will continue, or what the technology is capable of, and what its limitations are [73]. We now have models capable of ingesting vast amounts of data, the outcome of an experiment set in motion by Pasca et al. in the early days of "big data" [107, 108], demonstrating that vastly more data improves the quality of open-domain factoid question answering. It is still

[1] https://assets.researchsquare.com/files/rs-3945065/v1_covered_ce3bfe95-03d0-46c9-8d0a-411088d03f52.pdf?c=1707839911

[2] https://www.nytimes.com/2022/12/05/technology/chatgpt-ai-twitter.html

[3] https://www.technologyreview.com/2024/02/29/1089152/generative-ai-differentiating-disruptors-from-the-disrupted/ visited April 2024.

[4] Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, October 20, 2023. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

unclear how much more data can be ingested and how much more mileage there is to be gained from the current families of LLMs. Today's LLMs are exceptionally good at producing convincing natural language, but they are not reliable sources of truth [166]. Agent systems show promise, as the agentic components can be designed to provide credible, attributable information. It may be that the next big steps forward in generative AI use LLMs as smooth talkers who frequently make up information (i.e., employing them for managing dialogs, rewriting prompts to submit to agents, and producing user-friendly outputs) while employing collections of agents to provide credible, attributable, reliable information.[5]

However, even if the state of the art in language modeling were not to advance, the application of this incredibly powerful family of models to different domains will be ripe for innovation and exploration for a long time. We have the opportunity to design the future of information access, and we should think intentionally about what we want the future to be. Considering all of the tasks an AI could do (creative tasks, evaluations, interpersonal communications, exploration, formulaic tasks, repetitive tasks), we can decide what to offload to an intelligent workhorse and what to reserve for people. In this chapter, we discuss application domains that have been a focus of information retrieval (IR) in the past and discuss the open questions in the new information landscape. While multi-modal LLMs are an important part of the future of information access, in this chapter, we focus on text LLMs. At the end of the chapter, we discuss open problems shared by multiple domains.

9.1 Domains and Applications

9.1.1 Education

The use of generative AI in educational settings has been controversial [42, 157], echoing a similar discourse 25 years prior about the use of computers and the Internet [141]. The concerns center around the risk of exposing children to inappropriate content, causing harm to people's ability to learn to read, write, and converse, and in general contributing to anti-social behavior. Wartella and Jennings [141] noted that the same concerns were echoed in the 1940s with the introduction of television, in the 1920s with the introduction of radios, and in the 1900s with the introduction of film.

In the year following the release of ChatGPT,[6] a Pew Research Center study found roughly 13% of teens in the USA used ChatGPT for homework.[7] A Walton

[5] Note that a RAG system is a simple agent system where the RAG search system and knowledge base are an agentic component.

[6] https://openai.com/blog/chatgpt visited April 2024.

[7] https://www.pewresearch.org/short-reads/2023/11/16/about-1-in-5-us-teens-whove-heard-of-chatgpt-have-used-it-for-schoolwork/ visited April 2024.

Family Foundation study identified that 84% of teachers who used ChatGPT in their classroom said it had a positive impact.[8] In spite of early decisions to ban the technology from use by students,[9] many educators found ways to integrate it into the curriculum,[10] and the US Department of Education released guidance on the use of artificial intelligence in the classroom [19].

Information access is a key concern in the field of education, and information retrieval research from the mid-twentieth century onwards has focused on ways to improve information literacy and learning outcomes and to provide learning support such as designing search engines specifically for young people, adjusting search results according to reading level, and designing methods to derive or generate supplementary materials such as exam questions and adjunct questions to improve vocabulary and reading comprehension.

Search is an important instructional tool, and research has focused on understanding the specific needs of children, creating child-friendly search engine designs [13, 38, 49, 84]. A key part of providing search results for educational purposes includes estimating the reading level of a document [28, 29, 57]. Anuyah et al. [3] and Syed and Collins-Thompson [121] studied specific optimizations for classroom use, and Usta et al. [134] explored learning to rank for educational search. An informative overview of research in IR for children from 2000 to 2020 is in Huibers et al. [64].

In terms of online learning, Moraes et al. [97] found that short instructional videos are more effective than Search as an instructional technique (24% improvement in learning gains compared to search), but video instruction paired with Search is even better (yielding a 41% improvement over search alone). High-quality generative multi-modal models create the opportunity to design rich multi-modal learning experiences that combine video, Search, and other modalities to optimize learning outcomes.

Li et al. [80] point out that intelligent education systems have two main challenges. One is subject matter expertise, which entails giving complete and true information, with credible attribution, where so far LLMs have fallen short. The other is to tailor the content and presentation to the individual learner, requiring personalization, and an adaptive approach to session context. Lallé and Conati [77] propose an association rule mining approach to identify video-watching behaviors that are less conducive to learning, and adaptively tailor video recommendations to improve learning outcomes.

Retrieval practice (i.e., practice testing), in which the learner is asked to periodically recall information from memory that they have just read, is widely accepted

[8] https://www.waltonfamilyfoundation.org/learning/teachers-parents-report-positive-impact-of-chatgpt-on-teaching-and-learning visited April 2024.

[9] https://www.nbcnews.com/tech/tech-news/new-york-city-public-schools-ban-chatgpt-devices-networks-rcna64446 visited April 2024.

[10] https://www.politico.com/news/2023/08/23/chatgpt-ai-chatbots-in-classrooms-00111662 visited April 2024.

as improving reading comprehension. Francis Bacon [10] and John Locke [89] both advocated for retrieval practice in the 1600s, and it remains a regular practice in education today.

Prior to the development of LLMs, the research literature focused on statistical question formulation given an input text [56]. Brown et al. [18] use vocabulary assessment to determine a student's reading level and present a system to generate vocabulary questions based on question templates. An early example of automatic question generation for reading comprehension used a pre-defined vocabulary and pattern matching heuristics [144].

Generative AI has been employed to generate exam questions and adjunct questions to check students' progress as they read a text. Syed et al. [122] employ gaze tracking and automatic question generation to improve learning outcomes. Du et al. [39] employ a Recurrent Neural Network (RNN) encoder-decoder architecture to generate questions for reading comprehension, evaluated for the naturalness of the question and the difficulty to answer (but not on improvements to reading comprehension).

Somewhat surprisingly, in a study of online learning, Davis et al. [34] found that practice questions had no effect on information retention in online courses. Zhu et al. [164] found that adjunct questions increase the time spent reading, and while results are mixed whether fact questions improve reading comprehension, synthesis questions improved students' coverage of topics in essay questions. Synthesizing multiple sources of information and distilling them into a short text is a task made easier with generative AI. A natural follow-on to this line of research is to investigate which types of adjunct information (which types of questions or which types of rich multi-modal experiences) do improve information retention and synthesis and learning outcomes more generally.

ChatGPT is notable for the fluency of its dialog. This suggests that it might be ideal for learning a language. There is a lot of material online associated with basic language courses for languages most commonly spoken as a second language. The bias in generative AI toward English content will be a challenge for the use of generative AI in other languages,[11] as it may result in a bias toward second languages popular among English speakers. Amin [2] explores the topic of using ChatGPT for second-language learning. From an information retrieval perspective, there is an opportunity to generate dialogues for language learning, which are tailored not just to the learner's level of advancement but also to their topical interests, hobbies, or as a wholistic approach to learning a new topic while learning a new language.

An important aspect of intelligent tutoring is tailoring the information to the level of experience with a topic and reading level. Initial reports of style transfer for generative AI showed astonishing demonstrations of text generated in the style

[11] https://www.wired.com/story/chatgpt-non-english-languages-ai-revolution/ visited April 2024.

of Dr. Seuss,[12] Ernest Hemingway,[13] Shakespeare,[14] and others. There has been a lot of interest in the research community in style transfer for creative purposes, for improving communication (such as writing in a more formal or business style), and for mitigating social issues (such as improving the fluency of the text for second-language learners) [62, 68]. There has been work focusing on personalized style [12, 102] and writing in a simpler style (e.g., reducing technical jargon) [114].

Style transfer for reading level is an open problem, and generative information retrieval techniques open up many possibilities in personalizing the depth of information presented, the reading level of the student, the learning strategy, and the curriculum. This has the potential to expand the types of information available to students and language learners, where the system is dynamic in the content, style, and tone of the writing, rather than relying solely on pre-written documents.

9.1.2 Biomedical

There are many information needs in the biomedical domain, simple and complex, that require information retrieval solutions [59]. Prior to emergence of generative AI, search in the biomedical domain was a relatively mature technology. In addition to general commercial Web search engines, users (i.e., patients, clinicians, and researchers) turned to domain-specific search systems, such as PubMed, MedlinePlus, systems developed by the US National Library of Medicine, and others. Information-seeking tasks in the biomedical domain (question answering, summarization, information extraction, search) will be greatly facilitated by LLMs [125, 127]. LLMs are helpful in each step of the search process (querying, reformulation, retrieval, ranking, and distillation of relevant information) [163]. Further, a general-purpose LLM may be fine-tuned to a specific topic, allowing for highly specialized systems. For example, Liu et al. [86] describe fine-tuning LLMs for radiology; Tan et al. [123] fine-tune LLMs for ophthalmology; Tan et al. [124] fine-tune for traditional Chinese medicine.

