AI Open 4 (2023) 80–90

Contents lists available at ScienceDirect

AI Open

journal homepage: www.keaipublishing.com/en/journals/ai-open

Information Retrieval meets Large Language Models: A strategic report from Chinese IR community

Qingyao Ai [a], Ting Bai [b], Zhao Cao [c], Yi Chang [d], Jiawei Chen [e,*], Zhumin Chen [f], Zhiyong Cheng [g], Shoubin Dong [h], Zhicheng Dou [i], Fuli Feng [j], Shen Gao [f], Jiafeng Guo [k], Xiangnan He [j,*], Yanyan Lan [a], Chenliang Li [l], Yiqun Liu [a], Ziyu Lyu [m], Weizhi Ma [a], Jun Ma [f], Zhaochun Ren [f], Pengjie Ren [f], Zhiqiang Wang [n], Mingwen Wang [o], Ji-Rong Wen [i], Le Wu [p], Xin Xin [f], Jun Xu [i], Dawei Yin [q], Peng Zhang [r,*], Fan Zhang [l], Weinan Zhang [s], Min Zhang [a], Xiaofei Zhu [t]

[a] *Tsinghua University, China*
[b] *Beijing University of Posts and Telecommunications, China*
[c] *Huawei Technologies Ltd. Co, China*
[d] *Jilin University, China*
[e] *Zhejiang University, China*
[f] *Shandong University, China*
[g] *Shandong Artificial Intelligence Institute, China*
[h] *South China University of Technology, China*
[i] *Renmin University of China, China*
[j] *University of Science and Technology of China, China*
[k] *Institute of Computing Technology, Chinese Academy of Sciences, China*
[l] *Wuhan University, China*
[m] *Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, China*
[n] *Shanxi University, China*
[o] *Jiangxi Normal University, China*
[p] *Hefei University of Technology, China*
[q] *Baidu Inc, China*
[r] *Tianjin University, China*
[s] *Shanghai Jiao Tong University, China*
[t] *Chongqing University of Technology, China*

ARTICLE INFO

Keywords:
Information Retrieval
Language Models
Recommendation system

ABSTRACT

The research field of Information Retrieval (IR) has evolved significantly, expanding beyond traditional search to meet diverse user information needs. Recently, Large Language Models (LLMs) have demonstrated exceptional capabilities in text understanding, generation, and knowledge inference, opening up exciting avenues for IR research. LLMs not only facilitate generative retrieval but also offer improved solutions for user understanding, model evaluation, and user-system interactions. More importantly, the synergistic relationship among IR models, LLMs, and humans forms a new technical paradigm that is more powerful for information seeking. IR models provide real-time and relevant information, LLMs contribute internal knowledge, and humans play a central role of demanders and evaluators to the reliability of information services. Nevertheless, significant challenges exist, including computational costs, credibility concerns, domain-specific limitations, and ethical considerations. To thoroughly discuss the transformative impact of LLMs on IR research, the Chinese IR community conducted a strategic workshop in April 2023, yielding valuable insights. This paper provides a summary of the workshop's outcomes, including the rethinking of IR's core values, the mutual enhancement of LLMs and IR, the proposal of a novel IR technical paradigm, and open challenges.

* Corresponding authors.
E-mail addresses: sleepyhunt@zju.edu.cn (J. Chen), hexn@ustc.edu.cn (X. He), pzhang@tju.edu.cn (P. Zhang).

https://doi.org/10.1016/j.aiopen.2023.08.001
Received 13 July 2023; Accepted 1 August 2023
Available online 7 August 2023

1. Introduction

In the past few decades, Information Retrieval (IR) has experienced significant growth and development in both industry and academia. In early stage, IR research mainly focused on search, which aimed to assist users in finding relevant information (Kobayashi and Takeda, 2000a). In recent years, the scope of IR research has expanded to encompass a wide range of online applications and scenarios (Manoj and Elizabeth, 2008). This diversification is evident in the rise of recommendation systems as a prominent research area, as observed in flagship conferences ACM SIGIR from 2018 to 2023. Additionally, exciting research directions such as conversational systems, user modeling, and knowledge extraction, have also emerged (Faggioli et al., 2021; Li et al., 2022). These advancements reflect an evolution in the core value of IR, moving beyond merely retrieving relevant documents to meeting the information needs of users (Chen et al., 2020).

Large Language Models (LLMs) have demonstrated remarkable capabilities in various aspects of text understanding, generation, knowledge inference, and compositional generalization (Bubeck et al., 2023; Liu et al., 2023a). As a result, LLMs have the potential to open up new research directions in the field of IR. These include, but are not limited to:

- Enabling IR systems to generate content directly that satisfies user information needs, known as generative retrieval (Lee et al., 2022b) and generative recommendation (Wang et al., 2023a). An example of this is the New Bing,[1] which allows for content generation to meet user queries.
- Enhancing the understanding of user intents and behaviors by incorporating rich contextual information into IR system.
- Creating opportunities for the development of superior indexing systems that can handle dynamic, semantically-aware, and multi-modal data.
- Providing better methods for evaluating model accuracy and interpretability in IR tasks.
- Facilitating enhanced interactive experiences between users and IR systems. These advancements highlight the potential of LLMs to revolutionize various aspects of IR research and practice.

Reciprocally, IR technology plays a crucial role in supporting the development of LLMs (Zheng and Fischer, 2023; Jeronymo et al., 2023). The interaction among humans, IR models, and LLMs forms a new triangular IR paradigm, as depicted in Fig. 1. In this paradigm, the IR model acts as the means to acquire external knowledge, offering real-time, up-to-date, and relevant information to both LLMs and humans. LLMs contribute precise internal knowledge and information, leveraging their powerful reasoning abilities to provide high-quality responses. Humans, in their role as demanders and evaluators, play a central role in the retrieval process. It is important to emphasize that LLMs, without the inclusion of IR models, possess limitations in terms of short-term memory and reasoning. These limitations could result in the lack of effective integration of new and existing knowledge, hindering LLMs' ability to provide the information that humans need. The IR model, functioning as a "knowledge base" equipped with matching and retrieval capabilities, addresses these limitations by compensating for the lack of factual consistency and long-term memory in LLMs. This combination of "knowledge+reasoning" establishes a dual-wheel drive, enabling a more intelligent and reliable information service system for humans (Kim et al., 2022; Zheng and Fischer, 2023).

While LLMs for IR hold promise, they also face significant challenges that cannot be ignored:

- High computational costs: LLMs for IR require substantial computational resources, which can limit their use and research by small and medium-sized companies and institutions. Furthermore, deploying models on terminal devices and ensuring real-time performance can be challenging (Edalati et al., 2021).
- Credibility concerns: The credibility of LLMs for IR is relatively low, as they have been found to provide misleading explanations, incorrect answers, and unreliable information sources. This can undermine people's trust in LLMs (Liu et al., 2023b).
- Performance limitations in specific fields: LLMs may struggle to perform well in specific domains due to the lack of high-quality datasets, differences in data representation, and other related issues. The limited domain-specific knowledge hampers the rapid implementation and application of LLMs (Nori et al., 2023; Liu et al., 2023c).
- Ethical and moral considerations: The emergence of generative IR systems raises the requirements for ethical and moral evaluation standards. Ensuring fairness, impartiality, and ethics in the generated content is a significant challenge. It is inherently difficult to guarantee that LLMs meet these regulatory and ethical requirements (Zhang et al., 2023a; Blair-Stanek et al., 2023).

Given the flourishing research in the IR field and the promising opportunities brought about by LLMs, it is a favorable moment to envision the future advancements in IR. To this end, the Chinese IR community organized a strategic workshop on April 14th and 15th, 2023, aiming to explore future opportunities and challenges. This paper presents the key findings and conclusions drawn from the workshop, including the rethinking of the core values of IR in Section 2, a discussion of how LLMs and IR can mutually enhance each other in Section 3 and Section 4, the new IR paradigm of fusing humans, LLMs, and IR in Section 5, and a discussion of open challenges and future directions in Section 6.

2. Rethinking core values of IR

Before delving into the effects of LLMs on IR research, it is crucial to rethink the IR discipline itself. This contemplation leads us to ponder the following questions: Firstly, what is the fundamental contribution of IR, as a scientific discipline, to other domains? Secondly, considering its extensive development over the past decades, what are the boundaries and extensions of IR? These questions are intended to promote comprehension of the fundamental principles of IR, and foster the integration of LLMs into IR while preserving its essence.

