Klesel, Michael; Wittmann, H. Felix

Retrieval-Augmented Generation (RAG)

Business & Information Systems Engineering

Provided in Cooperation with:
Springer Nature

Suggested Citation: Klesel, Michael; Wittmann, H. Felix (2025) : Retrieval-Augmented Generation (RAG), Business & Information Systems Engineering, ISSN 1867-0202, Springer Fachmedien Wiesbaden GmbH, Wiesbaden, Vol. 67, Iss. 4, pp. 551-561, https://doi.org/10.1007/s12599-025-00945-3

Bus Inf Syst Eng 67(4):551–561 (2025)
https://doi.org/10.1007/s12599-025-00945-3

CATCHWORD

Retrieval-Augmented Generation (RAG)

Michael Klesel · H. Felix Wittmann

Received: 22 July 2024 / Accepted: 7 April 2025 / Published online: 1 June 2025

Keywords Retrieval-augmented generation · Artificial intelligence · Large language models · Information retrieval

1 Introduction

The necessity for information is a fundamental aspect of human nature, and as such, there are ongoing efforts to enhance information retrieval with information systems (Alavi and Leidner 2001; Alavi et al. 2024). Companies are particularly affected by this, as they have extensive data at their disposal, and employees need to access it. Unfortunately, current systems are not able to adequately meet employees' expectations. In fact, studies have shown that 79% of employees are dissatisfied with the user interfaces of enterprise search systems (Cleverley and Burnett 2019). This has led to a need for new approaches that can better address the information needs of organizations.

Conversational agents (CAs) powered by Artificial Intelligence (AI) and transformer-based large language models (LLMs) in particular (Vaswani et al. 2017) have revolutionized the way information can be accessed today. When compared to traditional enterprise systems, CAs offer two key benefits: Firstly, they enable users to pose questions in a natural and intuitive manner using natural language, receiving responses that are similarly conversational. Secondly, they are increasingly capable of tackling complex search tasks, facilitating problem-solving and decision-making in various domains (White 2024). For instance, individuals can use CAs to access recipe information for cooking (Jaber et al. 2024) and to obtain assistance with complex chemistry-related tasks (Bran et al. 2024) and geometric problems (Trinh et al. 2024).

In organizations, the need for information is often related to data about the organization that is not typically found on the Internet. For example, an employee may need a summary of a comprehensive requirements analysis or details of a contract. Since vanilla LLMs are unlikely to have used the necessary data, such as contract documents, as part of the training, the answers generated by LLMs are likely to be unreliable. Generally, LLMs may occasionally generate answers without a factual basis. This type of answer has been termed *hallucination* (Maynez et al. 2020; Ji et al. 2023), which is defined as "*content that is inconsistent with real-world facts or user inputs*" (Ji et al. 2023, p. 1).[1] Hallucinations are particularly critical, because they undermine the trustworthiness of the results and have been observed in various scenarios, such as multilingual use of LLMs (Guerreiro et al. 2023), or in context-specific situations, such as medicine (Pal et al. 2023).

Retrieval-augmented generation (RAG) has been proposed as a new framework for AI that seeks to integrate additional knowledge, such as organizational data, and generate results that can be linked to that knowledge (Lewis et al. 2020). This allows users to access information

Accepted after two revisions by Christine Legner

M. Klesel (✉) · H. F. Wittmann
Frankfurt University of Applied Sciences, Nibelungenplatz 1, Frankfurt, Germany
e-mail: michael.klesel@fra-uas.de

M. Klesel
Hessian Center for AI (hessian.AI), Darmstadt, Germany

[1] Recently, the literature has suggested *bullshit* as a more appropriate term, since there is no concept of truthfulness in the training of LLMs (Hicks et al. 2024). Although we agree that there is merit in proposing a new and arguably more appropriate term, we use hallucinations in this manuscript to ensure consistency with the relevant literature.

from within an organization and reduces the risk of hallucinations. This new architecture offers important advancements compared to previous architectures and presents new challenges for research and academia.