Users of IR systems, particularly academics, have concerns for authoritativeness (who authored a piece of information), timeliness (when it was authored), and context (of the questions and supporting evidence). Use cases for biomedical search include clinical (patient-care questions), research (methods and insights), and teaching (synthesizing knowledge for pedagogy). One basic question about generative AI and Search is how they compare in meeting information needs. In

[12] https://medium.com/@alezafreeman/i-asked-chatgpt-to-write-poem-like-dr-seuss-5e303ee4f4ee visited April 2024.

[13] https://medium.com/writers-blokke/chatgpt-vs-hemingway-editor-the-smackdown-da6f220a246c visited April 2024.

[14] https://www.zdnet.com/article/i-used-chatgpt-to-rewrite-my-text-in-the-style-of-shakespeare-c3po-and-harry-potter/ visited April 2024.

the biomedical domain, LLMs have been found to be highly effective in answering clinical questions [48], taking medical board exams [105], and solving clinical cases [132]. Several studies have assessed the value of information output by LLMs compared to search engines. Hopkins et al. [61] found that ChatGPT was more informative than Google snippets for four cancer questions. Van Bulck and Moons [135] compared the output of ChatGPT to Google, evaluated by 20 experts in the domains of congenital heart disease, atrial fibrillation, heart failure, and cholesterol. Responses by ChatGPT were deemed trustworthy and valuable, with few experts considering them dangerous. Comparing information from ChatGPT to information from Google Search, 40% rated ChatGPT as more valuable, 45% as similarly valuable, and 15% as less valuable, although few details were provided about the comparisons.

One role for IR in generative AI is to add more recent content to LLMs. Training LLMs is a very resource-intensive process and can only be done on an intermittent basis. Retrieval-Augmented Generation (RAG) allows LLMs to incorporate more recent information and attribute sources to their output [96, 145]. The small amount of research on RAG in biomedicine shows mixed results. Koopman and Zuccon [75] found that adding Web Search content to ChatGPT prompts reduced the accuracy of correct answers using Text Retrieval Conference (TREC) Health Misinformation Track data. However, Zakka et al. [159] developed an LLM framework called Almanac that employed RAG and was found to improve question answering over standard LLMs across axes of factuality, completeness, user preference, and adversarial safety.

In terms of the reverse process (i.e., generation-augmented retrieval), or Search systems improved by generative AI methods, there is likewise little research. Jin et al. [69] developed MedCPT, which uses an encoder model to train on 225 million query-click pairs in PubMed logs, leading to small improvements over BM25. Wang et al. [139] used GPT-4 to generate Boolean queries for systematic review search, obtaining improved precision but at a cost to recall. Jiang et al. [67] proposed improving dynamic retrieval of electronic health record notes by predicting which notes are most likely to be read.

Another long-standing biomedical IR use case is the matching of patients to clinical studies based on the data in their Electronic Health Record (EHR). This use case was developed in the Text REtrieval Conference (TREC) Medical Records Track in in 2011–2012 [136, 137] and subsequent work [21, 147]. More recent work incorporates LLMs via a variety of methods and datasets [37, 69, 76, 103, 133, 146].

One important task at the intersection of IR and generative AI assesses how well generative AI systems provide attribution for what they say. Rashkin et al. [115] developed approaches for measuring attribution in natural language generation models. Recent research has looked at LLMs generating text with citations [44], source attribution, conscious incompetence [81], and retrieving supporting evidence for generative question answering [65].

We do know that LLMs fall short when it comes to citations. One analysis of fabrication and errors in bibliographic citations asked ChatGPT to produce short literature reviews on 42 multidisciplinary topics [138]. It was found that 55% of GPT-3.5 citations and 18% of GPT-4 citations were fabricated and that 43% of

real (non-fabricated) GPT-3.5 citations and 24% of real GPT-4 citations included substantive errors. Another study prompting ChatGPT to cite articles about learning health systems found that GPT-3.5 cited 98% incorrect; GPT-4 cited more with only 20.6% incorrect [24]. This led Gusenbauer [52] to advocate that AI tools should be audited before their widespread use in scientific research, especially systematic reviews. Heidt [55] noted that LLMs may be useful in drawing connections in the scientific literature, but we must beware of biases in papers that may be perpetuated by each component of the system. Jin et al. [69] pointed out that LLMs do not consult any source of truth and proposed a retrieve-summarize-verify paradigm.

A recent study of resource attribution in the biomedical domain compared several commercial LLMs, one with RAG, in their ability to cite relevant references for their claims [150]. Clinical questions for prompting were generated from several well-known Web health information sources. The output was assessed by clinician experts for URL source validity, statement-level support of claims, and response-level support. The best LLM was Microsoft Copilot,[15] which includes RAG from the associated Bing search engine. Copilot had near-perfect URL source validity, 70% statement-level support, and 54% resource-level support. CoPilot had the lowest rate of citing no references in response to prompts. Other issues included grounded vs. correct claims and sources behind firewalls.

The research so far at the intersection of LLMs and IR in biomedicine shows much potential for generative AI in this process. However, it will be key to maintain authority, veracity, and timeliness of output from such systems.

9.1.3 Legal

IR has been pivotal in addressing legal domain challenges, serving as an essential tool for legal professionals to efficiently carry out a multitude of tasks. Retrieval algorithms, with the inclusion of a controlled vocabulary and a juridical thesaurus, have been developed to retrieve relevant jurisprudential precedence to assist the decision-making process [32]. In 2014, the Competition on Legal Information Extraction/Entailment (COLIEE[16]) was first introduced, inviting the development of informatics technologies on the legal question-answering task. With BERT published in late 2018, dense representations/encoders such as Doc2Vec, BERT, or ELMO, in conjunction with traditional methods such as BM25 and TF-IDF, were studied for retrieving legal cases that should be cited for a query case, as well as retrieving civil code articles for answering bar exam questions [113]. Echoing the debate on the effectiveness of early neural models in Search ranking [154], the vanilla application of transformers to legal retrieval did not improve on traditional methods [9].

[15] https://copilot.microsoft.com/ visited July 2024.

[16] https://webdocs.cs.ualberta.ca/~miyoung2/jurisin_task/index.html visited April 2024

The convergence of GenAI and information access has transformed the landscape of the legal domain, leading to substantial improvements on several legal tasks [120]. Tested on three multilingual datasets, Trautmann et al. [129] showed that foundation models paired with legal prompt engineering delivers promising results on the task of legal judgment prediction. Other studies reiterated the importance of prompting, where carefully designed few-shot [15] and chain-of-thought [158] prompts advanced the effectiveness in legal reasoning tasks. While LLMs are considered to encapsulate and compress world knowledge in a generic sense, the success of legal applications relies on the faithfulness and accuracy of the responses. Niklaus et al. [104] curated a large-scale legal instruction-tuning dataset, covering 17 jurisdictions, 24 languages, and 12 million examples, to update LLMs with domain knowledge. Their results suggest that domain-specific pre-training and instruction tuning improve performance on the LegalBench benchmarking task[17] [51], compared to general-purpose models. However, the effect did not generalize across the board, echoing the finding that accuracy and high-quality writing remain a challenge despite the advancements [130].

GenAI and LLMs have also been employed to provide legal advice, such as supporting law professors with service- and teaching-related tasks [111] and offering quasi-expert legal advice services [91]. The work of Wu et al. [151] shows the potential for LLMs and domain-specific models working in a collaborative mode to assist users in predicting legal judgments. They designed a staged framework, where domain-specific models were employed to provide labels and retrieve relevant legal precedent cases, after which LLMs leveraged the cases as in-context examples to make the final prediction.

From proof-of-concept studies to real-world practice, emerging start-ups have brought streamlined GenAI capabilities to the consumer space. Paxton[18] offers assisted contract review, legal document drafting, interactive file analysis and legal research, and automatic Boolean query generation for precision-focused retrieval. AI Lawyer[19] provides consumers with AI-assisted legal consultation while helping law professionals automate legal research and paperwork. Spellbook[20] embeds legal documents within Microsoft Word, facilitating the holistic integration of contract reviewing, legal term analysis, and assisted writing. Noting that truthfulness significantly influences the usability of AI-enhanced legal applications, Lexis+AI [21] includes linked legal citations as supportive evidence in the interactions with users in conversational search, drafting, summarization, and document analysis.

Existing legal applications continue to evolve as the legal landscape stands poised for innovation. Today, legal document understanding and composition, including contracts, wills, trusts, deeds, court orders, and court judgments, still require time-

[17] https://huggingface.co/datasets/nguha/legalbench visited July 2024

[18] https://www.paxton.ai/ visited April 2024.

[19] https://ailawyer.pro/ visited April 2024.

[20] https://www.spellbook.legal/ visited July 2024.

[21] https://www.lexisnexis.com/en-us/products/lexis-plus-ai.page visited April 2024.

consuming, back-and-forth processes to complete. The development of generative AI and IR techniques facilitates a future for the general public where complex legal jargon can be explained simply, relevant precedent can be retrieved for decision making purposes, cross-domain knowledge can be joined, and contextual customization can be baked in. For organizations, businesses, and law professionals, opportunities lie in legal risk assessments and research for a variety of tasks. Due to the complexity and dynamism of regulation, these advanced technologies may be helpful for analyzing regulatory frameworks and industry standards to identify potential risks and ensure compliance. This enables legal professionals to anticipate implications, to advise clients on risk mitigation strategies.