2.1. Scientific connotation of IR

Classical IR concerns the retrieval of the information that satisfies a user's information need from the target corpus. Therefore, user, ranking, and corpus are the fundamental concepts in the IR discipline. In the light of this, it is important to re-examine the scientific connotation of IR from the three aspects.

User perspective. Users are the core part of IR (Manning et al., 2008; Kobayashi and Takeda, 2000b), since an IR system is supposed to serve users by satisfying their information needs. Therefore, user understanding is a fundamental problem in IR, and the community has made decades-long progress in user understanding. The studies of user understanding in IR mainly concentrate on two sides, respectively user intent understanding and user behavior modeling.

- User intent understanding plays a central role in IR. Commercial search engines like Google[2] and Bing,[3] have evolved from

[1] https://www.bing.com/new

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

[3] https://www.bing.com

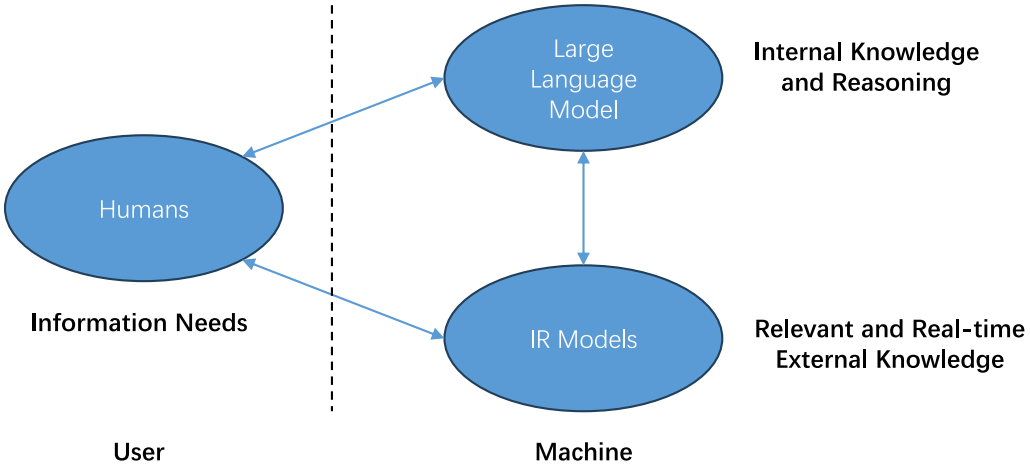

Fig. 1. Our proposed new IR technical paradigm introduces three key elements: Humans, IR models, and LLMs. The synergistic relationship among them not only facilitates mutual enhancement but also enables the fulfillment of new tasks as an organic whole.

keyword semantic matching to user intent optimization in recent decades. Understanding user intent is the first step for better user experiences, which has been broadly studied in various IR applications, e.g., web search (Jansen et al., 2007; Teevan et al., 2008), product search in e-commerce (Su et al., 2018), community question answering (Chen et al., 2012), etc.

- User behavior modeling is to comprehensively understand and model user behaviors when interacting with IR systems. IR community has studied diverse and multi-dimensional user behaviors including user interaction behaviors (Agichtein et al., 2006; Manavoglu et al., 2003), sequential behaviors (Pi et al., 2019; Yuan et al., 2020), mobility behaviors (Yuan et al., 2012; Zheng et al., 2010) and social behaviors (Jin et al., 2013). User behavior modeling helps to inject personalization into IR systems.

The techniques of user understanding in IR have provided valuable insights for other research domains. In natural language processing (NLP), user intent understanding is the preliminary step to make further actions in task-oriented dialogue systems (Alfieri et al., 2022). User profiling and personalization are also utilized to develop persona-based dialogue systems (Li et al., 2016; Zhong et al., 2020). In social computing, user modeling is an important area in the design of personalized incentive mechanisms for encouraging participation (Vassileva, 2012).

Ranking perspective. Ranking lies at the core part of IR research by providing an ordering of items towards user's information need (Guo et al., 2020; Trabelsi et al., 2021). The formats of items are various, including but not limited to textual documents, images, and videos. The problem is usually formulated as ranking a collection of items that match the given query or context, according to some criterion like relevance and recency. Most IR systems are supported by a ranking module.

Different techniques have been used for constructing ranking models, from traditional ranking models to machine learning methods. The traditional ranking models include vector space models and probabilistic models (e.g. BM25 (Robertson and Zaragoza, 2009)). The machine learning methods include the earliest learning to rank models (e.g. Rankboost (Freund et al., 2003) and LambdaMart (Burges, 2010)) and the recent neural ranking models such as Deep Structured Semantic Models (DSSM) (Huang et al., 2013) and DeepRank (Pang et al., 2017). Ranking models have been extensively exploited in many IR applications such as ad-hoc retrieval (Guo et al., 2016; Jiang et al., 2020; Fujiwara et al., 2013), recommender systems (Karatzoglou et al., 2013; Vargas and Castells, 2011), community question answering (Zhang et al., 2014; Yang et al., 2013), etc.

It is worth noting that ranking is not restricted to IR alone, as it also has applications in a diverse range of other research fields. For example, ChatGPT (Ouyang et al., 2022a) collects comparison samples and utilizes a ranking labeler to train the reward model when aligning with user intents. In computational biology, learning-to-rank algorithms have been exploited to rank the candidate 3-D structures in protein structure prediction (Duh and Kirchhoff, 2008).

Corpus perspective. For the corpus perspective, the connotation of IR lies in conducting effective search efficiently from a large volume of multi-modal information corpus (Kobayashi and Takeda, 2000a). Such connotation mainly lies in the following two folds:

- Effective corpus representation. Conventional IR systems represent the documents with bag-of-words or TF–IDF methods (Kobayashi and Takeda, 2000a). Deep learning methods (e.g., dense retrieval) empowers the IR system to represent multi-modal corpus such as images and videos with dense latent vectors, which has largely improved the retrieval results and broadened the application scope of IR (Karpukhin et al., 2020; Zhen et al., 2019). Recently, contrastive learning and self-supervised learning further improve the learning of corpus representation (Izacard et al., 2021).
- Efficient retrieval infrastructure. The infrastructure of IR involves fundamental facilities to construct an IR system, including but not limited to building dictionaries, indexing, scoring, distributed search, and caching (Manning et al., 2008). Such infrastructure determines how to organize the information source and thus largely affect the efficiency of IR (Manning et al., 2008). LLMs and generative IR (Tay et al., 2022) shed new lights to reform the retrieval infrastructure.

2.2. Boundaries and extensions of IR

Although the growth of Web and mobile technology has greatly advanced the development of IR, it also results in some limitations. A common misconception is that IR simply involves ranking a set of objects in cyberspace, which is an inaccurate and narrow view. As such, it is crucial to reconsider the boundaries and extensions of IR from various perspectives, including the input, process, output, and environment.

Ranking or generation? The primary focus of classical IR pertains to the ordering of items based on the information need, such as the query in search engines or the context of recent interactions in recommender systems. The advancement of large-scale generative models

such as GPT (Brown et al., 2020) and Diffusion Models (Rombach et al., 2022) has expanded the content space, allowing exploration beyond existing sets of content through generation. Integrating the generation channel for information seeking opens up new paradigms in IR systems, such as generative search and recommendation (Wang et al., 2023a). However, this extension presents challenges to building IR systems in infrastructure, algorithm, evaluation, *etc*. For instance, it calls for the infrastructure to accommodate generative models at a large scale and distribute the content generated on the fly in an efficient manner, especially for multi-media content such as micro-videos. Additionally, it calls for algorithm innovation on the alignment of generative models to IR tasks, the aggregation of ranking and generation results, and the joint optimization of ranking and generation models. Furthermore, evaluation techniques must also be extended to accommodate unlimited content and emphasize new perspectives such as fidelity checking.

For human or for machine? IR systems have been designed to assist *human* users to access information resources. However, with the increasing prevalence of AI models, particularly LLMs, the user base of IR systems could potentially be expanded to include intelligent *machines*. This has led to the emergence of a growing body of research referred to as IR-augmented methods or retrieval-enhanced machine learning (REML) (Zamani et al., 2022) in various communities such as computer vision (Gur et al., 2021), natural language processing (Borgeaud et al., 2021), and machine learning (Santoro et al., 2016). From a conceptual perspective, IR systems can be regarded as a memory and access system for external data, information, or knowledge, capable of serving as a generic supporting technique for either humans or machines. Nevertheless, the shift from human to machine poses a multitude of research challenges throughout the entire process of the IR system, including indexing, representation, retrieval, ranking, and feedback. Additionally, there are many research issues to be addressed regarding the learning, evaluation, and deployment of IR technology in conjunction with AI models. Despite these challenges, it is encouraging to anticipate that IR could become a ubiquitous component of the AI paradigm in the future, providing valuable support to society at large.