Previous catchword articles have already covered important aspects of AI, namely fair AI (Feuerriegel et al. 2020), AI as a Service (Lins et al. 2021), foundation models (Schneider et al. 2024), and generative AI (Feuerriegel et al. 2024). We contribute to this ongoing engagement with current AI developments by focusing on RAG. Specifically, we review the fundamental architecture of RAG and highlight some extensions that can enhance a plain vanilla RAG architecture. We showcase how RAG can be used in different use-case scenarios and summarize the most important advantages and challenges that should be considered when using RAG and RAG-specific extensions. Finally, we discuss important research avenues for the *BISE* community by highlighting implications that emerge as a consequence of using RAG architectures.

2 Retrieval-Augmented Generation (RAG)

2.1 Fundamental Framework

The core idea of RAG is to combine the generative capabilities of LLMs with external knowledge retrieved from a separate database (e.g., an organizational database) (Lewis et al. 2020). While Lewis et al. (2020) acknowledge previous work on the integration of external data (Guu et al. 2020; Karpukhin et al. 2020; Perez et al. 2019), they coined the term "*Retrieval-Augmented Generation* (RAG)" and proposed a general framework that leverages the strength of pretrained parametric memory (i.e., the LLM) with non-parametric memory (i.e., a separate database) as a new way to improve the performance for knowledge-intensive tasks.

Parametric memory refers to information that is stored in the parameters of a model. Rather than directly storing readily meaningful data, the parametric memory stores model parameters that can be used later to regenerate information. The more parameters a model has, the more information it can represent faithfully (Brown et al. 2020). Current models typically include the number of parameters in their name. For example, Mistral 7B (Jiang et al. 2023) is a model with seven billion ($7 \cdot 10^9$) parameters. While the number of parameters is a technical detail, it has important consequences. For example, model evaluation typically involves this number, since larger models require more resources to run. This is why models are usually compared with other models with the same number of parameters. On the other hand, if a small model performs well compared to a large model, the small model is usually preferable. In the context of RAG, it is important to note that the parametric memory only contains information that has been provided as part of the training.

In contrast, non-parametric memory, or external memory, refers to information that is outside of a model (e.g., information from a database). Therefore, this information is independent of the constraints of the model. Examples of non-parametric memory resources include Internet sites such as Wikipedia or domain-specific data (e.g., organizational data). In other words, non-parametric memory allows the integration of knowledge not previously used in the training process. For example, Veturi et al. (2024) use organizational data (e.g., policy documents) to enhance the performance of a frequently asked questions (FAQ) system.

In principle, the data used for training an LLM (e.g., a text corpus from Wikipedia) could also be used for the non-parametric memory. In fact, most current user interfaces use some kind of RAG architecture to source the factual details of the original data. For example, an FAQ system (Veturi et al. 2024) will not only provide an answer to a specific question, but also a link to the corresponding document (e.g., a policy document). Since most organizations use LLMs from a vendor (e.g., from Microsoft), RAG can be used to add internal data and therefore contextual knowledge. Similar to model training with LLMs, a RAG architecture requires a data collection phase where the external (i.e., non-parametric) data is stored in a dedicated database. It is important to note that this database is distinct from the parametric memory (i.e., the LLM).

Figure 1 provides an overview of the differences between a foundation model, LLM Fine-Tuning, and RAG. In the first scenario (*Foundation Model*), all training data results in an LLM and is part of the parametric model. Most commonly, models are trained on massive text corpora from the World Wide Web (Touvron et al. 2023), including the CommonCrawl dataset and the Pile (Gao et al. 2020). The model creation and training of the foundation model requires massive resources and infrastructure. Therefore, the creation is only feasible for very large organizations. Many of these models are accessible as a service and can thus be used by anyone. Foundation models are well-equipped for a wide range of applications that do not depend on organizational or private data (Schneider et al. 2024).

In the second scenario (*LLM Fine-Tuning*), additional domain-specific data, such as internal documents, are used to update the parameters of the LLM[2] and thereby improve

[2] This is commonly done using backpropagation (Rumelhart et al. 1986). Fine-tuning of LLMs has been greatly facilitated by techniques such as LoRA (Hu et al. 2022) that build on backpropagation. For a current overview, see for instance (Ding et al. 2023).