The degree to which AI technologies can replace law professionals is the subject of intense debate[22] [99]. Concerns arise with the reliance on LLMs in high-stakes circumstances, touching upon a wide array of aspects such as ethics, policy, doctrine, and more. To study when and why such technologies should or should not be used, Cheong et al. [26] convened workshops with 20 legal experts and elicited guidance on appropriate AI assistance for sample use cases. Their findings advocated for a focus on novel legal problems, for example, that users' interactions with LLMs are not protected by attorney-client confidentiality or bound to professional ethics that guard against conflicted counsel or poor-quality advice. This work sheds light on the need for more thinking into the design of AI in a professional context. For instance, instead of advising actions or decisions, AI agents might offer to polish users' questions and offer relevant facts. Muhlenbach and Sayn [99] delivered an ethical matrix highlighting ethical aspects and values in response to the introduction of AI-based models of court decisions. They concluded that, ultimately, while predictive justice tools have potential, they also have significant limitations, and legal professionals must be able to use AI tools as aids.

9.1.4 Finance

Retrieving and accessing financial information is a common task among finance professionals. Ariannezhad et al. [5] paint a rich picture of how users seek information among financial statements and disclosures hosted on the Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system, the primary resource for accessing company filings. They uncovered that the distributions of filing access patterns are skewed, where individual users are interested in a limited number of companies, shedding light on design options to better support user interactions (e.g., via filing recommendation). Financial news articles serve as another important source of information for investors. Lavrenko et al. [79] presented a pioneering work

[22] https://www.forbes.com/sites/forbestechcouncil/2023/05/25/will-ai-replace-lawyers/?sh=38f20bd83124 visited April 2024.

on employing language models for financial news recommendation, by correlating the content of news stories with trends in financial time series. They demonstrated that financial news articles could be used to predict forthcoming patterns in stock prices.

Traditional approaches to financial IR rely on pipelines of machine-learned or engineered components to address fine-grained tasks in the entire system [20, 85, 112, 119]. As financial documents pose domain-specific challenges, including the presence of a large amount of tables and (monetary) numbers, Plachouras et al. [112] proposed a search system with a query understanding component that conducts both entity tagging and intent ranking. They were also early adopters of natural language generation techniques to aggregate and synthesize retrieved information to answers, albeit following a simple template-based approach. Sumithra and Sridhar [119] leveraged named entity recognition and entity relation linking to process financial documents, leading to a triplet representation for matching natural language queries through query conversation. Ceccarelli et al. [20] analyzed the ranking of posts from the online social media platform X (formerly Twitter)[23] for the financial community, suggesting that factors such as popularity in traditional media, speculation, and capture of fleeting information are critical for a more relevant and engaging experience with the social media posts.

Machine Learning (ML) and language techniques have been used by financial services and applications for over a decade, from fraud detection [142], risk assessment [131], market analysis [1, 40], and expense predictions [156] to product and service recommendations. Agrawal et al. [1] proposed a technique for summarizing financial reports toward buy-or-sell goals based on hierarchical neural models, while Fan et al. [40] proposed an attentive graph convolutional network approach for organizing financial documents into pre-defined categories to facilitate investment advising services. Tsai and Wang [131] demonstrated that the use of text mining and learning-to-rank techniques can be effective for assessing the risks of publicly traded companies.

The financial industry is on the cusp of an unprecedented technological transformation. With generative language capabilities, Choi et al. [27] introduced an LLM-based conversational financial information retrieval model tailored for query intent classification and knowledge base labeling. Trained with internal data, BloombergGPT was introduced by Wu et al. [149], and Bloomberg has integrated conversational capabilities within its terminal for answering financial questions. Market sentiment tracking is crucial for decision-makers in the financial sector. Chen et al. [25] investigated how ChatGPT might capture corporate sentiment toward environmental policy based on their financial statements. Their findings suggested that sentiment scores generated by ChatGPT are predictive of a firm's risk-management capabilities and stock return performance. In forecasting, Lopez-Lira and Tang [90] explored the feasibility of predicting subsequent daily stock returns using ChatGPT with promising results.

[23] https://x.com/ visited July 2024.

In late 2018, Morgan Stanley released an AI-powered tool Next Best Action[24] for financial advisors, assisting them to provide personalized financial advice to clients based on life events. With recent advances in generative AI, the viability of building 24/7 end-user-friendly financial advisors that are up-to-date and equipped with an individual's context is closer to reality. Magnifi[25] uses ChatGPT and other technologies to provide personalized, data-driven investment advice, performing like a brokerage where one can directly trade stocks and exchange-traded funds (ETFs).[26] Intuit Assist[27] connects existing Intuit product offerings, assisting users or small businesses in filing taxes, selecting higher reward programs based on personal spending, and catching up on selling insights with recommended next steps. Extending the autonomous modality, it is not hard to imagine a future multi-agent environment where financial decisions can be made with minimal human intervention, in which actions are executed based on an individual's goals and risk tolerance.

Financial institutions deal with a wide array of data types, ranging from structured or time-series data (e.g., trades, prices, transactions), textual content (e.g., institutional guidance, news articles, reports), social media data, and graphical and visual information (e.g., satellite images of economic activities or videos of company press releases and media interviews). It remains an open research area how future information systems should integrate diverse data sources, in a user-friendly way with accurate, timely, and actionable financial analyses. On top of public information sources, it is anticipated that the user experience will be enhanced with personalization and understanding the user context.

These new opportunities come with elevated risk. Financial institutions operating AI-driven systems must navigate a complex regulatory landscape governing data privacy, consumer protection, and cybersecurity threats [152]. Although financial services accumulate a large volume of data, they are often stored in silos within organizations, and data sharing is strictly regulated. Synthetic data generation in the financial domain presents an opportunity to mitigate this limitation. Synthetic data that reflects real data distributions has the potential to advance the financial system's ability to comprehend and integrate cross-environment information, to better serve individuals and organizations. Assefa et al. [7] highlighted the importance of measuring the similarities between real and generated datasets while ensuring the generative process satisfies any privacy constraints. Koenecke and Varian [74] also discussed several ways to synthesize data for economic analyses, including using generative adversarial networks and LLMs.

[24] https://www.cnbc.com/2018/11/20/morgan-stanley-launches-new-advisory-technology-platform.html

[25] https://magnifi.com/

[26] https://www.cnbc.com/2023/04/27/chatgpt-meets-robinhood-new-app-features-ai-powered-portfolio-mentor-.html

[27] https://www.intuit.com/intuitassist/ visited April 2024

9.2 New Challenges

9.2.1 Agent Systems

Agent systems, interchangeably referred to as other terms such as LLM agents or agentic workflows, are believed to drive massive AI progress for the years to come.[28,29] They are on the rise in domains such as health [47], education [16, 33], legal [54], and finance [72, 82, 140]. With advanced language models sitting at the core, these systems are typically characterized by the autonomous nature with limited (or no) human intervention in completing tasks or goals. Generally speaking, their capabilities can be summarized as the following[30] [93, 153]:

- **Reflection and refinement** [50, 92]. An agent can reason about past actions, to learn from mistakes and refine themselves for future steps, with the goal of improving the outcome of a target task.
- **Planning** [63]. Related to the abundant literature in productivity [143, 160], an agent can break down large, high-complexity tasks into smaller, more manageable sub-tasks for a more successful execution. The act of planning outlines the steps needed, wherein alternative, parallel intermediate steps can be contrasted or debated to enhance goal completion.
- **Memory** [162]. An agent can access external memory in addition to its internal memory. In-context learning can be considered short-term memory for the agent, while long-term memory via external vector stores can provide the agent with the capability to retain and recall information over extended periods, as with retrieval-augmented generation.
- **Tool use** [45, 110]. An agent is commonly equipped with a variety of tools and can learn which to use to enhance the quality and accuracy of the outcome. Tools can be ML or non-ML functions, application programming interface (APIs), or services, such as calculators, code interpreters, holistic or vertical-specific search systems, and ticket-booking systems.
- **Multi-agent collaboration** [60, 148]. A multi-agent approach involves deploying multiple agents assigned with different roles and capabilities to execute and collectively complete a goal together. For instance, given the task of planning a wedding, a multi-agent approach may break down the task and distribute the load to a financial agent for budget control, a venue agent for time and location coordination, or a secretary agent that sits on the top of every detail for communications. A key distinction is whether agents communicate directly to each other or with a central orchestrator.

[28] https://www.deeplearning.ai/the-batch/how-agents-can-improve-llm-performance/ visited April 2024.

[29] https://www.technologyreview.com/2024/05/01/1091979/sam-altman-says-helpful-agents-are-poised-to-become-ais-killer-function/ visited April 2024.

[30] https://lilianweng.github.io/posts/2023-06-23-agent/ visited April 2024.

The precursor to agent systems in IR is aggregated search [78, 101], where a search query is submitted to multiple (disparate, multi-modal) information sources (or search verticals such as blogs, video, social media posts, audio, Web pages, etc.), and the results from each source are assembled into a single response to the query. The primary research questions as the field of aggregated search advanced were how to select the sources given a query; how to weight the results from each source (or how to compare their relevance to the query given that similarity scores from each source are not comparable); how to present them to the user in a coherent way; which items to interleave, merge, or summarize; and how to assess the quality of the aggregated result. In the new generative AI paradigm, we consider the search engine as the orchestrator and each source (and its vertical search system) an agent.