For finding or for decision? IR systems have long benefited people in finding desired information. For the extension of IR, the system should also support complicated, explainable, and long-term information seeking to help users with reasonable decision-making suggestions (Ford et al., 1999). To accomplish this goal, a retrieval system should be aware of the user context in which the information need is to be placed (Xu et al., 2023). For complex decision-making tasks, the system is expected to scaffold the task step-by-step (Borlund, 2013). The system should also provide transparent retrieval process and explainable results to provide reasonable decision-making supports (Yin et al., 2023). Besides, understanding how people make decisions is also important. Collaboration with other communities will likely be necessary to make progress in decision understanding. These communities may include human–computer interaction, behavioral economics, psychology, cognitive science, as well as specific domain communities like the clinical communities (Tsvetkov, 2015; Ingwersen, 1984). Finally, placing IR into LLMs to provide AI-Generated Actions (AIGA) is also an emerging topic for the extension of IR (Janner, 2023).

Going beyond cyberspace? IR systems have conventionally emphasized digital content within the realm of cyberspace. However, advancements in human–computer interaction (Esposito et al., 2021), unmanned vehicles (Mohsan et al., 2022), and robotics (Zhu et al., 2022) enable us to envision the retrieval of contents in the physical world. For instance, the extended IR system may explore information in the physical world such as the wind speed around and the source of noise. Further extensions may incorporate the search and delivery of physical items. These extensions build on the aforementioned ranking-generation hybrids, human–machine hybrids, and finding-decision hybrids. Moreover, new concepts around interaction interfaces, system architecture, and feedback mechanisms will arise. Towards these ends, we may open up more cross-domain research directions and entangle IR with various emerging techniques in other fields such as Metaverse (Mystakidis, 2022) and Embodied Artificial Intelligence (Shenavarmasouleh et al., 2022). As a result, there will be many research opportunities regarding the development, operation, maintenance, and regulation of such IR systems, increasing the value of the IR industry remarkably.

3. Large language models for IR

IR systems have long been divided into five major components: user modeling, indexing, matching/ranking, evaluation, and user interaction. By pre-training on large-scale corpora and fine-tuning to follow human instructions, LLMs have demonstrated superior capability in language understanding, generation, interaction, and knowledge reasoning (Ouyang et al., 2022b). Therefore, it is reasonable to anticipate that LLMs will considerably augment the major IR components from various perspectives. In this section, we discuss the possibilities of LLMs for each major IR component.

3.1. User modeling

User modeling aims to accurately represent users and their information needs by understanding their characteristics, preferences, and behaviors (John and Mooney, 2001; Hersh et al., 1994; Dennis et al., 2002; Guo et al., 2019). In view of the acknowledged capabilities of LLMs, we enumerate several potential directions that LLMs can enhance user modeling.

- **Language Understanding**. Most basically, LLMs can enhance the understanding of user queries, enabling more accurate analysis and leading to more relevant search results.
- **Behavior Understanding**. LLMs allow the analyses of user behaviors in the semantic level of item content, offering a comprehensive understanding of preferences and behaviors. By analyzing data like click-stream, search log, and interaction history, models can identify patterns and relationships, resulting in more accurate user models.
- **Personalization**. LLMs can build comprehensive user models, incorporating various characteristics and preferences. By analyzing social media activity, online behavior, and integrating with IoT devices, models can consider the factors like physical environment and emotional states, leading to more personalized recommendations.
- **Conversational Interfaces**. LLMs facilitate more natural and seamless interactions, generating context-aware responses. Conversational interfaces can incorporate sentiment analysis and emotional response generation, thereby enhancing personalized and engaging user experiences.
- **Hybrid Modeling**. Combining LLMs with rule-based or collaborative filtering models can yield stronger hybrid models. Such hybrid modeling unifies the strengths of different approaches, improving the personalization and accuracy of search and recommendation (Bao et al., 2023; Gao et al., 2023).

The advancement of LLMs is expected to significantly enhance the user modeling in IR systems. Nonetheless, certain challenges, such as data privacy, biases in data and models, and the need for extensive training data, need to be tackled. Addressing these challenges is imperative to fully exploit the potential benefits of LLMs for user modeling in IR, while mitigating any potential negative effects.

3.2. Indexing

The emergence of LLMs provides the basis for generative retrieval, a new retrieval paradigm where the indexing component is dramatically changed. Recently, dense retrieval has been extensively studied and the approaches based on the standard MIPS index and nearest neighbor search are common (Karpukhin et al., 2020). Given the success of Transformers as a good associative memory store or search index, Tay et al. (2022) propose a novel architecture called differentiable search index (DSI) where the index is stored in the model parameters. DSI uses LLMs to directly learn the mapping of queries to relevant document IDs. It enables the model's internal memory to act as an index, thus greatly simplifying the entire retrieval process. Inspired by the work of DSI, we identify some directions that LLMs will change the technology of indexing:

- **From Static to Dynamic.** Indexing systems based on LLMs need be dynamic, as all corpus information is encoded within the LLM parameters. Incremental index updates are a specific instance of model updates, as noted in the study by Sun et al. (2020).
- **From Keyword-based to Semantics-oriented.** Indexing systems based on LLMs should be semantic. Thanks to the powerful contextual modeling capabilities of LLMs, indexing systems based on LLMs are able to find the documents that are related to the query in a more nuanced way.
- **From Uni-modal to Multi-modal.** Indexing systems utilizing LLMs have the potential to be multi-modal. The development of multi-modal LLMs will facilitate the indexing systems capable of indexing various modalities of data in a unified manner, including but not limited to texts, images, and videos.

3.3. Matching/ranking

LLMs have demonstrated remarkable capability to understand and rank complex content, including both single-modal and multi-modal data. In this part, we focus on two topics: (1) If generative models can already provide exact answers, is ranking still necessary? (2) What are the future research directions in the ranking, and what problems remain to be solved?

While generative LLMs can provide coherent answers to user queries, users still need to know which document, image, or webpage is most relevant to their query, so as to verify the answer generated by the LLM. The ranking results from a retrieval system can improve the interpretability of LLM, and provide the trustworthy information to support the answer generation. The retrieval system typically acts as a plugin to LLMs (Feng et al., 2023; Qin et al., 2023), providing knowledge acquisition abilities.

In future research, LLMs offer many interesting directions to explore in ranking. Under the generative paradigm, we should prioritize ranking and improve the ranking of results to ensure a good user experience and high satisfaction. This is also a more essential goal of ranking in IR. When ranking search results, we should focus on the integrity of the returned results. Instead of simply returning a ranking list based on relevance, the model should return the information that is more integrated and relevant to users' real needs.

In terms of ranking evaluation, existing methods are designed for the form of returning a list of documents, which is no longer suitable under the generative paradigm. There are promising directions for future research in the field of ranking. An example is to leverage the LLMs as the human simulator to measure the user satisfaction and experience of a ranking method (Sun et al., 2022).

3.4. Evaluation

For validate the effectiveness of LLM-enhanced IR approaches, it is essential to develop proper metrics and datasets. Conventional IR metrics, such as Precision and Recall, Mean Reciprocal Rank (MRR, Craswell (2009)), Mean Average Precision (MAP, Zhu (2004)), Normalized discounted cumulative gain (nDCG, Järvelin and Kekäläinen (2002), Sun et al. (2023)), still play a critical role in IR model evaluation. However, with the wide application of LLMs in IR, tailored evaluation strategies become essential to justify the effectiveness of LLMs. To this end, some characteristics needed to be emphasized including:

- **Robustness**. Many models are sensitive to distributional differences between training and test data (Mitra and Craswell, 2017; Zhan et al., 2022). It is eager to see the generalization of LLM-enhanced IR models in out-of-distribution (OOD) scenarios.
- **Interpretability**. Neural IR models rely on dense document representations, and suffer from poor interpretability, which is different from previous sparse IR models (e.g., BM25) with explicit term matching (Llordes et al., 2023).
- **Efficiency**. LLMs may require additional training or fine-tuning to adapt to IR tasks, and storage is also a severe bottleneck in the training scenarios (Santhanam et al., 2022; He et al., 2022a).
- **Reliability**. LLMs are capable of directly retrieving knowledge from the model itself and generating results for users. However, they are vulnerable to adversarial inputs and some small character changes would negatively affect its reliability (Shen et al., 2023).