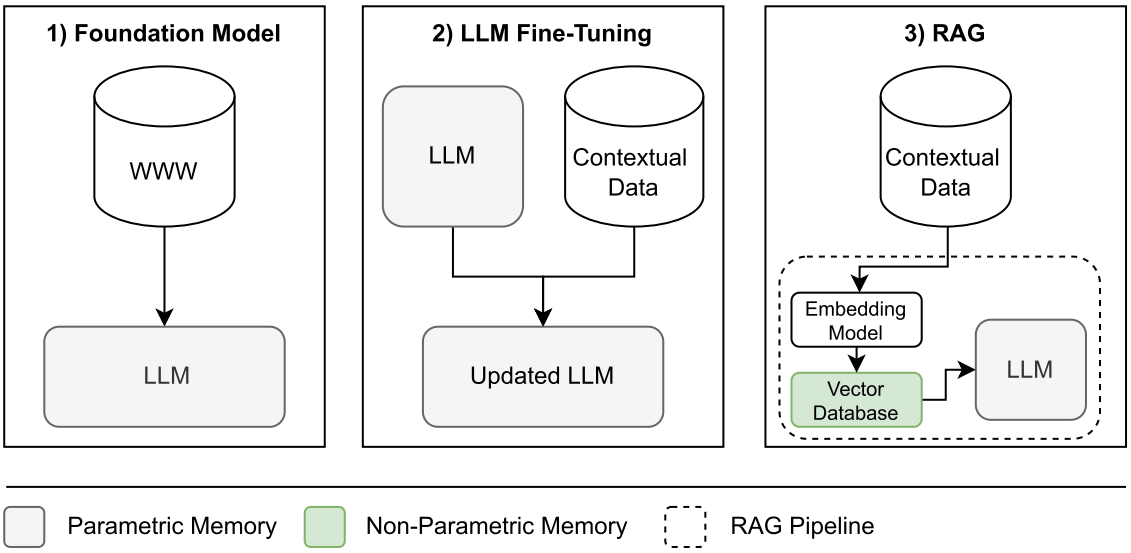

Fig. 1 Comparison of LLM approaches

the performance of an LLM with respect to the specific requirements and domain tasks at hand.

Fine-tuning is less expensive and requires fewer resources than building a foundation model. This allows smaller organizations and individuals to fine-tune foundation models for a specific context. This is particularly interesting for applications that require domain-specific knowledge that cannot be found in commonly used text corpora available on the Internet. For example, fine-tuning can be used to train a model that is able to answer questions that are specific to the domain of agriculture (Balaguer et al. 2024).

In the third scenario (*RAG*), a separate database with vector information (i.e., embedding) is generated using an embedding model and contextual data. This separate part is called the non-parametric memory. As we will explain later, the vector database is used for augmentation with the LLM to improve its result. The creation of a vector database is even less resource intensive compared to fine-tuning (Balaguer et al. 2024). Therefore, with sufficient technological capabilities, this is in principle feasible for many organizations.

Using a RAG architecture results in a pipeline, which is shown in Fig. 2. The RAG pipeline begins with a query and ends with a result. In between, there are three fundamental parts, namely *retrieval*, *augmentation*, and *generation*. Note that the augmentation is the output of the retriever and serves as the input for the generator.

Embedding Model The embedding model translates data of different modalities such as text, audio, images, or video into a vector. Additionally and importantly, the same embedding model that was used for the creation of the vector database must be used to translate the input query into a vector, as the similarity between query and chunks (or documents) in the database is measured using these vectors (Steck et al. 2024).

Retriever The retriever searches for the most relevant information in the vector database by calculating the similarity score between an input query and documents in the vector database. This is done using the vectors that were calculated with the embedding model and the vector that was calculated for the query. The retriever must use the same embedding model for the query as was previously used for the documents in the vector database. As a result, the retriever suggests a context, which is typically a list of retrieved chunks or documents.[3] Most commonly, this is done by selecting the top k (e.g., top 5) hits, ranked by the similarity score. This type of ranking is sometimes called the probability ranking principle (Robertson 1977). For example, if the query refers to a specific customer, the most relevant documents (e.g., top 5 documents) related to this customer are included in the context.

Augmentation The context is used to generate an extended query. For that reason, LLMs use a type of augmentation template that defines how a user query is augmented. Most importantly, the context is explicitly included in this query. For example, a basic augmentation template instructs the language model to use specific information, stating: 'Use the following context: [context].' It then asks the model to answer a question based on that information. 'Given the context information, answer the

[3] Because context windows keep increasing, more recent solutions can use whole documents rather than smaller text chunks.