More recently, Pan et al. [109] introduced an agent system for information seeking that employs an LLM as its cognitive core. Upon receiving a user query, the agent LLM orchestrates the following steps: (1) Updates and retrieves memory (e.g., past conversations or completed tasks); (2) conducts planning (i.e., a comprehensive prompt for generating sub-tasks and associated tools); (3) executes tools (i.e., invoking tools with arguments and updating memory); and (4) draws conclusions (i.e., a final response to the user's query). To fulfill user queries, an array of information-oriented tools were made available to the agent, including Web and Wikipedia search, Web content understanding (e.g., Web pages, aspects, videos), as well as time-aware functional tools such as calendar, holiday, and weather details. The experiments showed that agent systems typically yield better results (in terms of truthfulness and quality) than directly querying the LLM. Fine-tuning open-sourced models with generated templates designed for agent systems was shown more effective than using them out of the box.

Zong et al. [165] introduced a framework that uses an LLM-based agent with multiple roles for question-answering tasks based on knowledge bases. Specifically, the agent is assigned three distinct roles: a generalist adept at small tasks by the given examples, a decision-maker proficient at identifying options and selecting candidates, and an advisor skilled at providing answers using accessible knowledge. These roles collaborate to conduct question parsing, Uniform Resource Identifier (URI) linking, query construction, and answer generation. Their follow-up analysis suggested that the generalist benefits from quality data more than quantity; the decision-maker could be hurt by considering too many URI candidates; and the advisor should orchestrate to retry previous steps if the other two roles did not yield sufficient information. The design of future information acquisition agent systems includes investigations about how to store and retrieve memory, which roles should be created to balance effectiveness and efficiency, and how to navigate complex, multi-modal, open-domain, or domain-specific corpora to respond and execute. As agent systems for information seeking evolve, so do the open research questions.

Evaluating an agent system is intrinsically challenging and nuanced, leading to an increasing need for reliable benchmark paradigms. Arabzadeh et al. [4] proposed a separate LLM agent workflow designed for evaluation, to assess the alignment between the behavior of an agentic application and user goals. This framework, AgentEval, comprises two agents: CriticAgent suggests criteria based

on task descriptions and proposed solutions, and QuantifierAgent verifies how well the solutions align with the criteria.

Another line of work focuses on establishing realistic, challenging benchmark datasets, which aim to test agents' multi-step problem-solving abilities in multi-modal environments [35, 46, 87, 95]. Deng et al. [35] introduced MIND2WEB, consisting of instances that reflect the sequence of actions carried out by an agent in order to complete a given task on a Web site. For example, given a task *show me the reviews for the auto repair business closest to 10002*, the sequence of actions are identified as *search "auto repair"*, *click button on auto repair*, *type 10002 in textbox*, $\cdots$, *click read reviews*, etc. MIND2WEB covered 2000 open-ended tasks collected via human annotators on Amazon Mechanical Turk, from 137 Web sites spanning 31 domains such as airline travel, housing information, and more.

With a similar goal, Liu et al. [87] presented AGENTBENCH, a benchmark that consists of eight distinct environments to assess LLM-as-Agent's reasoning and decision-making abilities. Their results revealed a significant disparity in performance between proprietary and smaller open-sourced models, where the poor performance of the smaller models was due to poor long-term reasoning, decision-making, and instruction-following abilities.

Opportunities associated with agent systems are accompanied by risks [11]. The transmission of private information (e.g., user-provided texts or images) to both LLMs and tools poses privacy risks to users [161]. As systems are composed of multiple components, they are vulnerable to security risks, such as a malicious attacker injecting a backdoor into the LLM agents to adversely affect user experiences through interactive reasoning traces [155]. Agent systems also introduce safety concerns. Based on an emulated framework, Ruan et al. [117] identified safety risks correlated with under-specified information and erroneous tool executions. For example, an agent could use a fabricated recipient bank account to transfer money or deliberately miscalculate the amount of money to transfer to the intended recipient. Compared to LLMs, LLM-based agents are more prone to harmful behaviors due to domino effects. On AdvBench, Tian et al. [128] showed that Attack Success Rate (ASR) of harmful behaviors increases with the number of agents. In addition, when a higher-level agent disseminates harmful information, it significantly increases the likelihood of inducing similar harmful behaviors in lower-level agents. More broadly, the impact of AI agents automating complex, high-stakes tasks can disempower individuals in the space of decision-making or can create delayed, hard-to-notice harms given agents' optimization on long-horizon goals [22, 23].

9.2.2 Common Threads

IR up to now has been the study of how people access information and how to design systems to help them. The focus has been primarily on finding relevant documents, passages, or sentences and (if needed) extracting key nuggets of information.

Generative AI turns the problem on its head, as LLMs are capable of generating synthesized information. The challenge for IR now is how to generate relevant information to satisfy the user's needs, rather than identifying existing information. Due to the fluency of the language, the LLM sounds more credible than it is. As Metzler et al. [94] put it, LLMs are dilettantes where people need experts.

Previously information credibility was handled primarily by finding authoritative source documents [17, 71], and letting the user decide whether the information was credible. Topics such as information provenance and attribution have a new urgency, due to the LLMs propensity for inventing information from whole cloth [8].

Research on veracity was primarily centered on Question Answering. The veracity of answers to factoid questions was handled by finding the same answers from multiple sources [83] or from carefully constructed knowledge bases (augmented with data from external sources) [118] or from repositories of human-answered questions via community question answering sites such as Yahoo! Answers[31] [70]. In the current research, veracity is often measured as faithfulness (the answer is in agreement with a reference text), which has its precursors in passage retrieval for question answering [30, 31, 100, 116].

Generative IR requires new evaluation paradigms, because the information is generated rather than identified or extracted. Generative models are typically aligned to a policy determined by the model designer. The policy reflects what an appropriate response is. For example, model alignment may prevent the model from dispensing medical or legal advice or from using profanity.

Red teaming (a suite of techniques to evaluate a system's security based on how easy it is to break the security) shows promise as a method for evaluating a model's alignment [43]. However, red teaming, which consists of interactive sessions designed to elicit an inappropriate response from the model, itself is not scientific. Red teaming will identify gaps that can be patched, but it is not possible to measure whether the model as a whole is improving as a result of red teaming or how to measure the effectiveness of the red teaming itself. Often red teaming is measured in terms of success rate (how often did the red teamer convince the model to respond inappropriately), but since each red teaming session is interactive and based on the skill of the red teamer, it is not possible to determine whether the success rate was due to the red teamer's skill or due to the model's inherent weakness. It does not consider the relative risk of each gap to the business or the end user. This becomes even more critical as the industry turns to automating the red teaming process. Without a method to measure the effectiveness of a red teaming effort, there is no way to know whether the automation of the process is effective or to estimate its return on investment or to assess whether the automated red teaming is increasingly challenging to the system as the system matures. Putting

[31] Yahoo! Answers was a Web site where users could ask questions or answer the questions of other people and upvote questions or answers to increase their visibility. The site was shut down in 2021.

red teaming on a scientific footing, and exploring more effective ways to probe the model alignment, is a new challenge.

For many downstream tasks, there is no existing data for fine-tuning or evaluation, so the data must be generated, often by other LLMs [41]. For example, Hämäläinen et al. [53] explore the use of LLMs to generate synthetic user research data. Synthetic data has been investigated in a large number of domains (e.g., for tabular healthcare data [58], for traffic sign recognition [36], for finance applications [6], and more). Common approaches include prompting an LLM to generate the data directly (using zero shot or few shot prompting) or using the LLM to augment the data with similar examples.

Beyond these examples, we have seen in this chapter and previous chapters that there are several topics of research that cover multiple domains, such as fine-grained personalization, effective use of user session context, style transfer, subjective descriptions (e.g., of items to be recommended). As generative IR systems produce fluent dialog interactions, we have new paradigms for information access available to us, such as multi-modal systems, mixed initiative systems (where the system takes an action on behalf of the user), agent architectures for reasoning, planning and decision-making, and true interactive information seeking sessions, where the Generative IR system is more of a partner than a tool. All of these topics will present research opportunities for a long time to come.

References

1. Agrawal, Y., Anand, V., Arunachalam, S., Varma, V.: Hierarchical model for goal guided summarization of annual financial reports. In: Companion Proceedings of the Web Conference 2021, WWW '21, pp. 247–254 (2021)
2. Amin, M.Y.M.: Ai and ChatGPT in language teaching: Enhancing EFL classroom support and transforming assessment techniques. Int. J. Higher Educ. Pedagogies **4**(4), 1–15 (2023)
3. Anuyah, O., Milton, A., Green, M., Pera, M.S.: An empirical analysis of search engines' response to web search queries associated with the classroom setting. Aslib J. Inform. Manag. **72**(1), 88–111 (2020)
4. Arabzadeh, N., Kiseleva, J., Wu, Q., Wang, C., Awadallah, A., Dibia, V., Fourney, A., Clarke, C.: Towards better human-agent alignment: assessing task utility in LLM-powered applications (2024). arXiv preprint arXiv:2402.09015
5. Ariannezhad, M., Yahya, M., Meij, E., Schelter, S., de Rijke, M.: Understanding financial information seeking behavior from user interactions with company filings. In: Companion Proceedings of the Web Conference 2022, WWW '22, pp. 586–594 (2022)
6. Assefa, S.A., Dervovic, D., Mahfouz, M., Tillman, R.E., Reddy, P., Veloso, M.: Generating synthetic data in finance: opportunities, challenges and pitfalls. In: Proceedings of the First ACM International Conference on AI in Finance, pp. 1–8 (2020)
7. Assefa, S.A., Dervovic, D., Mahfouz, M., Tillman, R.E., Reddy, P., Veloso, M.: Generating synthetic data in finance: opportunities, challenges and pitfalls. In: Proceedings of the First ACM International Conference on AI in Finance, ICAIF '20. Association for Computing Machinery, New York (2021)
8. Azaria, A., Mitchell, T.: The internal state of an LLM knows when its lying (2023). arXiv preprint arXiv:2304.13734