3.5. Interaction

The interaction between users and traditional search engines typically involves three steps (Baeza-Yates et al., 1999). First, a user submits a query generally expressed as keywords. Then, the search engine processes the query, retrieves its index and presents the results in the form of a list. Finally, the user browses the results and clicks on the most relevant links to the query.

With the development of conversational technology, conversational search has become a new retrieval paradigm (Radlinski and Craswell, 2017; Ren et al., 2021). The retrieval process has become an iterative form of multiple rounds of dialogue. A user initiates the conversation by asking a question in natural language. The search engine processes the user's input, retrieves its index and generates a response. Then, the user may provide additional information or ask follow-up questions, and the search engine continues to process the user's input and generate responses until the user is satisfied.

LLMs (Liu et al., 2023a) have the potential to overturn the interaction between users and IR systems. For example, search engines powered by LLMs can serve as an AI assistant, such as Windows Copilot, Bing Chat. It can generate responses based on the conversation's context and the user's needs in real-time, without relying on pre-programmed responses. As long as the user inputs a question, it can help retrieve information, search knowledge, use Apps, and call plugins in the simplest way (Schick et al., 2023). The IR system powered by LLMs possesses the capability to comprehend and address complex queries, thereby enhancing the interaction process in terms of intuitiveness, personalization, efficiency, effectiveness, responsiveness, and friendliness.

4. IR for large language models

LLMs exhibit certain inherent limitations that are commonly encountered in generative language models. Firstly, they may occasionally generate erroneous or nonsensical responses, a phenomenon

referred to as "hallucinations".[4] Secondly, they are trained on fixed corpora, which restricts their ability to answer questions that require newly emerged knowledge or information after the training date (Komeili et al., 2021; Lazaridou et al., 2022); the stored knowledge within the model can be inevitably incomplete, outdated, or even incorrect (He et al., 2022a). Thirdly, they can hardly respond to "private" questions that require access to confidential data sources, which are not available during the training of LLMs.

Utilizing the retrieval capabilities of IR models presents a viable approach for addressing the above limitations of LLMs. By incorporating retrieval, we can leverage relevant knowledge from external knowledge bases during the generation process, thereby reducing the occurrence of hallucinations. Retrieval models possess strong ability to access up-to-date information, enabling LLMs to respond with fresh and relevant information. Moreover, the combination of retrieval and LLMs is particularly valuable in data-sensitive scenarios, where internal data cannot be utilized for language model training, factual accuracy is of utmost importance, or the retrieval pool may vary over time (Lee et al., 2022a; Petroni et al., 2019).

In this section, we discuss three directions that augment LLMs with retrieval in different phrases: pre-training, fine-tuning, and leveraging black-box LLMs such as ChatGPT that only offer APIs.

4.1. Pre-training LLMs with retrieval

There is limited work on pre-training LLMs that incorporate a retrieval module. One notable example is Atlas (Izacard et al., 2022), which pre-trains an encoder–decoder T5 with an extra retrieval module and demonstrates its ability on knowledge-intensive tasks with very few training samples. REALM enhances encoder-only language model training by incorporating a neural knowledge retriever that extracts information from a text-based knowledge database (Guu et al., 2020). RETRO is a retrieval-augmented decoder-only language model, where a chunked cross-attention module is employed to aggregate retrieved text from a retrieval pool containing trillions of tokens (Borgeaud et al., 2022).

In a recent study, based on the auto-regressive language model RETRO, Wang et al. (2023b) studies the question whether it is useful to pre-train LLMs with retrieval. As a result, the pre-trained retrieval-augmented language model RETRO outperforms the vanilla GPT on text generation tasks with higher factual accuracy, and on knowledge-intensive tasks. In addition, in open-domain question answering tasks, RETRO largely outperforms GPT which just incorporates retrieval at the fine-tuning stage. Note that there is a trade-off between the computation overhead from using retrieval and the performance. Wang et al. (2023b) suggest a flexible implementation that specifies the number of tokens to generate with the current retrieval result before the next retrieval.

In addition to incorporating a retrieval module into the pre-training of LLMs, it is also valuable to investigate how retrieval functionalities can enhance the pre-training process itself. Generally, the retrieved evidence serves as the *privileged information* within the input context. Such privileged information is often costly or impractical to obtain during the inference stage, though it is available during the training phase (Xu et al., 2020). By having both the retrieved evidence and the processed output answer (or subsequent tokens) accessible during training, it is possible to pre-train (or simply fine-tune) LLMs to improve their performance on knowledge-intensive tasks (Nakano et al., 2021; Schick et al., 2023).

[4] Introducing ChatGPT, https://openai.com/blog/chatgpt

4.2. Fine-tuning LLMs with retrieval adapters

Pre-training or fine-tuning the entire LLM with retrieval could enhance its retrieval capabilities. However, the considerable cost makes it prohibitive to use. Considering that LLMs serve as a foundation for downstream tasks, updating all parameters should be approached cautiously and infrequently. Therefore, an alternative is to fine-tune LLMs with search adapters, which is more efficient and cost-effective. Adapters are plug-and-play modules for LLM with only a small number of parameters. They enable LLMs to acquire task-specific capabilities without affecting the original parameters, while still achieving comparable or superior performance compared to updating the entire models.

There have been several studies on LLM adapters, which can be roughly classified into two categories: token-based methods and layer-based methods (He et al., 2022b; Hu et al., 2023). The former seeks to insert task-related anchor tokens into the input sequence for fine-tuning (Li and Liang, 2021; Liu et al., 2022, 2021). For example, Prompt tuning prepends several additional task-specific tunable tokens into the input sequence (Lester et al., 2021). The latter seeks to insert additional layers into the models and fine-tune these layers only (Houlsby et al., 2019; Pfeiffer et al., 2021; Zhang et al., 2023b). For example, Hu et al. (2021) propose LoRA which adds a trainable low-rank dense layer before the self-attention in transformers. LLMs with adapters have been demonstrated effective on some downstream tasks such as machine reading comprehension.

Nevertheless, the following research questions remain to be explored to empower LLMs with retrieval capabilities through adapters.

- What adapters are best suited for retrieval? What neural network modules are best suited for retrieval adapters?
- How to fine-tune the retrieval adapters? Which fine-tuning techniques can promote the retrieval capability in highest degree without hurting the inherent capabilities of LLM?
- Why do retrieval adapters work? How do we know the capability boundary of retrieval adapters, and what can we do to avoid potential failures?

Fine-tuning LLMs with retrieval adapters can provide benefits in various aspects. On one hand, it offers new opportunities for researchers with limited computation resources. Progress in scientific research cannot be achieved without the participation of researchers around the world. On the other hand, the lower cost of search adapters allows for wider applications, especially when data privacy is a top priority. Small-size institutions can have LLMs equipped with their own retrieval systems without exposing the raw data.

4.3. Augmenting black-box LLMs with retrieval

In many scenarios, LLMs can only be accessed through remote APIs, and the possibility of fine-tuning these models is restricted. The most prevalent approach to leveraging LLMs is treating them as black-box systems and using customized prompts that integrate evidence from external data sources. This method enables users to obtain desired outputs from the LLMs by providing tailored prompts that incorporate relevant external evidence.

There are some preliminary studies on this interesting topic. Shuster et al. (2022) proposed a modular system SeeKeR which searches for and chooses knowledge with internet search as a module during language generation. Similarly, Komeili et al. (2021) and Lazaridou et al. (2022) also introduced approaches which condition on the results from internet search engines (such as Google Search) to generate a response. He et al. (2022a) proposed a post-processing approach named "rethinking with retrieval (RR)" to solve the same problem. RR uses the chain-of-thought (COT) prompting to generate multiple reasoning paths, then retrieves external evidences based on the steps

in these paths. Experiments on three complex reasoning tasks and different datasets demonstrate that RR outperforms all baselines without additional pre-training or fine-tuning. Ram et al. (2023) introduced an in-context Retrieval-Augmented Language Modeling (RALM) method that simply uses off-the-shelf general purpose retrievers to retrieve documents and preprends retrieved results to the input of language models. Peng et al. (2023) proposed a system named LLM-Augmenter, which augments a fixed black-box LLM with a set of plug-and-play modules. Given a query, LLM-Augmenter first retrieves evidences from external data sources, then generates a prompt that contains the retrieved evidences for ChatGPT. Their experiments on the information seeking dialog and open-domain wiki question answering tasks show this method could significantly reduce the hallucination problem of LLMs.

Recently, OpenAI officially released the Retrieval Plugin,[5] which provides semantic search and retrieval functionality of personal or organizational documents. Relevant documents snippets can be retrieved from external sources, such as personal emails or internal organizational files, with this plugin.