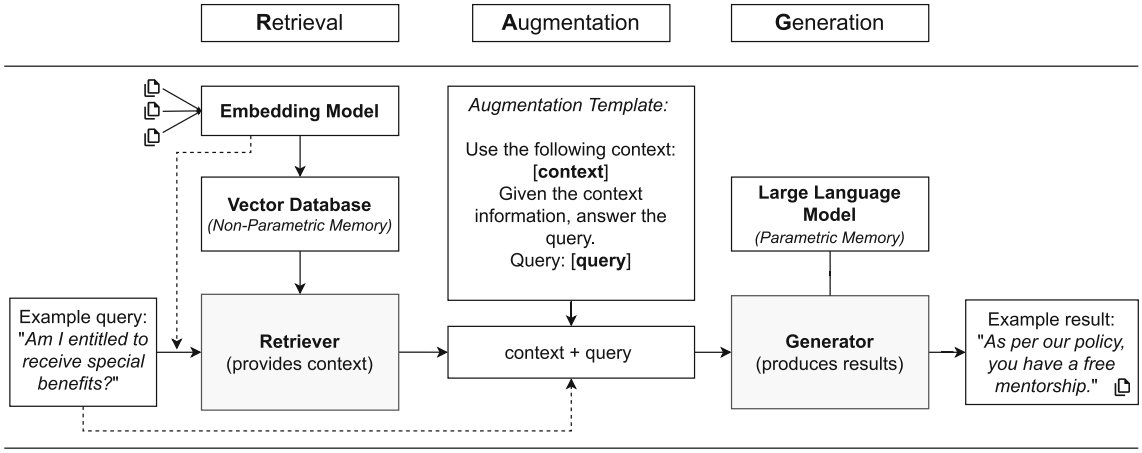

Fig. 2 A plain vanilla RAG architecture based on Lewis et al. (2020)

query: [query].' The [context] part can include links to the most relevant documents found in the vector database.

Generator The generator takes the query augmented with information from the database and generates a new result. Again, the generation process can use the information of the context and provide the link to the original document (e.g., a hyperlink to a document).

2.2 Enhanced RAG

Figure 2 provides an overview of what can be considered a plain vanilla RAG architecture. In addition, recent literature suggests several ways to improve the performance of a basic plain vanilla RAG architecture. For example, hierarchical information retrieval approaches allow for a deeper understanding and integration of information across documents, improving performance on complex, multi-step reasoning tasks. In particular, *RAPTOR* (Recursive Abstractive Processing for Tree-Organized Retrieval) recursively embeds, clusters, and summarizes text at multiple levels of abstraction (Sarthi et al. 2024).

Further, graph-based approaches such as GraphRAG offer significant improvements by extracting knowledge graphs and structuring them hierarchically to improve RAG-based tasks (Edge et al. 2024). GraphRAG can be used to extract a knowledge graph from text, building a hierarchy that is then used to leverage these graph-based structures to perform a RAG-based task.

In addition, retrieval-augmented thoughts (RAT) enhance augmentation through a zero-shot Chain of Thought (CoT), iteratively refining it with retrieved information. This method provides more contextual and coherent output, supporting tasks such as code generation and mathematical reasoning (Wang et al. 2024c). The underlying mechanism, CoT, which has become critical to improving LLM reasoning capabilities, was first developed by Google in 2022 (Wei et al. 2022) and has now been integrated into models such as GPT-4o.

Moreover, retrieval-augmented fine-tuning (RAFT) combines the advantages of RAG and fine-tuning, creating synthetic datasets for fine-tuning models to specific domains (Zhang et al. 2024). RAFT outperforms traditional RAG in specialized domains such as medicine. It involves the creation of a synthetic dataset of queries, relevant documents, and target responses. A model can be fine-tuned on this dataset to align it with the domain knowledge and style. RAFT allows the model to "study" the domain knowledge in advance, resulting in better performance than traditional RAG.

Innovative methods such as RA-ISF (retrieval-augmented iterative self-feedback) decompose tasks into submodules, enhancing factual reasoning and reducing hallucinations (Liu et al. 2024).

These examples are intended to reflect the potential that RAG has to offer. For a more comprehensive overview of recent enhancements, we refer to in-depth reviews of RAG (Zhao et al. 2024; Yu et al. 2024; Gao et al. 2023).

