Information Retrieval for Artificial General Intelligence: A New Perspective on Information Retrieval Research

ChengXiang Zhai
Siebel School of Computing and Data Science
Grainger College of Engineering
University of Illinois at Urbana-Champaign
Urbana, IL, USA
czhai@illinois.edu

Abstract

Traditionally, the users of an information retrieval (IR) system have been human users. We present a new perspective on IR research in which the users of an IR system are intelligent agents instead of human users. Extending the current work on retrieval-augmented generation (RAG), we identify five novel IR tasks that an intelligent agent must be able to perform in order to achieve Human-Level Artificial Intelligence, or Artificial General Intelligence (AGI), including 1) External Information Retrieval (EIR) to access new information unseen by the agent, 2) Provenance Information Retrieval (PIR) to trace the provenance of information, 3) Curriculum Information Retrieval (CIR) to actively acquire the most useful new data and information for lifelong learning, 4) Rule Information Retrieval (RIR) to perform reasoning and problem solving, and 5) Scenario Information Retrieval (SIR) to leverage past scenarios for problem solving and decision making. We compare these new IR tasks with the traditional IR tasks performed by an IR system that serves human users and systematically examine the challenges involved in the five new IR tasks, providing a roadmap for new IR research within the broader context of AGI development.

CCS Concepts

• **Information systems** → **Information retrieval**; **Retrieval tasks and goals**;

Keywords

Artificial General Intelligence, Intelligent Agent, Large Language Models, Retrieval-Augmented Generation, Neurosymbolic Architecture

ACM Reference Format:

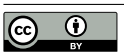
ChengXiang Zhai. 2025. Information Retrieval for Artificial General Intelligence: A New Perspective on Information Retrieval Research . In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), July 13–18, 2025, Padua, Italy.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3726302.3730349

1 Introduction

The recent breakthrough in Artificial Intelligence (AI) toward Human-Level AI [32], or Artificial General Intelligence (AGI) [21], notably the emergence of powerful large foundation models for both language modeling and visual modeling [6, 23, 38, 53], has impacted many research fields, including Information Retrieval (IR) [54]. The most noticeable impact on IR so far has been the explosive growth of work on applying new AI technologies such as Large Language Models (LLMs) to improve the current generation IR systems or develop more intelligent future IR systems [2, 3, 54, 57]. Considering the natural view that IR is an application of AI, such a trend of research is not surprising as IR naturally benefits from any progress in the foundational AI technologies that can empower an IR system. A recent workshop report [3] provides a comprehensive roadmap for many important future research questions.

However, while it is natural to regard IR as an application of AI, we argue that it is also appropriate to view IR as a foundation of AI when the users of IR systems are intelligent agents. In fact, the important role that IR can play in supporting and improving LLMs can already be clearly seen from the rapid growth of the work on retrieval-augmented generation (RAG) [7, 11, 15, 17, 33, 39, 48]. In this paper, we go further to argue that with a broader definition of IR that includes many interesting non-traditional forms of retrieval problems, IR is *essential* to achieving AGI.

Specifically, we discuss how an intelligent agent must be able to perform information retrieval on both *external* information sources to access new information unseen by the agent or trace the provenance of information and *internal* information sources to perform complex tasks such as problem solving, planning, and reasoning. Intelligent agents also benefit from using IR to actively acquire the most useful new data and information for lifelong learning. Specifically, we identify five novel IR tasks that an intelligent agent must be able to perform in order to achieve Human-Level Artificial Intelligence, or Artificial General Intelligence (AGI), including 1) External Information Retrieval (EIR) to access new information unseen by the agent, 2) Provenance Information Retrieval (PIR) to trace the provenance of information, 3) Curriculum Information Retrieval (CIR) to actively acquire the most useful new data and information for lifelong learning, 4) Rule Information Retrieval (RIR) to perform reasoning and problem solving, and 5) Scenario Information Retrieval (SIR) to leverage past scenarios for problem solving and decision making. We analyze the similarity and difference of these new tasks and the normal retrieval tasks performed by an IR system that serves human users. We systematically examine the challenges that these new retrieval tasks may involve and outline a roadmap for new IR research from the perspective of IR for AGI.



SIGIR '25, Padua, Italy

ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3730349

2 Paths to Artificial General Intelligence and IR

Achieving Artificial General Intelligence (AGI) [21], roughly understood as human-level intelligence, is generally regarded as the ultimate goal of AI research [32]. An intellligent system that achieves AGI would be able to understand, learn, and apply knowledge across a wide range of tasks, much like a human. The breakthrough of LLMs accelerated research progress toward AGI, making the pursuit of AGI seemingly a much more feasible goal than it used to be [32]. Indeed, today, intelligent agents powered by LLMs (e.g., ChatGPT) have already demonstrated human-level proficiency in natural language understanding and generation [1].

However, the current intelligent agents are still far from like humans. Human brains are known to have two somewhat disconnected systems, called System 1 and System 2, respectively [19]. System 1 is fast and intuitive, but not reliable, whereas System 2 is slow and rational. It is generally believed that System 1 is more like a neural network system, while System 2 is based on some kind of symbolic representation with capacity of logic reasoning [19]. The current LLMs approximate System 1 well, thus the current agents powered by LLMs can be regarded as simulating System 1 of a human brain well. However, they do not have the capacity of System 2 [51]; how to incorporate System 2 into an LLM-based agent is a difficult open challenge; this is partly because we do not yet have a good understanding of how the System 2 works in a human brain [16]. In particular, it is unclear how the neural system of a human brain is able to represent symbolic information and perform logic reasoning.

The main progress in understanding System 2 is reflected in the development of multiple cognitive architectures, such as ACT-R [4] and SOAR [31]. These architectures are based on a rule-based production system stored in the brain. Both ACT-R and SOAR also use memory to represent the current state of a (human-like) agent, which would be matched with the conditions of all the rules in the production system. Any rule whose condition is matched would then be activated, causing the action specified in the rule to be taken, which would lead to the change of the state, thus potentially activating another rule. Note that the activation of a production rule can be regarded as performing a special retrieval task, i.e., we would use the current state as a "query" to match with many conditions in the production system and attempt to "retrieve" the best matching rule. When the rules are inference rules, a sequence of them may be activated to perform complex reasoning.

Recent work has attempted to integrate LLMs with such a cognitive architecture, including both ACT-R [49] and SOAR [50]. A heuristic approach called LLM Sandwich uses LLMs to both acquire rules from text data to form a symbolically represented knowledge base and exploit the constructed knowledge base to solve a problem by interacting with users [12]. Regardless of which strategy to be used, future intelligent agents that attempt to achieve AGI all must be able to simulate a production system in some way, and how an LLM can be integrated with a production system remains one of the most important challenges to be studied in the future.