There are several key components in the pipeline of augmenting LLMs with retrieval, which are worth to study in the future.

- **Design of Retriever**. Intuitively, a better retriever could return results with higher quality, and thereby helps to generate better response. Should we use a dense retriever or a sparse retriever in generating grounding documents? Which kind of information to index, documents or passages?
- **Context Modeling**. How to generate an explicit keyword query or representation vector that can describe the current information need and retrieve useful results for language generation? ChatGPT-like LLMs are optimized for dialogue rather than search, hence how to derive search intent from the multi-turn chat history remains a challenge.
- **Selection of Grounding Documents**. In addition to pure relevance, other factors of diversity, informativeness, and freshness should also be considered when selecting a small set of documents.
- **Prompting Mechanism**. It remains a challenge to generate good prompts that can help improve generation quality with the given grounding documents.

5. LLMs + IR: New paradigm and framework

Traditional IR models are designed to meet human information needs through the interactions between users and systems. IR models do not preserve knowledge, instead, they search for external knowledge or information to meet the information needs efficiently. With the emergence of LLMs, we propose a new technical paradigm of IR: as shown in Fig. 1, the key elements are LLMs, IR models and humans.

- LLMs provide valuable (internal) knowledge and information to meet human information needs, and their reasoning capacity makes it easier to provide high-quality responses.
- IR models that search for external knowledge are indispensable, which provide up-to-date and relevant information for both LLMs and humans. While LLMs may suffice in directly generating answers to meet human information needs for simple questions, the significance of IR models becomes more pronounced when dealing with complex and difficult problems.
- Humans raise information needs in the paradigm and are the tutor for both LLMs and IR models. They endow the system with human values and behavioral characteristics, making it serve users better.

The synergistic relationship among them not only facilitates mutual enhancement but also enables the fulfillment of new tasks as an organic whole.

[5] ChatGPT Retrieval Plugin, https://github.com/openai/chatgpt-retrieval-plugin

5.1. Importance of the three modules

In accordance with the new IR technical paradigm, an important question arises: Are all three elements essential for this new paradigm? In the following subsection, we will provide our insights and responses.

5.1.1. Paradigm without LLMs

Without LLMs, the paradigm degrades to traditional IR, accompanied by inherent challenges:

- **Dependency on the Internet**. As IR models do not retain knowledge or information themselves, they rely on the Internet to acquire external knowledge, potentially limiting their applicability in certain scenarios.
- **Lacking reasoning ability.** Existing IR models mainly provide collected knowledge/information to fulfill human information needs, lacking the ability to assist users in comprehending the information. Better reasoning ability will deliver more user-friendly and valuable results for humans.

5.1.2. Paradigm without IR

In the era of LLMs, the presence of an IR system is critical as it addresses the limitations of generative language models. Without an IR system, LLMs encounter the following challenges:

- **Lacking factual consistency**. One inherent limitation of generative LLMs is the lack of factual consistency, resulting from their training data and the potential generation of false information. In contrast, the IR system frees from this issue by storing and offering factual information through keyword matching, thereby complementing the shortcomings of LLMs.
- **Lacking effective integration of new and existing knowledge.** Large Language models have the capacity to learn knowledge from large-scale data, such as commonsense information. However, relying solely on LLMs may neglect the crucial mechanism of effectively and efficiently integrating new and existing knowledge. In the human brain, memory involves a complex network of interconnected brain regions, including the hippocampus, prefrontal cortex, and other cortical areas. Retrieval processes are closely linked to memory consolidation and the reactivation of neural networks associated with previously encoded information. When we retrieve information from memory, this activity can strengthen the neural connections related to that information, making it more accessible for future retrieval and improving overall memory performance.

5.1.3. Paradigm without human

The absence of human input and feedback hinders both LLMs and IR models from delivering personalized information services. The associated challenges are:

- **Identical information services.** Large-scale dialogue systems use a generation-based approach to provide information to users. The generated content is not personalized, which can cause the issues that the content does not match the user's real intention.
- **Uncontrollable value of society.** Without human feedback, LLMs are unable to account for user personality and values, resulting in information services that do not adequately align with societal values.

5.2. Summary

The rise of dialogue-based LLMs, exemplified by platforms like ChatGPT from OpenAI and New Bing from Microsoft, has sparked a trend that could potentially replace traditional IR systems. However, as previously discussed, in the era of LLMs, human information needs and feedback, along with the continued relevance of traditional IR models, remain vital components in the development of trustworthy information service systems.

6. Challenges and future

While LLMs for IR hold promise, they also present numerous challenges and unanswered questions. In the final section of this article, we discuss some selected issues to outline future directions.

- **High Computational Costs**. The primary challenge in using LLMs is their high computational cost. This poses a significant barrier for small and medium-sized research laboratories and companies, hindering their integration of LLMs into daily workflows and products. Even large companies with ample computational resources face cost pressures when deploying LLMs for online search, recommendation, and advertisement services due to the immense volume of user requests. Common solutions include compressing LLMs, reducing their size from hundreds of billions to tens billions or even smaller, especially before online deployment. Additionally, efforts to develop more efficient and cost-effective hardware for training and inference are underway to address the cost challenge.
- **General-purpose v.s. Domain-specific**. LLMs have demonstrated impressive capabilities in general-purpose tasks like text generation and chatting, owing to their pre-training and fine-tuning on large-scale Internet corpora. However, it is widely recognized that LLMs face limitations when it comes to adapting to domain-specific tasks. On one hand, high-quality professional domain knowledge, which is often not abundantly available on the Internet, makes it prohibitive to pre-train and fine-tune LLMs. On the other hand, domain-specific knowledge is not always expressed in natural language; it may be represented as semi-structured or structured tables, heuristic rules, equations, and more. Enabling LLMs to effectively handle domain-specific tasks is crucial not only for the specific domains themselves but also for enhancing the overall capabilities and applications of LLMs.
- **Trustworthiness**. There is a widely acknowledged concern that LLMs currently lack the ability to provide reliable and trustworthy answers to user queries. While LLMs can generate explanations and cite sources, it has been observed that a significant portion of these explanations and citations are illogical, inappropriate, or even fake. This poses a substantial risk in real-world search and recommendation scenarios, as generating misleading explanations, answers, and information sources can have detrimental effects on the community at large. To address this issue and enhance the trustworthiness of LLMs, it is crucial to enable LLMs to have a clear understanding of their knowledge and limitations. One potential solution is allowing LLMs to decline providing an answer when uncertain.
- **Controllable Generation**. Considering the public nature of search engines and recommendation systems, it is important to address regulatory and ethical considerations such as fairness, impartiality, and human values when presenting content to users. While LLMs demonstrate proficiency in generating text, they often lack a deep understanding of the meaning behind the generated words. Ensuring that the generated content meets the necessary regulatory and ethical requirements remains a significant challenge, and has no effective solutions for now.
- **High-quality Data:** High-quality data plays a vital role in the development and improvement of LLMs. The success of LLMs heavily relies on the continuous provision of human-labeled data. It is crucial that the labeled data not only meets a certain quantity threshold but also maintains high quality. Obtaining high-quality data in real-world application scenarios involves multiple steps such as data cleaning, data labeling, and data quality evaluation. Professional data annotation providers play a crucial role in supporting these processes. Additionally, it is essential to develop advanced, professional, and sustainable data annotation approaches to meet the growing demand for high-quality data in LLM applications.
- **Long-context Dependency**. Existing LLMs have limited capacity to handle long contexts, while IR tasks rely on long-term context to effectively capture and understand user intent. It is crucial to enable LLM-enhanced IR systems to model users' long-term intent that spans a large range of period.
- **Serving Time Requirements**. The latency in serving LLM results significantly lags behind the time requirements of information retrieval (IR) systems. This presents efficiency challenges when integrating LLMs into IR, thereby impacting the online user experience.
- **Presentation Format**. Traditional IR systems present ranked lists of content, while LLMs excel at generating new information. How to design a new presentation format that effectively fulfills user needs in LLM-enhanced IR remains an open question.
- **Integrating Structural Information**. LLMs primarily rely on textual sequence information, whereas IR systems require the integration of structural information such as user–item interactions and web linkage data. Effectively leveraging this structural information in LLM-based IR systems is an unresolved issue.
- **Balance between Generative and Retrieved Data**. LLMs leverage deep learning and reinforcement learning to generate content at scale, but their generated content may have limitations in terms of freshness and credibility. In contrast, retrieval can provide the content from the Web, ensuring the latest information. Balancing these two types of data in real-life applications is a significant challenge for improving overall performance. One approach is to refine user needs and classify them into different groups, allowing for appropriate data generation methods or balancing ratios. In addition, retrieval can be used to provide additional information to enhance content generation, or assist in screening and filtering the generated content for information grounding.
- **Content Quality and Credibility**. While LLMs are effective in generating content, they can also produce low-quality or even misinformation-filled content. The proliferation of such content on the Internet can disrupt the existing data ecology and impact applications like search engine and recommendation system. Traditional quality assessment techniques like PageRank may not be effective in this context. It is difficult for existing techniques to identify low-quality or misleading content. In the era of AI-generated content, new mechanisms are needed to assess data quality and distinguish between generated and reliable content. One approach is manual or community review to ensure content accuracy, but it is time-consuming and does not scale well. Another approach is leveraging LLM-powered machine learning techniques to train models that can recognize AI-generated content and evaluate its quality. These models can then be applied to select, label, or filter content for different applications.
- **Content Creation Environment**. The proliferation of generated content introduces challenges for content creators, and may even reshape the content ecosystem. The presence of generated content intensifies competition within the content market, pushing content creators to continuously enhance their writing quality, foster innovation, and proactively adapt to industry changes. Moreover, it may influence users' perception of value and demand for content, prompting content creators to constantly adjust their writing approach and strategies. Despite the challenges, LLMs also provide new opportunities for content creators to collaborate and develop effective platforms that facilitate more productive and higher-quality content generation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Agichtein, E., Brill, E., Dumais, S., 2006. Improving web search ranking by incorporating user behavior information. 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. 19–26.