9. Gain, B., Bandyopadhyay, D., Saikh, T., Ekbal, A.: Iitp in coliee@icail 2019: Legal information retrieval using bm25 and bert (2019)
10. Bacon, F.: Novum organum, 1620. https://constitution.org/2-Authors/bacon/nov_org.htm
11. Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrell, T., Harari, Y.N., Zhang, Y.Q., Xue, L., Shalev-Shwartz, S., Hadfield, G., Clune, J., Maharaj, T., Hutter, F., Baydin, A.G., McIlraith, S., Gao, Q., Acharya, A., Krueger, D., Dragan, A., Torr, P., Russell, S., Kahneman, D., Brauner, J., Mindermann, S.: Managing extreme ai risks amid rapid progress. Science **384**(6698), 842–845 (2024)
12. Bhandarkar, A., Wilson, R., Swarup, A., Woodard, D.: Emulating author style: a feasibility study of prompt-enabled text stylization with off-the-shelf LLMs. In: Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024), pp. 76–82 (2024)
13. Bilal, D., Boehm, M.: Towards new methodologies for assessing relevance of information retrieval from web search engines on children's queries. Qual. Quant. Methods Libraries **2**(1), 93–100 (2013)
14. Birhane, A., Kasirzadeh, A., Leslie, D., Wachter, S.: Science in the age of large language models. Nat. Rev. Phys. **5**, 277–280 (2023). https://doi.org/10.1038/s42254-023-00581-4
15. Blair-Stanek, A., Holzenberger, N., Van Durme, B.: Can gpt-3 perform statutory reasoning? In: Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law, pp. 22–31 (2023)
16. Boiko, D.A., MacKnight, R., Gomes, G.: Emergent autonomous scientific research capabilities of large language models (2023). arXiv preprint arXiv:2304.05332
17. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Comput. Netw. ISDN Syst. **30** (1–7), 107–117 (1998)
18. Brown, J., Frishkoff, G., Eskenazi, M.: Automatic question generation for vocabulary assessment. In: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, pp. 819–826 (2005)
19. Cardona, M.A., Rodríguez, R.J., Ishmael, K.: Artificial intelligence and the future of teaching and learning: insights and recommendations. U.S. Department of Education, Office of Educational Technology, 2023. https://tech.ed.gov/files/2023/05/ai-future-of-teaching-and-learning-report.pdf
20. Ceccarelli, D., Nidito, F., Osborne, M.: Ranking financial tweets. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '16, pp. 527–528 (2016)
21. Chamberlin, S.R., Bedrick, S.D., Cohen, A.M., Wang, Y., Wen, A., Liu, S., Liu, H., Hersh, W.R.: Evaluation of patient-level retrieval from electronic health record data for a cohort discovery task. JAMIA open **3**(3), 395–404 (2020)
22. Chan, A., Salganik, R., Markelius, A., Pang, C., Rajkumar, N., Krasheninnikov, D., Langosco, L., He, Z., Duan, Y., Carroll, M., Lin, M., Mayhew, A., Collins, K., Molamohammadi, M., Burden, J., Zhao, W., Rismani, S., Voudouris, K., Bhatt, U., Weller, A., Krueger, D., Maharaj, T.: Harms from increasingly agentic algorithmic systems. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23, pp. 651–666. Association for Computing Machinery, New York (2023)
23. Chan, A., Ezell, C., Kaufmann, M., Wei, K., Hammond, L., Bradley, H., Bluemke, E., Rajkumar, N., Krueger, D., Kolt, N., et al.: Visibility into AI agents. In: The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT), pp. 958–973 (2024)
24. Chen, A., Chen, D.O.: Accuracy of chatbots in citing journal articles. JAMA Netw. Open **6**(8), e2327647 (2023). ISSN 2574-3805. https://doi.org/10.1001/jamanetworkopen.2023.27647
25. Chen, B., Wu, Z., Zhao, R.: From fiction to fact: the growing role of generative ai in business and finance. J. Chinese Econ. Bus. Stud. **21**(4), 471–496 (2023). https://doi.org/10.1080/14765284.2023.2245279
26. Cheong, I., Xia, K., Feng, K.K., Chen, Q.Z., Zhang, A.X.: I Am Not a Lawyer, But...: Engaging legal experts towards responsible LLM policies for legal advice. In: The 2024 ACM Conference on Fairness, Accountability, and Transparency, pp. 2454–2469 (2024)

27. Choi, S., Gazeley, W., Wong, S.H., Li, T.: Conversational financial information retrieval model (ConFIRM) (2023). arXiv preprint arXiv:2310.13001
28. Collins-Thompson, K., Callan, J.P.: A language modeling approach to predicting reading difficulty. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004, pp. 193–200 (2004)
29. Collins-Thompson, K., Callan, J.: Predicting reading difficulty with statistical language models. J. Am. Soc. Inform. Sci. Technol. **56**(13), 1448–1462 (2005)
30. Corrada-Emmanuel, A., Croft, W.B., Murdock, V.: Answer passage retrieval for question answering. Tech. Reports of CIIR UMass, 2003
31. Cui, H., Sun, R., Li, K., Kan, M.Y., Chua, T.S.: Question answering passage retrieval using dependency relations. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 400–407 (2005)
32. D'Agostini Bueno, T.C., von Wangenheim, C.G., da Silva Mattos, E., Hoeschl, H.C., Barcia, R.M.: Jurisconsulto: retrieval in jurisprudencial text bases using juridical terminology. In: Proceedings of the 7th International Conference on Artificial Intelligence and Law, ICAIL '99, pp. 147–155. Association for Computing Machinery, New York (1999)
33. Dan, Y., Lei, Z., Gu, Y., Li, Y., Yin, J., Lin, J., Ye, L., Tie, Z., Zhou, Y., Wang, Y., et al.: Educhat: a large-scale language model-based chatbot system for intelligent education (2023). arXiv preprint arXiv:2308.02773
34. Davis, D., Kizilcec, R.F., Hauff, C., Houben, G.J.: The half-life of MOOC knowledge: a randomized trial evaluating knowledge retention and retrieval practice in MOOCs. In: Proceedings of the 8th International Conference on Learning Analytics and Knowledge (LAK '18) (2018). https://doi.org/10.1145/3170358.3170383
35. Deng, X., Gu, Y., Zheng, B., Chen, S., Stevens, S., Wang, B., Sun, H., Su, Y.: Mind2web: towards a generalist agent for the web. In: Advances in Neural Information Processing Systems, vol. 36, pp. 28091–28114. Curran Associates (2023)
36. Dewi, C., Chen, R.C., Liu, Y.T., Tai, S.K.: Synthetic data generation using DCGAN for improved traffic sign recognition. Neural Comput. Appl. **34**(24), 21465–21480 (2022)
37. Dobbins, N.J., Han, B., Zhou, W., Lan, K.F., Kim, H.N., Harrington, R., Uzuner, z., Yetisgen, M.: LeafAI: query generator for clinical cohort discovery rivaling a human programmer. J. Am. Med. Inform. Assoc. **30**(12), 1954–1964 (2023). ISSN 1527-974X. https://doi.org/10.1093/jamia/ocad149
38. Druin, A., Foss, E., Hatley, L., Golub, E., Guha, M.L., Fails, J., Hutchinson, H.: How children search the internet with keyword interfaces. In: Proceedings of the 8th International Conference on Interaction Design and Children, pp. 89–96 (2009)
39. Du, X., Shao, J., Cardie, C.: Learning to ask: neural question generation for reading comprehension (2017). arXiv preprint arXiv:1705.00106
40. Fan, M., Cheng, D., Yang, F., Luo, S., Luo, Y., Qian, W., Zhou, A.: Fusing global domain information and local semantic information to classify financial documents. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, CIKM '20, pp. 2413–2420. Association for Computing Machinery, New York (2020)
41. Figueira, A., Vaz, B.: Survey on synthetic data generation, evaluation methods and gans. Mathematics **10**(15), 2733 (2022)
42. Fuchs, K.: Exploring the opportunities and challenges of NLP models in higher education: is ChatGPT a blessing or a curse? In: Frontiers in Education, vol. 8, pp. 1166682. Frontiers (2023)
43. Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K., et al.: Red teaming language models to reduce harms: methods, scaling behaviors, and lessons learned (2022). arXiv preprint arXiv:2209.07858
44. Gao, C.A., Howard, F.M., Markov, N.S., Dyer, E.C., Ramesh, S., Luo, Y., Pearson, A.T.: Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. NPJ Digit. Med. **6**(1), 75 (2023). ISSN 2398-6352. https://doi.org/10.1038/s41746-023-00819-6