Another major weakness of the current agents as compared with humans is that they are not as autonomous as humans, especially in terms of learning how to use a tool or even inventing a new tool for a particular purpose [40]. As a specific example, while RAG has been extensively studied with many algorithms for RAG proposed, there is generally much manual effort involved, and an intelligent agent cannot autonomously learn how to use a search engine to perform RAG, nor can they create a (new) search engine themselves. In the future, we may anticipate those intelligent agents would all have to learn how to effectively use a search engine [55] and even how to create their own search engines as needed. The current work on RAG has mostly focused on demonstrating the benefit of RAG for improving task performance, creating the impression that retrieval is only useful if it can indeed improve task performance. We note that there are two important reasons why an agent *must* use a retrieval system: 1) Because it is practically impossible to update an LLM every minute, using a search engine appears to be the only way for an LLM to get access to the most current information (e.g., the current price of a stock or a news article about an event that has just happened). Unless the agent does not need to know about such fresh information, it would have to use a search engine. This is also the reason why search engines will never disappear[55]. 2) An LLM must use a retrieval system to provide knowledge or information provenance. Unless an LLM can memorize all the original information, there would be no guarantee that the model would not hallucinate[1]. As long as there is any small chance of halluciation, it would cause potential concerns about the trustworthiness of an agent; the higher the stake of the application is, the more serious the concern would be. To address this concern, the agent must again use retrieval to provide information or knowledge provenance by retrieving the original support documents. This means that even after we train an LLM with a collection of documents, we still have to keep those documents and maintain a search engine to serve an LLM for this purpose. However, this kind of search engine would need to serve an intelligent agent as its user, and thus can be potentially quite different from those that serve human users today due to the difference between an intelligent agent and a human user.

The current agents are also unable to self-improve. For example, if we are to continue the training of any LLM, we would have to collect the data and set up the training process; unlike humans, LLMs lack the ability to actively search for useful new data to train themselves continuously. If we want an agent to do the same, the agent would need to have the capacity to find the most useful new data to continue training the LLM, which can be regarded as a retrieval problem where the ranking of documents may be based on how useful a document would be for training an LLM.

Finally, humans are known to use case-based Reasoning to solve a problem [44], where we would rely on our past experiences stored in our episode memory to assist in solving a new problem. To enable an intelligent agent to do the same, the agent would need to have memory to store the past scenarios in association with the reward information about what plans succeeded and what plans failed. Those past scenarios would have to be represented in some way to facilitate retrieval of scenarios similar to a current scenario and

[1]Note that hallucination is a consequence of the LLM's generalization capacity and thus desirable in many applications such as creative writing. The solution to the problem of hallucination is not to inhibit an LLM's capacity of generalization, but rather use a System 2-like module to regulate the LLM's behavior; such a regulation mechanism would allow hallucination in contexts like writing a poem, but inhibit it in contexts like answering a medical question.

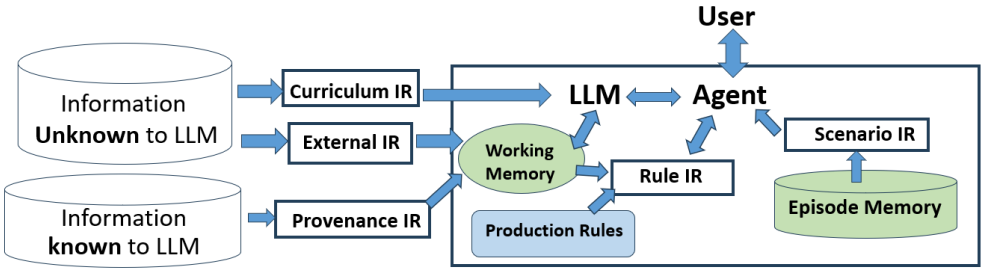


Figure 1: IR as Foundation for AI Agent with five IR tasks: External IR, Provenance IR, Curriculum IR, Rule IR, and Scenario IR.

optimize the decisions for the current scenario. In this situation, we have a decision-drive scenario information retrieval problem.

To summarize the discussion above, a truly intelligent agent that attempts to achieve AGI would have to rely on IR in at least five different scenarios, leading to five novel retrieval tasks as illustrated in Figure 1: 1) It needs IR to have access to the external information that the agent does not know (External Information Retrieval) , 2) It needs IR to provide knowledge provenance (Provenance Information Retrieval), 3) It needs IR to activate production rules (Rule Information Retrieval), 4) It needs IR to acquire new data for continued training (Curriculum Information Retrieval), and 5) It needs IR to retrieve past experience to facilitate problem solving in a current scenario (Scenario Information Retrieval).

It is from this perspective that we argue that IR can also be viewed as a foundation for AI when it serves intelligent agents, instead of humans, as users. This perspective is in contrast with the common view that AI provides a foundation for IR, which has led to the main stream of IR research influenced by the LLMs so far. The five new IR problems we identified have not yet been well-studied. A main goal of this paper is to systematically discuss all of them, examine to what extent the existing retrieval techniques can be applied to solve those new retrieval tasks, and discuss what new challenges we have to address in future research.

3 Information Retrieval for AGI

In this section, we provide a more detailed explanation of all the five novel retrieval tasks in the context of serving AGI, highlighting any similarities and differences from the traditional retrieval tasks where humans are the users.

3.1 External Information Retrieval (EIR)

As it is practically infeasible to feed an LLM instantly with all the real-time information, an LLM alone cannot answer any question about a current event (e.g., the current price of a stock or an event that has just happened), thus an agent must use a search engine to get access to such new information [54]; in general, an agent would benefit from using an IR system to manage a large amount of external information that its LLM has not yet been able to absorb. The retrieval problem in this context is quite similar to a traditional retrieval problem and thus can be solved by using an ordinary IR system that currently serves human users. However, since an AI agent is fundamentally different from a human user, a current system optimized for serving human users is not necessarily optimized to serve an AI user. For one thing, the Probability Ranking Principle [42] has long been the theoretical foundation of the current search engines, but it was based on the assumption of sequential browsing by human users. An AI agent clearly does not have to do "sequential browsing," making it unclear how we can establish a theoretical framework to optimize an IR system for serving an AI user. In this sense, this is a novel IR problem, which we call External Information Retrieval (EIR) since the goal is to enable the AI user to have access to information external to the agent.

In a broader context, the reliance of an agent on a retrieval system is a special case of its reliance on many other computational tools such as a database system or data mining tool. In general, a truly intelligent agent must be able to learn how to use such a tool and combine different tools to solve a complex problem [40].

Given the availability of a search engine, an LLM-powered agent should ideally optimize its training objective so that it would optimize its collaboration with a search engine. For example, it should know that if the information can be easily served by the search engine, which costs less than an LLM, the LLM would not need to be able to generate it (from scratch). In general, an agent may take advantage of a search engine as a more cost-effective tool to help manage all kinds of factual information and data by simply having the LLM write out the factual information or data known to the LLM to create a collection of documents. This way, the LLM would no longer need to use many parameters to memorize all the information (e.g., weather data or sports game statistics); instead, it can use the extra parameters to model information or knowledge that cannot be retrieved by the search engine. In other words, the LLM can avoid duplicated work that a search engine can already do. Formally, this means that we need to optimize a loss function with consideration of using a search engine to make a prediction (e.g., by using a reinforcement learning framework).