Alfieri, A., Wolter, R., Hashemi, S.H., 2022. Intent disambiguation for task-oriented dialogue systems. In: CIKM '22. Association for Computing Machinery, New York, NY, USA, pp. 5079–5080.

Baeza-Yates, R., Ribeiro-Neto, B., et al., 1999. Modern Information Retrieval, vol. 463.

Bao, K., Zhang, J., Zhang, Y., Wang, W., Feng, F., He, X., 2023. Tallrec: An effective and efficient tuning framework to align large language model with recommendation. In: Recsys. ACM.

Blair-Stanek, A., Holzenberger, N., Durme, B.V., 2023. Can GPT-3 perform statutory reasoning? arXiv preprint arXiv:2302.06100.

Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., van den Driessche, G., Lespiau, J., Damoc, B., Clark, A., de Las Casas, D., Guy, A., Menick, J., Ring, R., Hennigan, T., Huang, S., Maggiore, L., Jones, C., Cassirer, A., Brock, A., Paganini, M., Irving, G., Vinyals, O., Osindero, S., Simonyan, K., Rae, J.W., Elsen, E., Sifre, L., 2021. Improving language models by retrieving from trillions of tokens. arXiv preprint arXiv:2112.04426.

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., 2022. Improving language models by retrieving from trillions of tokens. In: International Conference on Machine Learning. PMLR, pp. 2206–2240.

Borlund, P., 2013. Interactive information retrieval: An introduction. J. Inf. Sci. Theory Pract. 1 (3), 12–32.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., 2020. Language models are few-shot learners. In: NeurIPS. Curran Associates, Inc., pp. 1877–1901.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S., et al., 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. arXiv preprint arXiv:2303.12712.

Burges, C.J.C., 2010. From RankNet to LambdaRank to LambdaMART: An overview.

Chen, Z., Cheng, X., Dong, S., Dou, Z., Guo, J., Huang, X., Lan, Y., Li, C., Li, R., Liu, T.-Y., Liu, Y., Ma, J., Qin, B., Wang, M., rong Wen, J., Xu, J., Zhang, M., Zhang, P., Zhang, Q., 2020. Information retrieval: A view from the Chinese IR community. Front. Comput. Sci. 15, 1–15.

Chen, L., Zhang, D., Mark, L., 2012. Understanding user intent in community question answering. In: Proceedings of the 21st International Conference on World Wide Web. In: WWW '12 Companion, Association for Computing Machinery, New York, NY, USA, pp. 823–828.

Craswell, N., 2009. Mean reciprocal rank. In: Encyclopedia of Database Systems.

Dennis, S., Bruza, P., McArthur, R., 2002. Web searching: A process-oriented experimental study of three interactive search paradigms. J. Assoc. Inf. Sci. Technol. 53 (2), 120–133.

Duh, K., Kirchhoff, K., 2008. Learning to rank with partially-labeled data. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '08, Association for Computing Machinery, New York, NY, USA, pp. 251–258.

Edalati, A., Tahaei, M., Rashid, A., Nia, V.P., Clark, J.J., Rezagholizadeh, M., 2021. Kronecker decomposition for GPT compression. arXiv preprint arXiv:2110.08152.

Esposito, D., Centracchio, J., Andreozzi, E., Gargiulo, G.D., Naik, G.R., Bifulco, P., 2021. Biosignal-based human–machine interfaces for assistance and rehabilitation: A survey. Sensors 21 (20), 6863.

Faggioli, G., Ferrante, M., Ferro, N., Perego, R., Tonellotto, N., 2021. Hierarchical dependence-aware evaluation measures for conversational search. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1935–1939.

Feng, J., Tao, C., Geng, X., Shen, T., Xu, C., Long, G., Zhao, D., Jiang, D., 2023. Knowledge refinement via interaction between search engines and large language models. arXiv preprint arXiv:2305.07402.

Ford, N., Miller, D., O'rourke, A., Ralph, J., Turnock, E., Booth, A., 1999. Information retrieval for evidence-based decision making. J. Doc. 55 (4), 385–401.

Freund, Y., Iyer, R., Schapire, R.E., Singer, Y., 2003. An efficient boosting algorithm for combining preferences. J. Mach. Learn. Res. 4 (null), 933–969.

Fujiwara, Y., Nakatsuji, M., Shiokawa, H., Mishima, T., Onizuka, M., 2013. Efficient Ad-Hoc search for personalized PageRank. In: Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data. SIGMOD '13, Association for Computing Machinery, New York, NY, USA, pp. 445–456.

Gao, Y., Sheng, T., Xiang, Y., Xiong, Y., Wang, H., Zhang, J., 2023. Chat-REC: Towards interactive and explainable LLMs-augmented recommender system. CoRR abs/2303.14524.

Guo, Y., Cheng, Z., Nie, L., Wang, Y., Ma, J., Kankanhalli, M.S., 2019. Attentive long short-term preference modeling for personalized product search. ACM Trans. Inf. Syst. 37 (2), 19:1–19:27.

Guo, J., Fan, Y., Ai, Q., Croft, W.B., 2016. A deep relevance matching model for Ad-Hoc retrieval. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. CIKM '16, Association for Computing Machinery, New York, NY, USA, pp. 55–64.

Guo, J., Fan, Y., Pang, L., Yang, L., Ai, Q., Zamani, H., Wu, C., Croft, W.B., Cheng, X., 2020. A deep look into neural ranking models for information retrieval. Inf. Process. Manage. 57 (6), 102067.

Gur, S., Neverova, N., Stauffer, C., Lim, S.-N., Kiela, D., Reiter, A., 2021. Cross-modal retrieval augmentation for multi-modal classification. In: Findings of the Association for Computational Linguistics. EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, pp. 111–123. http://dx.doi.org/10.18653/v1/2021.findings-emnlp.11, URL https://aclanthology.org/2021.findings-emnlp.11.

Guu, K., Lee, K., Tung, Z., Pasupat, P., Chang, M.-w., 2020. REALM: Retrieval-augmented language model pre. In: ICML.

He, H., Zhang, H., Roth, D., 2022a. Rethinking with retrieval: Faithful large language model inference. arXiv preprint arXiv:2301.00303.

He, J., Zhou, C., Ma, X., Berg-Kirkpatrick, T., Neubig, G., 2022b. Towards a unified view of parameter-efficient transfer learning. In: International Conference on Learning Representations.

Hersh, W.R., Buckley, C., Leone, T.J., Hickam, D.H., 1994. OHSUMED: An interactive retrieval evaluation and new large test collection for research. In: Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. ACM/Springer, pp. 192–201.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S., 2019. Parameter-efficient transfer learning for NLP. In: International Conference on Machine Learning. pp. 2790–2799.

Hu, Z., Lan, Y., Wang, L., Xu, W., Lim, E.-P., Lee, R.K.-W., Bing, L., Poria, S., 2023. LLM-adapters: An adapter family for parameter-efficient fine-tuning of large language models. arXiv preprint arXiv:2304.01933.

Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., 2021. Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685.