45. Gao, S., Dwivedi-Yu, J., Yu, P., Tan, X.E., Pasunuru, R., Golovneva, O., Sinha, K., Celikyilmaz, A., Bosselut, A., Wang, T.: Efficient tool use with chain-of-abstraction reasoning (2024). arXiv preprint arXiv:2401.17464
46. Ge, Y., Hua, W., Mei, K., jianchao ji, Tan, J., Xu, S., Li, Z., Zhang, Y.: OpenAGI: when LLM meets domain experts. In: Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2023). https://openreview.net/forum?id=gFf0a0ZxJM
47. Gebreab, S.A., Salah, K., Jayaraman, R., ur Rehman, M.H., Ellaham, S.: LLM-based framework for administrative task automation in healthcare. In: 2024 12th International Symposium on Digital Forensics and Security (ISDFS), pp. 1–7. IEEE (2024)
48. Goodman, R.S., Patrinely, J.R., Stone, C.A., Zimmerman, E., Donald, R.R., Chang, S.S., Berkowitz, S.T., Finn, A.P., Jahangir, E., Scoville, E.A., Reese, T.S., Friedman, D.L., Bastarache, J.A., van der Heijden, Y.F., Wright, J.J., Ye, F., Carter, N., Alexander, M.R., Choe, J.H., Chastain, C.A., Zic, J.A., Horst, S.N., Turker, I., Agarwal, R., Osmundson, E., Idrees, K., Kiernan, C.M., Padmanabhan, C., Bailey, C.E., Schlegel, C.E., Chambless, L.B., Gibson, M.K., Osterman, T.J., Wheless, L.E., Johnson, D.B.: Accuracy and reliability of chatbot responses to physician questions. JAMA Netw. Open **6**(10), e2336483 (2023). ISSN 2574-3805. https://doi.org/10.1001/jamanetworkopen.2023.36483
49. Gossen, T., Höbel, J., Nürnberger, A.: A comparative study about children's and adults' perception of targeted web search engines. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1821–1824 (2014)
50. Gou, Z., Shao, Z., Gong, Y., yelong shen, Yang, Y., Duan, N., Chen, W.: CRITIC: large language models can self-correct with tool-interactive critiquing. In: The Twelfth International Conference on Learning Representations (2024). https://openreview.net/forum?id=Sx038qxjek
51. Guha, N., Nyarko, J., Ho, D., Ré, C., Chilton, A., Chohlas-Wood, A., Peters, A., Waldon, B., Rockmore, D., Zambrano, D., et al.: LegalBench: a collaboratively built benchmark for measuring legal reasoning in large language models. In: Advances in Neural Information Processing Systems, vol. 36 (2024)
52. Gusenbauer, M.: Audit ai search tools now, before they skew research. Nature **617**(7961), 439–439 (2023)
53. Hämäläinen, P., Tavast, M., Kunnari, A.: Evaluating large language models in generating synthetic HCI research data: a case study. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, pp. 1–19 (2023)
54. Hamilton, S.: Blind judgement: agent-based supreme court modelling with gpt (2023). arXiv preprint arXiv:2301.05327
55. Heidt, A.: Artificial-intelligence search engines wrangle academic literature. Nature **620**(7973), 456–457 (2023)
56. Heilman, M.: Automatic factual question generation from text. Ph.D. Thesis, Carnegie Mellon University, 2011
57. Heilman, M., Collins-Thompson, K., Eskenazi, M.: An analysis of statistical models and features for reading difficulty prediction. In: Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications, pp. 71–79 (2008)
58. Hernandez, M., Epelde, G., Alberdi, A., Cilla, R., Rankin, D.: Synthetic data generation for tabular health records: a systematic review. Neurocomputing **493**, 28–45 (2022)
59. Hersh, W.: Search still matters: information retrieval in the era of generative AI. J. Am. Med. Inform. Assoc. ocae014 (2024). ISSN 1527-974X. https://doi.org/10.1093/jamia/ocae014
60. Hong, S., Zheng, X., Chen, J., Cheng, Y., Wang, J., Zhang, C., Wang, Z., Yau, S.K.S., Lin, Z., Zhou, L., et al.: MetaGPT: meta programming for multi-agent collaborative framework (2023). arXiv preprint arXiv:2308.00352
61. Hopkins, A.M., Logan, J.M., Kichenadasse, G., Sorich, M.J.: Artificial intelligence chatbots will revolutionize how cancer patients access information: ChatGPT represents a paradigm-shift. JNCI Cancer Spectr. **7**(2), pkad010 (2023). ISSN 2515-5091. https://doi.org/10.1093/jncics/pkad010

62. Hu, Z., Lee, R.K.W., Aggarwal, C.C., Zhang, A.: Text style transfer: A review and experimental evaluation. ACM SIGKDD Explorations Newslett. **24**(1), 14–45 (2022)
63. Huang, X., Liu, W., Chen, X., Wang, X., Wang, H., Lian, D., Wang, Y., Tang, R., Chen, E.: Understanding the planning of LLM agents: a survey (2024). arXiv preprint arXiv:2402.02716
64. Huibers, T., Landoni, M., Murgia, E., Pera, M.S.: IR for children 2000–2020: where are we now? In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2689–2692 (2021)
65. Huo, S., Arabzadeh, N., Clarke, C.L.A.: Retrieving supporting evidence for generative question answering. In: Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region, pp. 11–20 (2023). arXiv:2309.11392 [cs]
66. Jansen, B.J., Spink, A., Saracevic, T.: Real life, real users, and real needs: a study and analysis of user queries on the web. Inform. Process. Manag. **36**(2), 207–227 (2000)
67. Jiang, S., Shen, Z., Agrawal, M., Lam, B., Kurtzman, N., Horng, S., Karger, D., Sontag, D.: Conceptualizing machine learning for dynamic information retrieval of electronic health record notes. In: Machine Learning for Healthcare (2023). https://www.mlforhc.org/s/ID147_Research-Paper_2023.pdf
68. Jin, D., Jin, Z., Hu, Z., Vechtomova, O., Mihalcea, R.: Deep learning for text style transfer: a survey. Comput. Linguist. **48**(1), 155–205 (2022)
69. Jin, Q., Wang, Z., Floudas, C.S., Sun, J., Lu, Z.: Matching Patients to Clinical Trials with Large Language Models (2023). http://arxiv.org/abs/2307.15051. arXiv:2307.15051 [cs] version: 2
70. Jurczyk, P., Agichtein, E.: Discovering authorities in question answer communities by using link analysis. In: Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, pp. 919–922 (2007)
71. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. J. ACM **46**(5), 604–632 (1999)
72. Koa, K.J., Ma, Y., Ng, R., Chua, T.S.: Learning to generate explainable stock predictions using self-reflective large language models. In: Proceedings of the ACM on Web Conference 2024, WWW '24. ACM, New York (2024). https://doi.org/10.1145/3589334.3645611
73. Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., Bielaniewicz, J., Gruza, M., Janz, A., Kanclerz, K., et al.: ChatGPT: Jack of all trades, master of none. Inform. Fusion **99**, 101861 (2023)
74. Koenecke, A., Varian, H.: Synthetic data generation for economists (2020). arXiv preprint arXiv:2011.01374
75. Koopman, B., Zuccon, G.: Dr ChatGPT tell me what I want to hear: how different prompts impact health answer correctness. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 15012–15022 (2023)
76. Kusa, W., Mendoza, s.E., Knoth, P., Pasi, G., Hanbury, A.: Effective matching of patients to clinical trials using entity extraction and neural re-ranking. J. Biomed. Inform. **144** (2023)
77. Lallé, S., Conati, C.: A data-driven student model to provide adaptive support during video watching across MOOCs. In: Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part I 21, pp. 282–295. Springer, Berlin (2020)
78. Lalmas, M.: Aggregated search. In: Advanced Topics in Information Retrieval, pp., 109–123. Springer, Berlin (2011)
79. Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D., Allan, J.: Language models for financial news recommendation. In: Proceedings of the Ninth International Conference on Information and Knowledge Management, CIKM '00, pp. 389–396. Association for Computing Machinery, New York (2000)
80. Li, Q., Fu, L., Zhang, W., Chen, X., Yu, J., Xia, W., Zhang, W., Tang, R., Yu, Y.: Adapting large language models for education: foundational capabilities, potentials, and challenges (2023). arXiv preprint arXiv:2401.08664