3.2 Provenance Information Retrieval (PIR)

The hallucination problem of LLMs [45] is a major barrier in using LLMs to directly interact with end users since an error might potentially have a very high cost. To address this problem, the agent must provide knowledge provenance by citing the sources that support its generated response, enabling the user to verify the information as needed. This problem is a special retrieval problem, where the collection contains all the documents that have been used to train the LLM and the goal is to retrieve the documents that can support a generated response by LLM. We call this retrieval problem Provenance Information Retrieval (PIR). PIR is an interesting new IR problem for two reasons: 1) The notion of query is not clear in PIR, and a main challenge is how to formulate appropriate queries based on a generated response. 2) All the documents in the collection were seen by the LLM during training, thus we may not need a

separate IR system; instead, the LLM may be trained to memorize all the IDs of relevant documents and thus can generate the IDs directly for provenance IR.

3.3 Curriculum Information Retrieval (CIR)

Curriculum Information Retrieval (CIR) refers to the problem of retrieving the most useful new content from a collection of candidate documents for continuing the training of an LLM so that the agent can continuously learn over time. For example, the whole Web can be the collection, and the target information to be retrieved to further train an LLM. Here again, a main challenge is how to formulate the query as the information need is complicated. The notion of relevance is also complicated since it is unclear how we can quantify the benefit for learning. Ideally, we can have a sequence of documents in ascending order of difficulty so that the model can first learn from easier ones before proceeding to learn from difficult ones. This kind of ordering is intuitively beneficial, which is the idea of Curriculum Learning in ML [47]. We thus call this task curriculum information retrieval (CIR).

3.4 Rule Information Retrieval (RIR)

While the retrieval problems discussed so far involve an external collection of documents, the next two retrieval problems, Rule IR and Scenario IR, both involve an internal collection of information inside an agent. We first introduce the problem of Rule Information Retrieval (RIR), where the collection is the set of production rules in a production system of an agent (assuming that future agents need to have such a system to implement System 2 of a human brain). The query is the current memory state of an agent. We would like to rank all the production rules based on how well the condition of each rule matches with the current memory state. Compared with a standard retrieval, the items to be ranked are quite different, while the matching can be much more complicated, depending on how information is represented in the memory and how conditions of rules are specified.

3.5 Scenario Information Retrieval (SIR)

Next, Scenario Information Retrieval (SIR) is another retrieval problem that involves retrieving information inside an agent. It refers to the problem of retrieving relevant past scenarios from the episode memory of the agent based on how well they match with a current scenario. The relevant scenarios retrieved would be used to guide the agent in choosing the actions in the current scenario. This enables the agent to learn from the past experience to optimize a decision or solve a problem in the current scenario, thus we can also view the problem of Scenario Information Retrieval as decision-driven or problem-solving Information Retrieval. As a scenario is not a well-defined concept, a major challenge in SIR lies in how to represent scenarios and match a query scenario with the stored scenarios.

4 A New Perspective on IR Research

For over seven decades, IR research has focused on IR systems that serve humans as users [28]. We propose a new perspective on IR research, where the focus is on IR systems that serve intelligent agents as users (i.e., AI users). Although there are still many difficult challenges to be addressed in the traditional perspective of IR serving human users, the current IR systems are already sufficiently mature and serve many people well in the world every day. Moreover, the emergence of powerful LLMs has now made it possible to answer a user's question directly, further reducing any pain point in information access. In contrast, there are many interesting new difficult challenges to be tackled in IR serving AI agents. As more progress is made toward AGI over time, we can expect to see even more new challenges that the IR community can study in using IR to serve AI agents. We thus believe that by taking this new perspective, IR research may enter a new era with numerous interesting new research challenges to be studied in the general context of IR for AGI.

As human users and AI users are fundamentally different, the requirements for an IR system to serve them also differ significantly, leading to different formulations of IR problems that naturally require new algorithms and new evaluation methods. In addition, when an IR system serves an agent system, it can go beyond serving the agent with an external document collection to further provide retrieval support inside an agent, to enable, e.g., case-based reasoning or execution of complex production rules, where the source of information is internal to an agent; such retrieval scenarios are quite different from the usual retrieval scenarios where the goal is to satisfy a user's information need. Below, we systematically examine some important new research challenges associated with the five new IR problems in the context of IR for AGI.

4.1 Formalization of New Retrieval Problems

We first discuss the challenges in formalizing each of those five new retrieval tasks.

External Information Retrieval (EIR): The EIR problem generally occurs when the agent needs to access external information to finish a task (e.g., answer a user's question). While the problem is quite similar to a traditional retrieval problem in the sense that the collection of documents is exactly the same, there are at least two challenges in defining the problem. The first is how to define the query as it would depend on the task. In the case of answering a user's question, the user's question can be naturally treated as the query, but in many other cases, we might have to take a user's task description as a "query"; needless to say, in such a case, the retrieval problem would be much harder since there will be a gap between the task description and an effective query that can work well for retrieving relevant information. Sometimes, it may also be necessary to use multiple queries to finish a task, resulting in further complications. The second challenge here is to define the ideal retrieval results. While a ranked list with snippets is optimal according to the Probability Ranking Principle when serving human users, it is unlikely optimal when serving an AI agent since the sequential browsing assumption used in PRP [42] no longer holds. The current work on Retrieval-Augmented Generation (RAG) has increased our understanding of how to tackle this challenge [11, 15, 17, 39, 48]. For example, one of the most effective RAG techniques, KAG [35], introduced the idea of generating "LLM-friendly" knowledge graphs by using retrieved documents and providing multi-layer representation of the retrieval results. However, it remains unclear what is an ideal "retrieval result" for an agent and to what extent such an ideal result depends on the specific tasks.

Provenance Information Retrieval (PIR): The PIR task is generally required in almost all the scenarios when an agent generates

a response; since a user may ask for the source information at any time to verify the response generated by the agent, the agent must be "prepared" to do PIR at any time. As a retrieval problem, the collection used by PIR is the same as in standard retrieval, which is often the collection of all the training documents used for training an LLM and thus can be potentially very large such as the whole Web. However, it is unclear what a query is in this scenario. If a user asks for sources about a specific statement, then that statement can be used naturally as a query, but if the user does not explicitly ask about any specific statement but instead has doubt on the reliability of the whole response, the agent would have to figure out which part of the response would benefit from citing a reference to back up the claim. In such a case, the problem seems to be similar to citation recommendation [18].

An important difference between EIR and PIR is that the information to be retrieved in PIR is *known* to the agent whereas in EIR, it is generally unknown, which is why the agent needs to retrieve it in the first place. This difference has implications on how to solve the retrieval problem. For example, in PIR, since the agent (LLM) has already seen the information, it might be possible that the LLM itself can remember the document ID associated with the claim and thus can directly *generate* the document IDs that can support a claim without requiring it to use a traditional IR system. However, it is possible that the LLM's memory of the document IDs is inaccurate, in which case, it might generate a wrong ID (e.g., hallucination) and thus would not be able to actually retrieve the support evidence. This is a major challenge that needs to be addressed since not only the agent needs to provide a reference, but the reference must also be actually available to the user for verification (a faked document title or ID would thus not work).