Huang, P.-S., He, X., Gao, J., Deng, L., Acero, A., Heck, L., 2013. Learning deep structured semantic models for web search using clickthrough data. In: CIKM '13. Association for Computing Machinery, New York, NY, USA, pp. 2333–2338.

Ingwersen, P., 1984. Psychological aspects of information retrieval. Soc. Sci. Inf. Stud. 4 (2–3), 83–95.

Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., Grave, E., 2021. Towards unsupervised dense information retrieval with contrastive learning. arXiv preprint arXiv:2112.09118.

Izacard, G., Lewis, P., Lomeli, M., Hosseini, L., Petroni, F., Schick, T., Dwivedi-Yu, J., Joulin, A., Riedel, S., Grave, E., 2022. Atlas: Few-shot learning with retrieval augmented language models. arXiv preprint arXiv. 2208.

Janner, M., 2023. Deep generative models for decision-making and control. arXiv preprint arXiv:2306.08810.

Jansen, B.J., Booth, D.L., Spink, A., 2007. Determining the user intent of web search engine queries. In: Proceedings of the 16th International Conference on World Wide Web. WWW '07, Association for Computing Machinery, New York, NY, USA, pp. 1149–1150.

Järvelin, K., Kekäläinen, J., 2002. Cumulated gain-based evaluation of IR techniques. ACM Trans. Inf. Syst. (TOIS) 20 (4), 422–446.

Jeronymo, V., Bonifacio, L., Abonizio, H., Fadaee, M., Lotufo, R., Zavrel, J., Nogueira, R., 2023. Inpars-v2: Large language models as efficient dataset generators for information retrieval. arXiv preprint arXiv:2301.01820.

Jiang, Y., Zhang, P., Gao, H., Song, D., 2020. A quantum interference inspired neural matching model for Ad-Hoc retrieval. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '20, Association for Computing Machinery, New York, NY, USA, pp. 19–28.

Jin, L., Chen, Y., Wang, T., Hui, P., Vasilakos, A.V., 2013. Understanding user behavior in online social networks: A survey. IEEE Commun. Mag. 51 (9), 144–150. http://dx.doi.org/10.1109/MCOM.2013.6588663.

John, R.I., Mooney, G.J., 2001. Fuzzy user modeling for information retrieval on the world wide web. Knowl. Inf. Syst. 3 (1), 81–95.

Karatzoglou, A., Baltrunas, L., Shi, Y., 2013. Learning to rank for recommender systems. In: Proceedings of the 7th ACM Conference on Recommender Systems. RecSys '13, Association for Computing Machinery, New York, NY, USA, pp. 493–494.

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., Yih, W.-t., 2020. Dense passage retrieval for open-domain question answering. arXiv preprint arXiv:2004.04906.

Kim, S.Y., Park, H., Shin, K., Kim, K.-M., 2022. Ask me what you need: Product retrieval using knowledge from GPT-3. arXiv preprint arXiv:2207.02516.

Kobayashi, M., Takeda, K., 2000a. Information retrieval on the web. ACM Comput. Surv. (CSUR) 32 (2), 144–173.

Kobayashi, M., Takeda, K., 2000b. Information retrieval on the web. ACM Comput. Surv. 32 (2), 144–173.

Komeili, M., Shuster, K., Weston, J., 2021. Internet-augmented dialogue generation. arXiv preprint arXiv:2107.07566.

Lazaridou, A., Gribovskaya, E., Stokowiec, W., Grigorev, N., 2022. Internet-augmented language models through few-shot prompting for open-domain question answering. arXiv preprint arXiv:2203.05115.

Lee, N., Ping, W., Xu, P., Patwary, M., Fung, P.N., Shoeybi, M., Catanzaro, B., 2022a. Factuality enhanced language models for open-ended text generation. Adv. Neural Inf. Process. Syst. 35, 34586–34599.

Lee, H., Yang, S., Oh, H., Seo, M., 2022b. Generative retrieval for long sequences. arXiv preprint arXiv:2204.13596.

Lester, B., Al-Rfou, R., Constant, N., 2021. The power of scale for parameter-efficient prompt tuning. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 3045–3059.

Li, J., Galley, M., Brockett, C., Spithourakis, G.P., Gao, J., Dolan, W.B., 2016. A persona-based neural conversation model. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers. The Association for Computer Linguistics.

Li, X.L., Liang, P., 2021. Prefix-tuning: Optimizing continuous prompts for generation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 4582–4597.

Li, S., Xie, R., Zhu, Y., Ao, X., Zhuang, F., He, Q., 2022. User-centric conversational recommendation with multi-aspect user modeling. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 223–233.

Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., et al., 2023a. Summary of chatgpt/gpt-4 research and perspective towards the future of large language models. arXiv preprint arXiv:2304.01852.

Liu, X., Ji, K., Fu, Y., Tam, W., Du, Z., Yang, Z., Tang, J., 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 61–68.

Liu, H., Ning, R., Teng, Z., Liu, J., Zhou, Q., Zhang, Y., 2023b. Evaluating the logical reasoning ability of chatgpt and GPT-4. arXiv preprint arXiv:2304.03439.

Liu, Z., Yu, X., Zhang, L., Wu, Z., Cao, C., Dai, H., Zhao, L., Liu, W., Shen, D., Li, Q., Liu, T., Zhu, D., Li, X., 2023c. DeID-GPT: Zero-shot medical text de-identification by GPT-4. arXiv preprint arXiv:2303.11032.

Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z., Tang, J., 2021. GPT understands, too. arXiv preprint arXiv:2103.10385.

Llordes, M., Ganguly, D., Bhatia, S., Agarwal, C., 2023. Explain like I am BM25: Interpreting a dense model's ranked-list with a sparse approximation. CoRR abs/2304.12631.

Manavoglu, E., Pavlov, D., Giles, C., 2003. Probabilistic user behavior models. In: Third IEEE International Conference on Data Mining. pp. 203–210.

Manning, C.D., Raghavan, P., Schütze, H., 2008. Introduction to Information Retrieval. Cambridge University Press, Cambridge, UK.

Manoj, M., Elizabeth, J., 2008. Information retrieval on internet using meta-search engines: A review. J. Sci. Ind. Res. 67 (10).

Mitra, B., Craswell, N., 2017. Neural models for information retrieval. CoRR abs/1705.01509.

Mohsan, S.A.H., Khan, M.A., Noor, F., Ullah, I., Alsharif, M.H., 2022. Towards the unmanned aerial vehicles (UAVs): A comprehensive review. Drones 6 (6), 147.

Mystakidis, S., 2022. Metaverse. In: Encyclopedia. vol. 2, (no. 1), MDPI, pp. 486–497.

Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., et al., 2021. Webgpt: Browser-assisted question-answering with human feedback. arXiv preprint arXiv:2112.09332.

Nori, H., King, N., McKinney, S.M., Carignan, D., Horvitz, E., 2023. Capabilities of GPT-4 on medical challenge problems. arXiv preprint arXiv:2303.13375.

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., 2022a. Training language models to follow instructions with human feedback. In: Advances in Neural Information Processing Systems, 35, pp. 27730–27744.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L.E., Simens, M., Askell, A., Welinder, P., Christiano, P.F., Leike, J., Lowe, R.J., 2022b. Training language models to follow instructions with human feedback. In: NeurIPS.

Pang, L., Lan, Y., Guo, J., Xu, J., Xu, J., Cheng, X., 2017. DeepRank: A new deep architecture for relevance ranking in information retrieval. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. CIKM '17, Association for Computing Machinery, New York, NY, USA, pp. 257–266.

Peng, B., Galley, M., He, P., Cheng, H., Xie, Y., Hu, Y., Huang, Q., Liden, L., Yu, Z., Chen, W., Gao, J., 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. arXiv preprint arXiv:2302.12813.

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., Miller, A., 2019. Language models as knowledge bases? In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. EMNLP-IJCNLP, pp. 2463–2473.

Pfeiffer, J., Kamath, A., Rücklé, A., Cho, K., Gurevych, I., 2021. AdapterFusion: Non-destructive task composition for transfer learning. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. pp. 487–503.

Pi, Q., Bian, W., Zhou, G., Zhu, X., Gai, K., 2019. Practice on long sequential user behavior modeling for click-through rate prediction. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. KDD '19, Association for Computing Machinery, New York, NY, USA, pp. 2671–2679.

Qin, Y., Cai, Z., Jin, D., Yan, L., Liang, S., Zhu, K., Lin, Y., Han, X., Ding, N., Wang, H., Xie, R., Qi, F., Liu, Z., Sun, M., Zhou, J., 2023. Webcpm: Interactive web search for Chinese long-form question answering. In: Proceedings of ACL 2023. Association for Computational Linguistics.