81. Li, X., Cao2, Y., Pan, L., Ma, Y., Sun, A.: Towards Verifiable Generation: A Benchmark for Knowledge-aware Language Model Attribution (2023). http://arxiv.org/abs/2310.05634
82. Li, Y., Yu, Y., Li, H., Chen, Z., Khashanah, K.: TradingGPT: multi-agent system with layered memory and distinct characters for enhanced financial trading performance (2023). arXiv preprint arXiv:2309.03736
83. Lin, J.: An exploration of the principles underlying redundancy-based factoid question answering. ACM Trans. Inform. Syst. **25**(2), 6–es (2007)
84. Lingnau, A., Ruthven, I., Landoni, M., van der Sluis, F.: Interactive search interfaces for young children-the PuppyIR approach. In: 2010 10th IEEE International Conference on Advanced Learning Technologies, pp. 389–390. IEEE (2010)
85. Liu, Y.W., Liu, L.C., Wang, C.J., Tsai, M.F.: Fin10k: a web-based information system for financial report analysis and visualization. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, CIKM '16, pp. 2441–2444 (2016)
86. Liu, Z., Zhong, A., Li, Y., Yang, L., Ju, C., Wu, Z., Ma, C., Shu, P., Chen, C., Kim, S., et al.: Tailoring large language models to radiology: A preliminary approach to llm adaptation for a highly specialized domain. In: International Workshop on Machine Learning in Medical Imaging, pp. 464–473. Springer, Berlin (2023)
87. Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., Gu, Y., Ding, H., Men, K., Yang, K., Zhang, S., Deng, X., Zeng, A., Du, Z., Zhang, C., Shen, S., Zhang, T., Su, Y., Sun, H., Huang, M., Dong, Y., Tang, J.: AgentBench: evaluating LLMs as agents. In: The Twelfth International Conference on Learning Representations (2024). https://openreview.net/forum?id=zAdUB0aCTQ
88. Lock, S.: What is ai chatbot phenomenon ChatGPT and could it replace humans (2022). https://www.theguardian.com/technology/2022/dec/05/what-is-ai-chatbot-phenomenon-chatgpt-and-could-it-replace-humans
89. Locke, J.: An essay concerning humane understanding, 1690. https://www.gutenberg.org/files/10615/10615-h/10615-h.htm#chap2.10
90. Lopez-Lira, A., Tang, Y.: Can ChatGPT forecast stock price movements? Return predictability and large language models (2023). https://doi.org/10.1080/14765284.2023.2245279
91. Macey-Dare, R.: ChatGPT & generative ai systems as quasi-expert legal advice lawyers-case study considering potential appeal against conviction of tom hayes (2023). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4342686
92. Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., Gupta, S., Majumder, B.P., Hermann, K., Welleck, S., Yazdanbakhsh, A., Clark, P.: Self-refine: iterative refinement with self-feedback. In: Thirty-seventh Conference on Neural Information Processing Systems (2023). https://openreview.net/forum?id=S37hOerQLB
93. Masterman, T., Besen, S., Sawtell, M., Chao, A.: The landscape of emerging AI agent architectures for reasoning, planning, and tool calling: a survey (2024). arXiv preprint arXiv:2404.11584
94. Metzler, D., Tay, Y., Bahri, D., Najork, M.: Rethinking search: making domain experts out of dilettantes. In: ACM SIGIR Forum, vol. 55, pp. 1–27. ACM New York (2021)
95. Mialon, G., Fourrier, C., Wolf, T., LeCun, Y., Scialom, T.: GAIA: a benchmark for general AI assistants. In: The Twelfth International Conference on Learning Representations (2024). https://openreview.net/forum?id=fibxvahvs3
96. Monigatti, L.: Retrieval-Augmented Generation (RAG): From Theory to LangChain Implementation (2023). https://towardsdatascience.com/retrieval-augmented-generation-rag-from-theory-to-langchain-implementation-4e9bd5f6a4f2
97. Moraes, F., Putra, S.R., Hauff, C.: Contrasting search as a learning activity with instructor-designed learning. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, pp. 167–176 (2018)
98. Mouton, C.A., Lucas, C., Guest, E.: The operational risks of ai in large-scale biological attacks. RAND Corporation (2024). https://www.rand.org/pubs/research_reports/RRA2977-2.html

99. Muhlenbach, F., Sayn, I.: Artificial intelligence and law: what do people really want? Example of a French multidisciplinary working group. In: Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law, ICAIL '19, pp. 224–228. Association for Computing Machinery, New York (2019)
100. Murdock, V., Croft, W.B.: A translation model for sentence retrieval. In: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, pp. 684–691 (2005)
101. Murdock, V., Lalmas, M.: Workshop on aggregated search. In: ACM SIGIR Forum, vol. 42, pp. 80–83. ACM, New York (2008)
102. Neelakanteswara, A., Chaudhari, S., Zamani, H.: RAGs to style: Personalizing LLMs with style embeddings. In: Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024), pp. 119–123. Association for Computational Linguistics (2024)
103. Nievas, M., Basu, A., Wang, Y., Singh, H.: Distilling large language models for matching patients to clinical trials. J. Am. Med. Inform. Assoc. (2024)
104. Niklaus, J., Zheng, L., McCarthy, A.D., Hahn, C., Rosen, B.M., Henderson, P., Ho, D.E., Honke, G., Liang, P., Manning, C.: FLawN-T5: an empirical examination of effective instruction-tuning data mixtures for legal reasoning (2024). arXiv preprint arXiv:2404.02127
105. Nori, H., Lee, Y.T., Zhang, S., Carignan, D., Edgar, R., Fusi, N., King, N., Larson, J., Li, Y., Liu, W., et al.: Can generalist foundation models outcompete special-purpose tuning? Case study in medicine (2023). arXiv preprint arXiv:2311.16452
106. OpenAI: Introducing chatgpt (2022). https://openai.com/blog/chatgpt
107. Paşca, M.: Organizing and searching the world wide web of facts–step two: harnessing the wisdom of the crowds. In: Proceedings of the 16th International Conference on World Wide Web, pp. 101–110 (2007)
108. Paşca, M., Lin, D., Bigham, J., Lifchits, A., Jain, A.: Organizing and searching the world wide web of facts-step one: the one-million fact extraction challenge. In: Proceedings of AAAI, vol. 6, pp. 1400–1405 (2006)
109. Pan, H., Zhai, Z., Yuan, H., Lv, Y., Fu, R., Liu, M., Wang, Z., Qin, B.: KwaiAgents: generalized information-seeking agent system with large language models (2023). arXiv preprint arXiv:2312.04889
110. Patil, S.G., Zhang, T., Wang, X., Gonzalez, J.E.: Gorilla: large language model connected with massive APIs (2023). arXiv preprint arXiv:2305.15334
111. Pettinato Oltz, T.: ChatGPT, professor of law, 2023. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4347630
112. Plachouras, V., Smiley, C., Bretz, H., Taylor, O., Leidner, J.L., Song, D., Schilder, F.: Interacting with financial data using natural language. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '16, pp. 1121–1124 (2016)
113. Rabelo, J., Kim, M.Y., Goebel, R., Yoshioka, M., Kano, Y., Satoh, K.: A summary of the COLIEE 2019 competition. In: New Frontiers in Artificial Intelligence: JSAI-IsAI International Workshops, JURISIN, AI-Biz, LENLS, Kansei-AI, Yokohama, Japan, November 10–12, 2019, Revised Selected Papers, pp. 34–49. Springer, Berlin (2019)
114. Raheja, V., Kumar, D., Koo, R., Kang, D.: Coedit: text editing by task-specific instruction tuning (2023). arXiv preprint arXiv:2305.09857
115. Rashkin, H., Nikolaev, V., Lamm, M., Aroyo, L., Collins, M., Das, D., Petrov, S., Tomar, G.S., Turc, I., Reitter, D.: Measuring Attribution in Natural Language Generation Models (2022). http://arxiv.org/abs/2112.12870
116. Roberts, I., Gaizauskas, R.: Evaluating passage retrieval approaches for question answering. In: European Conference on Information Retrieval, pp. 72–84. Springer, Berlin (2004)
117. Ruan, Y., Dong, H., Wang, A., Pitis, S., Zhou, Y., Ba, J., Dubois, Y., Maddison, C.J., Hashimoto, T.: Identifying the risks of LM agents with an LM-emulated sandbox. In: The Twelfth International Conference on Learning Representations (2024). https://openreview.net/forum?id=GEcwtMk1uA