A further challenge in PIR is what should the retrieval results look like. A simple solution of ranking the documents would not be optimal since ideally, we can locate which passages support a claim. A more complicated case is when the claim must be supported by combining information from multiple documents, in which case, the retrieval algorithm would need to model *collective relevance*, which goes beyond scoring each document independently and may require searching through a potentially very large space of combinations of multiple documents. While such a challenge of collective relevance also exists in the case when a retrieval system serves human users, it is more crucial in PIR as otherwise the user would not be able to trust the response from an agent.

Curriculum Information Retrieval (CIR): CIR occurs when an agent would like to actively find new data (e.g., from the Web) to allow its LLM to acquire new knowledge or fine-tune itself for a particular domain or a certain kind of tasks. While the current agents and LLMs cannot yet self-improve themselves, it is reasonable to anticipate future agents and LLMs would be able to do so. In that case, they would have to perform CIR. The collection of CIR can be any available information space, e.g., the whole Web, and the target information to be retrieved can also be informally defined as the best documents to allow an LLM to acquire the most useful new knowledge, but there are many challenges to be addressed in order to formalize this retrieval problem:

First, given a collection of candidate documents, what should be the target documents to be retrieved and used for training? From the perspective of Curriculum Learning [47], ideally, it is a set of documents ordered in some way to optimize the learning efficiency of an LLM (e.g., enabling the LLM to first be exposed to relatively easy documents before being exposed to more challenging documents). If we view those documents that are most useful for training as "relevant documents," the notion of relevance here is very complicated since it cannot be assessed independently by judging each individual document as relevant or not, or assessing the usefulness of each document, but instead, the whole list needs to be considered. That is, as a retrieval problem, we will perhaps need to use a listwise learning to rank approach to solve it; pairwise or pointwise would be inherently non-optimal [36].

Second, how much redundancy should we allow in the retrieved documents? When serving users, we know that we generally want to eliminate redundancy, but for training LLMs, it may be beneficial to have some level of redundancy; indeed, if there is no redundancy, an LLM may not be able to learn the underlying semantic patterns effectively. However, if there is too much redundancy, it might also cause the LLMs to be exposed to information that it has already seen many times, thus not learning from any additional exposure. However, due to interference of optimizing multiple components in the loss function of an LLM (e.g., predicting the next word and answering a question), it might also be beneficial to allow the LLM to periodically "review" some known content that the model might have "forgot" due to additional training. This can be achieved by including some new content similar to the "forgotten content" in the CIR results to re-expose the LLM to such content. For example, such "reviewing" can encourage adjusting the gradients toward minimizing the errors on some words, whose errors had been minimized earlier but increased later due to tradeoff between errors on different words. From retrieval perspective, even if we use listwise approaches [36], it is still quite challenging to define and model the optimality of a list. Note that this challenge is related to some work done in the IR community on search and learning (see, e.g., [14]), where the goal is to scaffold learning for human users.

Third, a simplified view of CIR is to define it as an online single document retrieval tast, i.e., we can define the problem as retrieving a single best document as the document to learn from at any time. Using this formulation, the problem of CIR can be solved by iteratively retrieving the "best next document" for an LLM to use for learning. The advantage of this formulation is that the search space would be significantly smaller and the problem would be closer to a traditional retrieval problem. With this formulation, redundancy can be addressed by using a greedy algorithm such as Maximal Marginal Relevance (MMR) [10]. However, by retrieving a single document, we essentially use a greedy algorithm to optimize a whole list, which may not be optimal if we makes a mistake early on when choosing an initial document. Thus there is an inevitable tradeoff between optimality and efficiency. This can be addressed by treating each single document retrieval step as an action taken by an agent and using reinforcement learning to learn an optimal policy for choosing the best next document based on a reward designed on a whole list of documents used for training.

Fourth, to optimize the learning efficiency, cohesion of the document list may also be important. That is, ideally, the documents that an LLM would see should be presented in an order to ensure some level of cohesion so that we can minimize the fluctuation of the gradients during the training process. Furthermore, the knowledge

gap between what an LLM has already known and a new training document should also be considered; ideally there is not a big gap and there is some kind of cohesion as the model attempts to digest the new document in the sense that it can invoke some closely related old knowledge. This kind of cohesion may be used to determine which documents are "easier" and which are "harder" for a particular model. How to quantify this is a technical challenge, though a simple solution may assume that a text that can be predicted accurately by an LLM to be easier than one that cannot be predicted well [52].

Finally, since the benefit of additional training is to increase an LLM's performance on future tasks, CIR would also need to favor content similar to the content of the anticipated future tasks; this way, the agent can maximize the utility of additional training and avoid being trained on content that it might never see in the future. Thus optimal CIR would also require an estimate of the future tasks and domains that the LLM is expected to operate on, which would naturally be easier in the case of training a domain-specific foundation model; for example, favoring content with high educational value has been shown to be beneficial for developing a foundation model in astronomy [37]. For this reason, we can anticipate future real world applications to use increasingly more specialized foundation models customized to a specific domain and a specific type of tasks; such specialized foundation models would be much more cost-effective than a general foundation model and much easier to maintain/improve over time using CIR.

In sum, it appears that the CIR problem is quite complex and there are multiple objectives to be optimized in CIR, related to many other problems such as active machine learning [46], curriculum learning [47], semantic cohesion, redundancy, ranking algorithms, and utility optimization over a horizon.

Rule Information Retrieval (RIR): The RIR problem is relevant to the implementation of System 2 of the human brain in an intelligent agent. While the current agents cannot yet have this kind of neurosymbolic architecture, it is an actively researched area [25]. It is reasonable to anticipate the future agents to implement cognitive architectures that are known to reflect how humans solve problems (e.g., ACT-R [4] or SOAR [31]), where the basic mechanism is a production system with conditioned rules. A rule would be activated when the condition of the rule is matching well with the current state of the agent as reflected in its memory. The RIR problem can thus be formulated as the problem of scoring all the production rules based on how well their conditions match with the current state and then "retrieving" the matching rules for execution.

As a retrieval problem, RIR is quite different from a regular retrieval problem in that the collection here is a collection of production rules. It is also different in that the query may be quite complicated as it is some form of representation in the memory. Although we do not yet have a good understanding of how information is stored in the memory exactly, we may still view this as an interesting novel information retrieval problem since the goal here is to retrieve the best rules for executation. Perhaps the most challenging aspect of this problem is the notion of relevance, i.e., how we should match the current state with the condition. Because the information we are dealing with here may not be even interpretable (e.g., the memory may be represented as embedding vectors), any traditional matching techniques based on discrete representation are unlikely applicable, but neural ranking models [24] may be appropriate after being trained with a sample of memory states and the corresponding "relevant" conditions. In general, RIR can only be studied in the context of a concrete production rule system.

Scenario Information Retrieval (SIR): The SIR problem occurs when the agent uses case-based reasoning to solve a problem in a human-like manner [44**?**]. The collection here is the past scenarios (cases) stored in the episode memory of the agent. The query is a current scenario (case). The retrieval task is to retrieve a set of best matching scenarios from the memory, which could then be used to figure out what to do in the current case, i.e., optimize the decisions when solving a problem in the current scenario. In general, there is also reward information associated with the past scenarios (the agent would assign reward to those scenarios to distinguish the scenarios that it likes from those that it does not like). From retrieval perspective, it means that the information items in the collection are annotated with reward information and the retrieval algorithm is supposed to consider such reward information. Specifically, the agent would prefer retrieving similar scenarios that have the highest reward. Thus the notion of relevance includes two aspects: 1) How similar is a past scenario to the current scenario? 2) How high was the reward associated with a past scenario? How to optimize the combination of these two factors would be a challenge that needs to be studied. Moreover, an additional challenge is the number of relevant scenarios to be retrieved. This challenge is essentially similar to the problem of optimizing the number of neighbors in the K-Nearest Neighbor classifier.