Radlinski, F., Craswell, N., 2017. A theoretical framework for conversational search. In: Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval. pp. 117–126.

Ram, O., Levine, Y., Dalmedigos, I., Muhlgay, D., Shashua, A., Leyton-Brown, K., Shoham, Y., 2023. In-context retrieval-augmented language models. arXiv preprint arXiv:2302.00083.

Ren, P., Chen, Z., Ren, Z., Kanoulas, E., Monz, C., De Rijke, M., 2021. Conversations with search engines: SERP-based conversational response generation. ACM Trans. Inf. Syst. (TOIS) 39 (4), 1–29.

Robertson, S., Zaragoza, H., 2009. The probabilistic relevance framework: BM25 and beyond 3 (4). pp. 333–389.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2022. High-resolution image synthesis with latent diffusion models. In: CVPR. IEEE, pp. 10684–10695.

Santhanam, K., Saad-Falcon, J., Franz, M., Khattab, O., Sil, A., Florian, R., Sultan, M.A., Roukos, S., Zaharia, M., Potts, C., 2022. Moving beyond downstream task accuracy for information retrieval benchmarking. CoRR abs/2212.01340.

Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., Lillicrap, T., 2016. Meta-learning with memory-augmented neural networks. In: Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48. ICML '16, JMLR.org, pp. 1842–1850.

Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., Scialom, T., 2023. Toolformer: Language models can teach themselves to use tools. arXiv preprint arXiv:2302.04761.

Shen, X., Chen, Z., Backes, M., Zhang, Y., 2023. In ChatGPT we trust? Measuring and characterizing the reliability of ChatGPT. arXiv preprint arXiv:2304.08979.

Shenavarmasouleh, F., Mohammadi, F.G., Amini, M.H., Reza Arabnia, H., 2022. Embodied AI-driven operation of smart cities: A concise review. Cyberphys. Smart Cities Infrastruct.: Optim. Oper. Intell. Decis. Making 29–45.

Shuster, K., Komeili, M., Adolphs, L., Roller, S., Szlam, A., Weston, J., 2022. Language models that seek for knowledge: Modular search & generation for dialogue and prompt completion. arXiv preprint arXiv:2203.13224.

Su, N., He, J., Liu, Y., Zhang, M., Ma, S., 2018. User intent, behaviour, and perceived satisfaction in product search. In: WSDM '18. Association for Computing Machinery, New York, NY, USA, pp. 547–555.

Sun, W., Guo, S., Zhang, S., Ren, P., Chen, Z., de Rijke, M., Ren, Z., 2022. Metaphorical user simulators for evaluating task-oriented dialogue systems. ACM Trans. Inf. Syst..

Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A., Hardt, M., 2020. Test-time training with self-supervision for generalization under distribution shifts. In: International Conference on Machine Learning. PMLR, pp. 9229–9248.

Sun, W., Yan, L., Ma, X., Ren, P., Yin, D., Ren, Z., 2023. Is ChatGPT good at search? Investigating large language models as re-ranking agent. arXiv preprint arXiv:2304.09542.

Tay, Y., Tran, V., Dehghani, M., Ni, J., Bahri, D., Mehta, H., Qin, Z., Hui, K., Zhao, Z., Gupta, J., et al., 2022. Transformer memory as a differentiable search index. Adv. Neural Inf. Process. Syst. 35, 21831–21843.

Teevan, J., Dumais, S.T., Liebling, D.J., 2008. To personalize or not to personalize: Modeling queries with variation in user intent. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Association for Computing Machinery, pp. 163–170.

Trabelsi, M., Chen, Z., Davison, B.D., Heflin, J., 2021. Neural ranking models for document retrieval. Inf. Retr. 24 (6), 400–444.

Tsvetkov, V.Y., 2015. Cognitive science of information retrieval. Eur. J. Psychol. Stud. (1), 37–44.

Vargas, S., Castells, P., 2011. Rank and relevance in novelty and diversity metrics for recommender systems. In: Proceedings of the Fifth ACM Conference on Recommender Systems. RecSys '11, Association for Computing Machinery, New York, NY, USA, pp. 109–116.

Vassileva, J., 2012. Motivating participation in social computing applications: A user modeling perspective. User Model. User Adapt. Interact. 22 (1–2), 177–201.

Wang, W., Lin, X., Feng, F., He, X., Chua, T.-S., 2023a. Generative recommendation: Towards next-generation recommender paradigm. arXiv preprint arXiv:2304.03516.

Wang, B., Ping, W., Xu, P., McAfee, L., Liu, Z., Shoeybi, M., Dong, Y., Kuchaiev, O., Li, B., Xiao, C., et al., 2023b. Shall we pretrain autoregressive language models with retrieval? A comprehensive study. arXiv preprint arXiv:2304.06762.

Xu, C., Li, Q., Ge, J., Gao, J., Yang, X., Pei, C., Sun, F., Wu, J., Sun, H., Ou, W., 2020. Privileged features distillation at taobao recommendations. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 2590–2598.

Xu, D., Schnabel, T., Cui, X., Dean, S., Deshmukh, A., Yang, B., Yu, S., 2023. Foreword for workshop on decision making for information retrieval and recommender systems. In: Companion Proceedings of the ACM Web Conference 2023. pp. 920–920.

Yang, L., Qiu, M., Gottipati, S., Zhu, F., Jiang, J., Sun, H., Chen, Z., 2013. CQArank: Jointly model topics and expertise in community question answering. In: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. CIKM '13, Association for Computing Machinery, New York, NY, USA, pp. 99–108.

Yin, H., Sun, Y., Xu, G., Kanoulas, E., 2023. Trustworthy recommendation and search: Introduction to the special issue-part 1. ACM Trans. Inf. Syst. 41 (3), 1–5.

Yuan, F., He, X., Karatzoglou, A., Zhang, L., 2020. Parameter-efficient transfer from sequential behaviors for user modeling and recommendation. In: SIGIR '20. Association for Computing Machinery, New York, NY, USA, pp. 1469–1478.

Yuan, J., Zheng, Y., Xie, X., 2012. Discovering regions of different functions in a city using human mobility and POIs. In: KDD '12. Association for Computing Machinery, New York, NY, USA, pp. 186–194.

Zamani, H., Diaz, F., Dehghani, M., Metzler, D., Bendersky, M., 2022. Retrieval-enhanced machine learning. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '22, Association for Computing Machinery, New York, NY, USA, pp. 2875–2886. http://dx.doi.org/10.1145/3477495.3531722.

Zhan, J., Xie, X., Mao, J., Liu, Y., Guo, J., Zhang, M., Ma, S., 2022. Evaluating interpolation and extrapolation performance of neural retrieval models. In: Proceedings of the 31st ACM International Conference on Information & Knowledge Management. pp. 2486–2496.

Zhang, J., Bao, K., Zhang, Y., Wang, W., Feng, F., He, X., 2023a. Is ChatGPT fair for recommendation? Evaluating fairness in large language model recommendation. In: Recsys. ACM.

Zhang, Q., Chen, M., Bukharin, A., He, P., Cheng, Y., Chen, W., Zhao, T., 2023b. Adaptive budget allocation for parameter-efficient fine-tuning. arXiv preprint arXiv:2303.10512.

Zhang, K., Wu, W., Wu, H., Li, Z., Zhou, M., 2014. Question retrieval with high quality answers in community question answering. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. CIKM '14, Association for Computing Machinery, New York, NY, USA, pp. 371–380.

Zhen, L., Hu, P., Wang, X., Peng, D., 2019. Deep supervised cross-modal retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10394–10403.

Zheng, J., Fischer, M., 2023. BIM-GPT: A prompt-based virtual assistant framework for BIM information retrieval. arXiv preprint arXiv:2304.09333.

Zheng, Y., Xie, X., Ma, W.-Y., 2010. GeoLife: A collaborative social networking service among user, location and trajectory. IEEE Data(base) Eng. Bull..

Zhong, P., Zhang, C., Wang, H., Liu, Y., Miao, C., 2020. Towards persona-based empathetic conversational models. In: Webber, B., Cohn, T., He, Y., Liu, Y. (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020. Association for Computational Linguistics, pp. 6556–6566.

Zhu, M., 2004. Recall, precision and average precision. University of Waterloo,

Zhu, M., Biswas, S., Dinulescu, S.I., Kastor, N., Hawkes, E.W., Visell, Y., 2022. Soft, wearable robotics and haptics: Technologies, trends, and emerging applications. Proc. IEEE 110 (2), 246–272.