118. Savenkov, D., Agichtein, E.: When a knowledge base is not enough: Question answering over knowledge bases with external text data. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 235–244 (2016)
119. Sumithra, M.K., Sridhar, R.: Information retrieval in financial documents. In: Singh, P.K., Noor, A., Kolekar, M.H., Tanwar, S., Bhatnagar, R.K., Khanna, S. (eds.) Evolving Technologies for Computing, Communication and Smart World, pp. 265–274. Springer, Singapore (2021). ISBN 978-981-15-7804-5
120. Sun, Z.: A short survey of viewing large language models in legal aspect (2023). arXiv preprint arXiv:2303.09136
121. Syed, R., Collins-Thompson, K.: Optimizing search results for human learning goals. Inform. Retrieval J. **20**, 506–523 (2017)
122. Syed, R., Collins-Thompson, K., Bennett, P.N., Teng, M., Williams, S., Tay, D.W.W., Iqbal, S.: Improving learning outcomes with gaze tracking and automatic question generation. In: Proceedings of the Web Conference 2020, pp. 1693–1703 (2020)
123. Tan, T.F., Elangovan, K., Jin, L., Jie, Y., Yong, L., Lim, J., Poh, S., Ng, W.Y., Lim, D., Ke, Y., et al.: Fine-tuning large language model (llm) artificial intelligence chatbots in ophthalmology and llm-based evaluation using gpt-4 (2024). arXiv preprint arXiv:2402.10083
124. Tan, Y., Zhang, Z., Li, M., Pan, F., Duan, H., Huang, Z., Deng, H., Yu, Z., Yang, C., Shen, G., et al.: Medchatzh: A tuning LLM for traditional Chinese medicine consultations. Comput. Biol. Med. **172**, 108290 (2024)
125. Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K., Gutierrez, L., Tan, T.F., Ting, D.S.W.: Large language models in medicine. Nat. Med. **29**(8), 1930–1940 (2023)
126. Thurlow, C., Brown, A.: Generation txt? The sociolinguistics of young people's text-messaging. Discourse Anal. Online **1**(1), 30 (2003)
127. Tian, S., Jin, Q., Yeganova, L., Lai, P.T., Zhu, Q., Chen, X., Yang, Y., Chen, Q., Kim, W., Comeau, D.C., Islamaj, R., Kapoor, A., Gao, X., Lu, Z.: Opportunities and challenges for ChatGPT and large language models in biomedicine and health. Briefings Bioinform. **25**(1) (2023)
128. Tian, Y., Yang, X., Zhang, J., Dong, Y., Su, H.: Evil geniuses: delving into the safety of LLM-based agents (2024). https://arxiv.org/abs/2311.11855
129. Trautmann, D., Petrova, A., Schilder, F.: Legal prompt engineering for multilingual legal judgement prediction (2022). https://arxiv.org/abs/2212.02199
130. Trozze, A., Davies, T., Kleinberg, B.: Large language models in cryptocurrency securities cases: can a gpt model meaningfully assist lawyers? Artif. Intell. Law 1–47 (2024)
131. Tsai, M.-F., Wang, C.-J.: Risk ranking from financial reports. In Serdyukov, P., Braslavski, P., Kuznetsov, S.O., Kamps, J., Rüger, S., Agichtein, E., Segalovich, I., Yilmaz, E. (eds.) Advances in Information Retrieval, pp. 804–807. Springer, Berlin, Heidelberg (2013)
132. Tu, T., Palepu, A., Schaekermann, M., Saab, K., Freyberg, J., Tanno, R., Wang, A., Li, B., Amin, M., Tomasev, N., Azizi, S., Singhal, K., Cheng, Y., Hou, L., Webson, A., Kulkarni, K., Mahdavi, S.S., Semturs, C., Gottweis, J., Barral, J., Chou, K., Corrado, G.S., Matias, Y., Karthikesalingam, A., Natarajan, V.: Towards Conversational Diagnostic AI (2024). http://arxiv.org/abs/2401.05654
133. Unlu, O., Shin, J., Mailly, C.J., Oates, M.F., Tucci, M.R., Varugheese, M., Wagholikar, K., Wang, F., Scirica, B.M., Blood, A.J., Aronson, S.J.: Retrieval augmented generation enabled generative pre-trained transformer 4 (GPT-4) performance for clinical trial screening (2024). medRxiv: The Preprint Server for Health Sciences
134. Usta, A., Altingovde, I.S., Ozcan, R., Ulusoy, Ö.: Learning to rank for educational search engines. IEEE Trans. Learn. Technol. **14**(2), 211–225 (2021)
135. Van Bulck, L., Moons, P.: What if your patient switches from Dr. Google to Dr. ChatGPT? A vignette-based survey of the trustworthiness, value and danger of ChatGPT-generated responses to health questions. Eur. J. Cardiovasc. Nursing **23**(1), 95–98 (2024) ISSN 1873-1953. https://doi.org/10.1093/eurjcn/zvad038

136. Voorhees, E.M.: The TREC medical records track. In: Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics, BCB'13, pp. 239–246 (2013)
137. Voorhees, E.M., Hersh, W.: Overview of the TREC 2012 medical records track. In: The Twenty-First Text REtrieval Conference (TREC 2012) Proceedings (2012)
138. Walters, W.H., Wilder, E.I.: Fabrication and errors in the bibliographic citations generated by ChatGPT. Sci. Rep. **13**(1), 14045 (2023). ISSN 2045-2322. https://doi.org/10.1038/s41598-023-41032-5
139. Wang, S., Scells, H., Koopman, B., Zuccon, G.: Can ChatGPT Write a Good Boolean Query for Systematic Review Literature Search? (2023). http://arxiv.org/abs/2302.03495
140. Wang, S., Yuan, H., Ni, L.M., Guo, J.: QuantAgent: Seeking holy grail in trading by self-improving large language model (2024). arXiv preprint arXiv:2402.03755
141. Wartella, E.A., Jennings, N.: Children and computers: new technology. old concerns. The future of children 31–43 (2000)
142. West, J., Bhattacharya, M.: Intelligent financial fraud detection: a comprehensive review. Comput. Secur. **57**, 47–66 (2016)
143. White, R.W., Nouri, E., Woffinden-Luey, J., EncarnacióN, M., Jauhar, S.K.: Microtask detection. ACM Trans. Inf. Syst. **39**(2) (2021). ISSN 1046-8188. https://doi.org/10.1145/3432290
144. Wolfe, J.H.: Automatic question generation from text-an aid to independent study. In: Proceedings of the ACM SIGCSE-SIGCUE Technical Symposium on Computer Science and Education, pp. 104–112 (1976)
145. Wolfe, C.R.: A Practitioners Guide to Retrieval Augmented Generation (RAG) (2024). https://cameronrwolfe.substack.com/p/a-practitioners-guide-to-retrieval
146. Wornow, M., Lozano, A., Dash, D., Jindal, J., Mahaffey, K.W., Shah, N.H.: Zero-Shot Clinical Trial Patient Matching with LLMs (2024). http://arxiv.org/abs/2402.05125
147. Wu, S., Liu, S., Wang, Y., Timmons, T., Uppili, H., Bedrick, S., Hersh, W., Liu, H.: Intrainstitutional EHR collections for patient-level information retrieval. J. Assoc. Inform. Sci. Technol. **68**(11), 2636–2648 (2017)
148. Wu, Q., Bansal, G., Zhang, J., Wu, Y., Zhang, S., Zhu, E., Li, B., Jiang, L., Zhang, X., Wang, C.: AutoGen: enabling next-gen LLM applications via multi-agent conversation framework (2023a). arXiv preprint arXiv:2308.08155
149. Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., Mann, G.: BloombergGPT: A large language model for finance (2023). arXiv preprint arXiv:2303.17564
150. Wu, K., Wu, E., Cassasola, A., Zhang, A., Wei, K., Nguyen, T., Riantawan, S., Riantawan, P.S., Ho, D.E., Zou, J.: How well do LLMs cite relevant medical references? An evaluation framework and analyses (2024). arXiv preprint arXiv:2402.02008
151. Wu, Y., Zhou, S., Liu, Y., Lu, W., Liu, X., Zhang, Y., Sun, C., Wu, F., Kuang, K.: Precedent-enhanced legal judgment prediction with LLM and domain-model collaboration. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 12060–12075. Association for Computational Linguistics (2023)
152. Wylde, V., Rawindaran, N., Lawrence, J., Balasubramanian, R., Prakash, E., Jayal, A., Khan, I., Hewage, C., Platts, J.: Cybersecurity, data privacy and blockchain: a review. SN Comput. Sci. **3**(2), 127 (2022)
153. Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., et al.: The rise and potential of large language model based agents: a survey (2023). arXiv preprint arXiv:2309.07864
154. Yang, W., Lu, K., Yang, P., Lin, J.: Critically examining the "neural hype" : weak baselines and the additivity of effectiveness gains from neural ranking models. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '19. ACM, New York (2019)
155. Yang, W., Bi, X., Lin, Y., Chen, S., Zhou, J., Sun, X.: Watch out for your agents! Investigating backdoor threats to LLM-based agents (2024). arXiv preprint arXiv:2402.11208

156. Yao, L., Lin, Y., Mo, Y., Wang, F.: Performance evaluation of financial industry related expense forecasting using various regression algorithms for machine learning. Highlights Sci. Eng. Technol. **57**, 235–241 (2023)
157. Yu, H.: Reflection on whether chat gpt should be banned by academia from the perspective of education and teaching. Front. Psychol. **14**, 1181712 (2023)
158. Yu, F., Quartey, L., Schilder, F.: Legal prompting: teaching a language model to think like a lawyer (2022). arXiv preprint arXiv:2212.01326
159. Zakka, C., Shad, R., Chaurasia, A., Dalal, A.R., Kim, J.L., Moor, M., Fong, R., Phillips, C., Alexander, K., Ashley, E., Boyd, J., Boyd, K., Hirsch, K., Langlotz, C., Lee, R., Melia, J., Nelson, J., Sallam, K., Tullis, S., Vogelsong, M.A., Cunningham, J.P., Hiesinger, W.: Almanac—retrieval-augmented language models for clinical medicine. NEJM AI **1**(2), AIoa2300068 (2024)
160. Zhang, Y., Jauhar, S.K., Kiseleva, J., White, R., Roth, D.: Learning to decompose and organize complex tasks. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 2726–2735 (2021)
161. Zhang, X., Xu, H., Ba, Z., Wang, Z., Hong, Y., Liu, J., Qin, Z., Ren, K.: Privacyasst: safeguarding user privacy in tool-using large language model agents. IEEE Trans. Depend. Secure Comput. (2024)
162. Zhang, Z., Bo, X., Ma, C., Li, R., Chen, X., Dai, Q., Zhu, J., Dong, Z., Wen, J.R.: A survey on the memory mechanism of large language model based agents (2024). https://arxiv.org/abs/2404.13501
163. Zhu, Y., Yuan, H., Wang, S., Liu, J., Liu, W., Deng, C., Dou, Z., Wen, J.R.: Large language models for information retrieval: a survey (2023). arXiv preprint arXiv:2308.07107
164. Zhu, P., Câmara, A., Roy, N., Maxwell, D., Hauff, C.: On the effects of automatically generated adjunct questions for search as learning. In: Proceedings of the 2024 Conference on Human Information Interaction and Retrieval, pp. 266–277 (2024)
165. Zong, C., Yan, Y., Lu, W., Shao, J., Huang, E., Chang, H., Zhuang, Y.: Triad: a framework leveraging a multi-role LLM-based agent to solve knowledge base question answering (2024). https://arxiv.org/abs/2402.14320
166. Zuccon, G., Koopman, B., Shaik, R.: ChatGPT hallucinates when attributing answers. In: Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region, pp. 46–51 (2023)