As such a retrieval problem is primarily driven by decision making, i.e., choosing an optimal action in the current scenario, an alternative way to frame this retrieval problem can be to retrieve "actions" directly from the past scenarios. This would require a somewhat different representation of the episode memory and lead to a problem setup more similar to the RIR problem. Specifically, if the past scenarios are organized based on actions, we can view the scenarios as conditions, forming many production rules. Then the SIR problem may also be viewed as matching the current scenario, which is in the current memory, with the conditions of the production rules. The matched rules with high reward can then be retrieved for execution.

Note that when using reinforcement learning to enable an agent to self-improve itself [20], we often store a sequence of tuples in the form of (state, action, reward), and the SIR problem is generally relevant as a way to facilitate policy learning. A key challenge from retrieval perspective is how to compute the similarity of two scenarios especially because a scenario might be quite complicated with a high-dimensional space representation.

4.2 Evaluation of New Retrieval Problems

Another general challenge in studying these new retrieval problems is how to evaluate them. Evaluation of IR systems that serve human users has traditionally been done using human annotators to make relevance judgments and create reusable test collections [43], or more generally using user simulation [9]. When the user of an IR system is an AI agent, how should we evaluate it? Intuitively, we should estimate the utility of the retrieved results from the perspective of the agent. However, the current agents and LLMs are mostly a blackbox, so their behaviors are somewhat unpredictable; they

might also change their behaviors after being fine-tuned with any additional data. This means that there will be a new kind of uncertainty associated with any judgments that an agent can provide about retrieval results. Handling this uncertainty and interpreting an AI agent's judgment appropriately would be a main challenge in evaluating all those new retrieval tasks.

However, since an agent usually performs a task, we can often use the agent's task performance as a way to *indirectly* evaluate any retrieval results. The idea here is to vary any retrieval system that supports an agent, and check its impact on the task performance. This evaluation methodology is the one used frequently in the current work on RAG and can be a general evaluation methodology for these new retrieval tasks.

If we frame the EIR, PIR and CIR problems as a conventional document retrieval problem (note that they do not have to be framed that way due to an agent's ability to digest other forms of results), we would be able to use the traditional Cranfield evaluation methodology [43] to evaluate them. However, while the relevance judgments for EIR and PIR can be made by human assessors, it would likely be inappropriate to for humans to make relevance judgments for CIR. Fortunately, we can directly evaluate the quality of a retrieved (ordered) list of documents in CIR by actually training (fine-tuning) an LLM with the retrieved documents and checking whether training would improve the performance of the LLM on various tasks. This evaluation method enables us to immediately make progress in studying CIR algorithms. However, such an evaluation strategy is computationally expensive, especially if we want to experiment with many variations of the retrieved documents and their orders, thus how to efficiently evaluate CIR is a major challenge. One solution is to first experiment with small models to explore a large space of CIR algorithms efficiently and then verify any findings on much larger models as suggested in the work [52].

The RIR and SIR problems are envisioned to be retrieval problems more relevant to the future human-like agents and LLMs. However, they can also be studied now in various application contexts. For example, SIR can be studied if we can collect interaction data from an agent that operates in a task environment (e.g., a web agent that helps a user finish a task), in which case, we would have a collection of the scenarios. We can use some collected scenarios to simulate "future scenarios" and evaluate SIR to check the optimality of the retrieval algorithm based on the observed reward or task performance of the agent. RIR can also be studied if we add rules to an agent for regulating their behaviors. Some existing work on integrating cognitive architectures such as ACT-R and SOAR with LLMs may be leveraged to evaluate RIR.

4.3 Retrieval Algorithms

The retrieval algorithms needed for the new IR problems generally vary according to the specific retrieval task, opening up many interesting new research directions on IR models and algorithms. There are several factors that would affect what kind of algorithms we would need for each retrieval problem. The first is how the agent would interact with the retrieval system. Since an agent does not have to "sequentially browse" a ranked list as human users have to do, it may benefit from a different kind of retrieval results being generated by the retrieval system than a regular ranked list. Another factor is the notion of the query. A simple general solution is to prompt LLMs to generate a keyword query. Such a query would then allow us to use a conventional IR algorithm to process the query and generate conventional retrieval results. However, forcing the agent to use a keyword query likely is too restrictive as the agent can interact with an IR system in a conversational manner (e.g., combining results from multiple queries) or even interact in a shared embedding representation space.

However, existing retrieval algorithms are still expected to be useful for some of the new problems, particularly, EIR, PIR and CIR, since they all involve a conventional document collection. For example, traditional retrieval algorithms can be used as a supporting component to enable further processing of the retrieved results (e.g., KAG [35], as a special case of EIR, would further construct a knowledge graph based on the retrieved documents) It is also natural to use a generative information model [34] for PIR, where the model could directly generate relevant document IDs. An interesting general challenge here is how an agent would optimize its collaboration with a supporting retrieval system. Since for any information need, there tends to exist an ideal query that has the "right" keywords to retrieve relevant documents [30], we can have the LLM do more in terms of query formulation so that it would be able to generate an effective query close to the ideal query; this would alleviate the burden on the IR system and a basic keyword matching retrieval system would suffice.

The queries and collections in RIR and SIR are both quite different from those in a conventional IR problem, thus new algorithms must be developed. Note that the items retrieved in RIR and SIR may be represented as non-interpretable embedding vectors inside a future AI agent, thus a neural ranking algorithm [24] may be expected to be more effective than a traditional keyword-based model.

4.4 Feedback

Feedback from human users, especially implicit feedback via clickthroughs [13], is known to be highly useful for improving an IR system that serves human users. It can also be exploited to improve IR systems that serve AI agents; indeed, as an agent can understand natural languages well, a user can provide even more informative feedback by directly expressing it in a natural language, which also enables reinforcement learning with human feedback (RLHF) for improving an intelligent agent in general [8].

For example, in EIR and PIR, a user's feedback while interacting with the retrieved results can be directly passed to the underlying IR system that assists the agent. Conversations with a user can help clarify the user's information need, enabling the agent to formulate a more effective query. The agent can further collaborate with a retrieval system to perform pseudo-relevance feedback to refine a query. The current work on RAG has already explored this kind of ideas [26]. A general opportunity here is for the agent and the retrieval system to collaborate on constructing an ideal query for a retrieval problem. For example, a very quick initial retrieval of just a few documents can be done first, and an LLM can then analyze the initial results to suggest revision of the query as needed; the process can be repeated until the results are satisfactory.

A more general challenge is how to optimize the collaboration of an agent and a retrieval system in terms of both effectiveness and efficiency. While there has already been work done in the context of many variants of RAG, we are far from having any theoretically

motivated optimal retrieval model presumably due to the challenge in defining an "ideal retrieval result" for an agent, which varies according to different tasks. However, since all the retrieval problems require some kind of ranking, learning to rank techniques can generally be expected to be useful for optimizing the ranking results of all the new retrieval tasks [36]. Also, dense retrieval techniques [24] may have greater potential for adapting to some of the new retrieval problems discussed in this paper than the classic keyword-matching retrieval models due to the possibility of further training them specifically for a somewhat different relevance criterion. When we use dense retrieval models, it is further possible that the retrieval system can be more tightly integrated with the LLM of the agent (e.g., via mapping of their embedding representations).

4.5 Indexing and Retrieval Systems

Finally, there will be potentially new challenges in implementing a retrieval system for all those new retrieval tasks. EIR, PIR, and CIR all involve very large collections, thus efficient indexing is clearly required, and the current indexing methods can provide a baseline solution especially if a keyword query is available. However, we cannot assume that we always have a keyword query available and may need to rely on dense vector representation for retrieving a document, in which case, we would need to consider using more efficient dense index methods (e.g., [56]).

As RIR and SIR are new retrieval problems that have not yet been well-studied, how to create an index to enable fast matching is in general an open challenge. This challenge can be difficult due to the complexity of the scenario and memory representation. Before RIR and SIR can be appropriately studied, we may need to understand how a human-like agent stores information in its working memory and episode memory. Nevertheless, both retrieval tasks will be needed in the future in order to build human-like intelligent agents.

5 Discussion

A main point we have made is that IR can be viewed as both an application layer on top of AI (when serving human users) and a foundation layer below AI (when serving intelligent agents). Their mutual support relationship suggests a natural path forward in which IR and AI would be increasingly integrated. In fact, progress in the research of all five identified retrieval problems would benefit from the interdisciplinary collaboration of researchers from IR and AI. In the following, we discuss some specific interdisciplinary research issues in the context of the five IR tasks.

RAG = EIR + PIR: In the current work on RAG, the collection used for retrieval may include both new information to which an LLM has not been exposed and information already known to the LLM during the training period without distinguishing them. However, we argue that it is important to distinguish EIR from PIR because they are, in nature, different problems, and thus require different approaches. In EIR, because the LLM has not seen the information before, it must rely on a separate (external) search engine to retrieve relevant information. Thus, it is a special case that an LLM agent learns to use a software tool, often called tool learning [40, 41]. A main research question in EIR is how to optimize the collaboration of an LLM and a search engine, in particular, how to divide the retrieval task between the two. At one extreme ("heavy search"), an LLM can directly pass a user's question as a query to a search engine, thus placing most of the burden on the search engine (e.g., to fully understand the question); at the other extreme ("light search"), an LLM can do as much as it can to frame one or multiple effective queries, to be executed by the search engine, in which case, a basic keyword matching search engine may suffice. Given that an LLM is more capable of understanding natural language, "light search" appears to make more sense than "heavy search", and is thus generally preferred. We may even go further to reduce the work on the side of a search engine by having many specialized (vertical) search engines, each being focused on searching over a certain type of information. The LLM can then learn how to use all these specialized search engines as different tools, i.e., it can learn when to query which engine and how to combine results from multiple search engines. It is also possible to further integrate the LLM with a search engine (tool) by allowing them to potentially communicate more efficiently in their respective spaces of embedding representations bypassing the use of a human-interpretable query.

The optimal solution to PIR is completely different from the solution to EIR because in PIR, the information is known to the LLM, which means that we do not necessarily need a separate search engine. Instead, the source information (e.g., document IDs) can be encoded with special tokens during the pre-training stage, thus enabling the LLM to remember where it has "seen" particular information just as humans can often recall roughly which part of a book has mentioned a particular point. This means that while PIR is in nature a retrieval problem, it can be potentially solved via associative memory in an LLM. However, we still need to actually retrieve any relevant evidence information according to the "remembered source" by the LLM to verify a response generated by an LLM (e.g., the LLM needs to show to its user the actual evidence). It is also worth pointing out that in PIR, the amount of information to be retrieved is often restricted to a small amount of directly relevant information in support of a response (known item search). In contrast, in EIR, the amount of information to be retrieved can be potentially very large depending on the information need by the LLM and the human user it serves.

CIR vs. Active Learning and Curriculum Learning: The general goal of CIR in optimizing the data to be used for training a machine learning algorithm is the same as that of active learning [46] and curriculum learning [47], but CIR provides a different and complementary perspective than either active learning or curriculum learning. In active learning, the specific problem is to choose unlabeled data instances for humans to label so that the labeled data would be most useful for (further) training a machine learning system, thus the existing research in active learning can inform how to design an effective algorithm for CIR. For example, a commonly used strategy in active learning is to choose the data instances that an ML program is least certain about for prediction to obtain human annotations. This heuristic is directly applicable to CIR in that we can favor those data that our LLM is not confident about, meaning the data that have high perplexity according to the current LLM. However, CIR emphasizes modeling the collective value of a set of instances and thus requires modeling of their dependency and interactions, where as in active learning, each candidate data point tends to be evaluated independently. In curriculum learning, the optimization of the ordering of a set of already selected data instances is the focus, where the general goal is to expose a learning

algorithm to easier cases first before hard cases, thus intuitively scaffolding learning. CIR provides a conrete way to support curriculum learning by framing the problem of ordering cases as a retrieval problem, thus opening up opportunities to apply many ranking algorithms studied in IR to curriculum learning. Moreover, CIR also covers the challenge of selecting the set of training data instances in the first places, whereas curriculum learning assumes that a training dataset is already given.

RIR and SIR: Complex Queries and Reasoning: RIR and SIR are both retrieval problems involving internal information inside an intelligent agent. Because we do not yet have a good understanding of the exact architecture of a humna-like agent, the formulation of these problems faces inevitable uncertainties. However, we can be sure that in both RIR and SIR, the notion of a query is likely complicated as the query refers to some kind of representation in the memory of an agent. In this sense, we might also view both as special cases of a more general task, which may be called Memory Information Retrieval (MIR). Despite the similarity, it is important to distinguish RIR from SIR for two reasons: 1) The retrieved object is different in nature: In RIR, we attempt to retrieve a production rule [27], which is symbolic and used in human cognitive processes [4], whereas in SIR, it is a past scenario, which is non-symbolic and can be regarded as a state representation that drives decision making. 2) RIR is needed for reasoning, planning and problem solving in general, where a sequence of rules can be repeatedly applied to any initial condition (initial state), as many times as needed, to derive a conclusion or reach a goal state for solving a problem [27]. RIR can facilitate both regulation and explanation of the behavior of an agent, crucial to increasing the trustworthiness of an agent. SIR is needed for supporting case-based reasoning [29], where the retrieved past scenarios can be used to assess the optimality of a candidate option in decision making and support reinforcement learning with a policy to favor actions associated with positive past cases [5]. We thus may view SIR as "state retrieval," where a current state is used to retrieve similar past states to guide the agent to choose an optimal action in the current state. Note that a unique characteristic of SIR is that the "query" and "document" are symmetric in that they are both scenarios (cases), whereas they are usually asymmetric in other retrieval problems. Thus we can also expect the effective algorithms for SIR to be different from those for other retrieval tasks.

6 Related Work

The perspective suggested in this paper is inspired partly by the rapid growth of work on Retrieval-Augmented Generation (RAG) [33] in multiple research communities, including Information Retrieval [39, 48], Natural Language Processing [15], Computer Vision [7], Data Mining [17], and Artificial Intelligence [11]. As an interdisciplinary research topic, RAG explicitly ties IR with many other fields in AI. The significant benefits from using IR to enhance LLMs in a wide spectrum of tasks demonstrated the important role that IR plays in AI. However, most existing work on RAG was motivated by the need to use external information to help address the problem of hallucination and improve quality of results generated by LLMs in general. This paper goes beyond the current work on RAG by making a much stronger argument that IR not only enhances agents powered by LLMs, but also enables the agents to approach Artificial General Intelligence (AGI) [32]. Our argument is partly based on the observation made in the previous work [54] that search engines will never disappear and future agents would need to learn how to use a search engine. We go beyond this point to further argue about the many different ways for an IR system to support an AI Agent, especially in the context of enabling the AI agent to perform case-based reasoning and incorporate a symbolic cognitive architecture such as ACT-R [4] or SOAR [31]. Our perspective is also partly inspired by research findings about how human brain works, particularly the dual-process and dual-system theory [19], and how biological intelligence might have been acquired via algorithms similar to reinforcement learning [22].

7 Conclusion and Outlook

We presented a new perspective on IR research, focusing on IR systems serving intelligent agents as users, and identified five novel retrieval tasks in this perspective, including 1) External Information Retrieval (EIR), 2) Provenance Information Retrieval (PIR), 3) Curriculum Information Retrieval (CIR), 4) Rule Information Retrieval (RIR), and 5) Scenario Information Retrieval (SIR). We discussed the similarity and difference of these new tasks and the normal retrieval tasks performed by an IR system that serves human users and systematically examined the challenges that these new retrieval tasks may involve, providing a roadmap for new IR research in the broad context of IR for AGI.

The new perspective that we presented raised an interesting question about the future of the IR research community. As we anticipate future IR systems to evolve into personalized intelligent task agents [54], it appears that the boundary between the IR field and AI would become less clear in the future, especially because IR techniques would be needed to support all kinds of agents. It is thus possible that we will see that IR and AI are increasingly integrated in the future. Considering that IR is also needed to achieve AGI, it is unclear whether there will be any meaningful boundary, nor do we want to have any meaningful boundary between IR and AI. Indeed, research work on RAG has already been published in all kinds of venues in multiple communities. Could it be the case that in the future, most IR research could also be published in all those venues, especially if we take the new perspective presented in this paper? As we expect the traditional view of IR (serving human users) to become increasingly mature, should the IR community also significantly broaden the scope of IR research to include both IR to serve human users and IR to serve intelligent agents?

8 Acknowledgments

This work is supported in part by the National Science Foundation (NSF) and the Institute of Education Sciences (IES) under Grant DRL-2229612, by the National Institutes of Health (NIH) under Grant 1R01LM014250-01A1, and by IIDAI at UIUC. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the NSF, IES, or NIH. The author is grateful to the three anonymous reviewers for their useful comments on an initial draft and to Ke Yang, Ekaterina Gribkova, Rhanor Gillette, Eric Modesitt, Kevin Ros, Chenkai Sun, Adam Davies, Bhavya, Chad Lane, Blair Lehman, and Diego Zapata-Rivera, for the useful discussion of some topics related to the perspective presented in this paper.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Qingyao Ai, Ting Bai, Zhao Cao, Yi Chang, Jiawei Chen, Zhumin Chen, Zhiyong Cheng, Shoubin Dong, Zhicheng Dou, Fuli Feng, et al. 2023. Information retrieval meets large language models: a strategic report from chinese ir community. *AI Open* 4 (2023), 80–90.

[3] James Allan, Eunsol Choi, Daniel P Lopresti, and Hamed Zamani. 2024. Future of Information Retrieval Research in the Age of Generative AI. *arXiv preprint arXiv:2412.02043* (2024).

[4] John R Anderson. 2014. *Rules of the mind*. Psychology Press.

[5] Mattia Atzeni, Shehzaad Dhuliawala, Keerthiram Murugesan, and Mrinmaya Sachan. 2021. Case-based reasoning for better generalization in textual reinforcement learning. *arXiv preprint arXiv:2110.08470* (2021).

[6] Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. 2025. Foundation Models Defining a New Era in Vision: a Survey and Outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025).

[7] Sukanya Bag, Ayushman Gupta, Rajat Kaushik, and Chirag Jain. 2024. RAG Beyond Text: Enhancing Image Retrieval in RAG Systems. In *2024 International Conference on Electrical, Computer and Energy Technologies (ICECET*. IEEE, 1–6.

[8] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862* (2022).

[9] Krisztian Balog and ChengXiang Zhai. 2024. User Simulation for Evaluating Information Access Systems. *Foundations and Trends in Information Retrieval* (2024). forthcoming.

[10] Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. 335–336.

[11] Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 17754–17762.

[12] Jennifer Chu-Carroll, Andrew Beck, Greg Burnham, David OS Melville, David Nachman, A Erdem Özcan, and David Ferrucci. 2024. Beyond LLMs: Advancing the Landscape of Complex Reasoning. *arXiv preprint arXiv:2402.08064* (2024).

[13] Aleksandr Chuklin, Ilya Markov, and Maarten de Rijke. 2015. *Click Models for Web Search*. Morgan & Claypool. doi:10.2200/S00654ED1V01Y201507ICR043

[14] Kevyn Collins-Thompson, Preben Hansen, and Claudia Hauff. 2017. Search as learning (dagstuhl seminar 17092). (2017).

[15] Célien Donzé, Jonathan Guerne, Henrique Marques Reis, and Pedro Costa. 2024. RAG: Unveiling the Power of Retrieval-Augmented Generation. In *Proceedings of the 9th edition of the Swiss Text Analytics Conference*, Capol Corsin, Cieliebak Mark, Weichselbraun Albert, Musat Claudiu, Maier Elisabeth, and Zimmermann Lucas (Eds.). Association for Computational Linguistics, Chur, Switzerland, 228–229. https://aclanthology.org/2024.swisstext-1.51/

[16] Jonathan St BT Evans and Keith E Stanovich. 2013. Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science* 8, 3 (2013), 223–241.

[17] Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 6491–6501.

[18] Michael Färber and Adam Jatowt. 2020. Citation recommendation: approaches and datasets. *International Journal on Digital Libraries* 21, 4 (2020), 375–405.

[19] Keith Frankish. 2010. Dual-process and dual-system theories of reasoning. *Philosophy Compass* 5, 10 (2010), 914–926.

[20] Samuel J Gershman and Nathaniel D Daw. 2017. Reinforcement learning and episodic memory in humans and animals: an integrative framework. *Annual review of psychology* 68, 1 (2017), 101–128.

[21] Ben Goertzel and Cassio Pennachin. 2007. *Artificial general intelligence*. Vol. 2. Springer.

[22] Ekaterina D Gribkova, Girish Chowdhary, and Rhanor Gillette. 2024. Cognitive mapping and episodic memory emerge from simple associative learning rules. *Neurocomputing* 595 (2024), 127812.

[23] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948* (2025).

[24] Jiafeng Guo, Yixing Fan, Liang Pang, Liu Yang, Qingyao Ai, Hamed Zamani, Chen Wu, W Bruce Croft, and Xueqi Cheng. 2020. A deep look into neural ranking models for information retrieval. *Information Processing & Management* 57, 6 (2020), 102067.

[25] Kyle Hamilton, Aparna Nayak, Bojan Božić, and Luca Longo. 2024. Is neuro-symbolic AI meeting its promises in natural language processing? A structured review. *Semantic Web* 15, 4 (2024), 1265–1306.

[26] Oz Huly, Idan Pogrebinsky, David Carmel, Oren Kurland, and Yoelle Maarek. 2024. Old IR Methods Meet RAG. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 2559–2563. doi:10.1145/3626772.3657935

[27] Gary Jones and Frank E Ritter. 2003. Production systems and rule-based inference. *Encyclopedia of cognitive science* 3 (2003), 741–747.

[28] Karen Sparck Jones and Peter Willett. 1997. *Readings in information retrieval*. Morgan Kaufmann.

[29] Janet Kolodner. 2014. *Case-based reasoning*. Morgan Kaufmann.

[30] Saar Kuzi, Abhishek Narwekar, Anusri Pampari, and ChengXiang Zhai. 2019. Help me search: Leveraging user-system collaboration for query construction to improve accuracy for difficult queries. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1221–1224.

[31] John E Laird, Allen Newell, and Paul S Rosenbloom. 1987. Soar: An architecture for general intelligence. *Artificial intelligence* 33, 1 (1987), 1–64.

[32] Yann LeCun. 2022. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review* 62, 1 (2022), 1–62.

[33] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.

[34] Xiaoxi Li, Jiajie Jin, Yujia Zhou, Yuyao Zhang, Peitian Zhang, Yutao Zhu, and Zhicheng Dou. 2024. From matching to generation: A survey on generative information retrieval. *arXiv preprint arXiv:2404.14851* (2024).

[35] Lei Liang, Mengshu Sun, Zhengke Gui, Zhongshu Zhu, Zhouyu Jiang, Ling Zhong, Yuan Qu, Peilong Zhao, Zhongpu Bo, Jin Yang, et al. 2024. Kag: Boosting llms in professional domains via knowledge augmented generation. *arXiv preprint arXiv:2409.13731* (2024).

[36] Tie-Yan Liu et al. 2009. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval* 3, 3 (2009), 225–331.

[37] Eric Modesitt, Ke Yang, Spencer Hulsey, Chengxiang Zhai, and Volodymyr Kindratenko. 2024. ORBIT: Cost-Effective Dataset Curation for Large Language Model Domain Adaptation with an Astronomy Case Study. *arXiv preprint arXiv:2412.14436* (2024).

[38] OpenAI. 2023. GPT-4 Technical Report. https://doi.org/10.48550/arXiv.2303.08774. arXiv:2303.08774.

[39] Fabio Petroni, Federico Siciliano, Fabrizio Silvestri, and Giovanni Trappolini. 2024. IR-RAG@ SIGIR24: Information retrieval's role in RAG systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3036–3039.

[40] Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Xuanhe Zhou, Yufei Huang, Chaojun Xiao, et al. 2024. Tool learning with foundation models. *Comput. Surveys* 57, 4 (2024), 1–40.

[41] Changle Qu, Sunhao Dai, Xiaochi Wei, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, Jun Xu, and Ji-Rong Wen. 2025. Tool learning with large language models: A survey. *Frontiers of Computer Science* 19, 8 (2025), 198343.

[42] Stephen E Robertson. 1977. The probability ranking principle in IR. *Journal of documentation* 33, 4 (1977), 294–304.

[43] Mark Sanderson et al. 2010. Test collection based evaluation of information retrieval systems. *Foundations and Trends® in Information Retrieval* 4, 4 (2010), 247–375.

[44] Stephen Slade. 1991. Case-based reasoning: A research paradigm. *AI magazine* 12, 1 (1991), 42–42.

[45] Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, et al. 2024. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561* 3 (2024).

[46] Alaa Tharwat and Wolfram Schenck. 2023. A survey on active learning: State-of-the-art, practical challenges and research directions. *Mathematics* 11, 4 (2023), 820.

[47] Xin Wang, Yudong Chen, and Wenwu Zhu. 2021. A survey on curriculum learning. *IEEE transactions on pattern analysis and machine intelligence* 44, 9 (2021), 4555–4576.

[48] Zihan Wang, Xuri Ge, Joemon M Jose, Haitao Yu, Weizhi Ma, Zhaochun Ren, and Xin Xin. 2024. R3AG: First Workshop on Refined and Reliable Retrieval Augmented Generation. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*. 307–310.

[49] Siyu Wu, Alessandro Oltramari, Jonathan Francis, C Lee Giles, and Frank E Ritter. 2024. Cognitive LLMs: Towards Integrating Cognitive Architectures and

Large Language Models for Manufacturing Decision-making. *arXiv preprint arXiv:2408.09176* (2024).

[50] Siyu Wu, Rodrigo F Souza, Frank E Ritter, and Walter T Lima Jr. 2023. Comparing LLMs for Prompt-Enhanced ACT-R and Soar Model Development: A Case Study in Cognitive Simulation. In *Proceedings of the AAAI Symposium Series*, Vol. 2. 422–427.

[51] Violet Xiang, Charlie Snell, Kanishk Gandhi, Alon Albalak, Anikait Singh, Chase Blagden, Duy Phung, Rafael Rafailov, Nathan Lile, Dakota Mahan, et al. 2025. Towards System 2 Reasoning in LLMs: Learning How to Think With Meta Chain-of-Though. *arXiv preprint arXiv:2501.04682* (2025).

[52] Ke Yang, Volodymyr Kindratenko, and ChengXiang Zhai. 2024. TinyHelen's First Curriculum: Training and Evaluating Tiny Language Models in a Simpler Language Environment. *arXiv preprint arXiv:2501.00522* (2024).

[53] Yutao Yang, Jie Zhou, Xuanwen Ding, Tianyu Huai, Shunyu Liu, Qin Chen, Yuan Xie, and Liang He. 2025. Recent advances of foundation language models-based continual learning: A survey. *Comput. Surveys* 57, 5 (2025), 1–38.

[54] ChengXiang Zhai. 2024. Large language models and future of information retrieval: opportunities and challenges. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 481–490.

[55] ChengXiang Zhai. 2024. Large Language Models and Future of Information Retrieval: Opportunities and Challenges. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 481–490. doi:10.1145/3626772.3657848

[56] Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. 2022. 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* (Virtual Event, AZ, USA) *(WSDM '22)*. Association for Computing Machinery, New York, NY, USA, 1328–1336. doi:10.1145/3488560.3498443

[57] Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji-Rong Wen. 2023. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107* (2023).