3 Opportunities and Challenges of RAG

3.1 RAG Use Cases

In this section, we provide an overview of example use cases where RAG can be used to substantially enhance the performance of specific tasks (see Table 1).

RAG has led to the development of sophisticated question-answering agents, with one notable application being in the domain of FAQs. Through the integration of domain-specific knowledge, RAG-based FAQ systems have the capacity to generate accurate and reliable responses to commonly asked questions that can be posed

Table 1 Example use cases with RAG

Task	Example use case	References
FAQs	RAG can be used to develop a sophisticated question-answering agent. This can be implemented, for example, for FAQs. By incorporating specific knowledge, a RAG-based FAQ system is able to accurately answer specific questions posed in natural language.	Veturi et al. (2024)
Real-time information retrieval	RAG can be used to integrate additional data sources to provide real-time information retrieval. For example, the ChatGPT Android app retrieves information from the web in real time to improve the timeliness and accuracy of results. External and current material can also be incorporated by allowing users to upload documents "on the fly" and to use this information in the generation process.	Amri et al. (2024) and Khan et al. (2024)
Context-specific answers	RAG can be used to enhance the development of applications that require specific information. For example, AI-empowered programming assistants that are tailored to course-specific content can use additional data sources (such as course materials) to improve the accuracy of results and include a reference to the relevant material.	Wei et al. (2024), Kazemitabaar et al. (2024), Rai et al. (2024) and Strobel and Banh (2024)
Enhanced content generation	RAG can be used to guide the content generation process by incorporating additional data. In this way, the quality and relevance of the content generated reflects current knowledge and trends relevant to the generation of content such as commentary.	Wu et al. (2024) and Wang et al. (2024b)
Federated search	RAG can be used to enhance the capabilities of sourcing relevant information across heterogeneous data sources in combination with an LLM.	Wang et al. (2024a)

in natural language by the user, replacing the traditional method of handling FAQs – by using a set of static "canned" questions and answers – and thereby enhancing the overall user experience (Veturi et al. 2024).

RAG can also enhance real-time information retrieval, thereby improving the timeliness and the accuracy of information provided. For instance, the ChatGPT Android application sources up-to-date information from the web in real time. Additionally, RAG facilitates the incorporation of external and contemporary materials by allowing users to upload documents on demand. This capability is particularly beneficial in scenarios where the most current information is crucial (Amri et al. 2024; Khan et al. 2024).

In applications that require context-specific answers, RAG is a valuable tool. For example, AI-powered programming assistants tailored to course-specific content can draw on additional data sources, such as course materials, to improve the accuracy of their output. This approach not only increases the accuracy of results, but also ensures that generated answers are directly linked to relevant reference materials (Wei et al. 2024; Kazemitabaar et al. 2024; Rai et al. 2024; Strobel and Banh 2024).

RAG significantly enhances the content creation process by incorporating the latest knowledge and trends into the output. This capability is particularly relevant to the creation of content such as commentaries, where the inclusion of up-to-date information is critical. By guiding the content generation process with additional data, RAG ensures that the quality and relevance of the output is maintained at a high level (Wu et al. 2024; Wang et al. 2024b).

Finally, RAG enhances federated search by improving the ability to retrieve relevant information from heterogeneous data sources in conjunction with LLMs. This combination enables a more comprehensive and efficient retrieval process, ensuring that users can seamlessly access relevant information from a wide range of sources. This is particularly beneficial in complex information environments where data is distributed across multiple platforms (Wang et al. 2024a).

3.2 Opportunities of RAG

By incorporating external data, the implementation of RAG provides an *enhanced contextual understanding* (Lewis et al. 2020). As a result, queries that require specific knowledge that was not present in the LLM's training, and therefore is not reflected in the LLM's parameters, can be meaningfully processed. Using foundation models is often problematic, because the training data is outdated. For example, a model with training data from 2023 and earlier cannot answer questions related to the 2024 European elections. With RAG, more recent data, such as the official elections website, can be added, which may then be referred to as non-parametric memory. In doing so, a RAG-based system has the potential to retrieve accurate information about the election.

When LLMs try to answer questions for a specific domain that is not part of the training data, *hallucinations* are likely. One way to address this problem is to fine-tune the LLM, which is less expensive than building a foundation model but requires considerable resources nonetheless. Studies have shown that fine-tuning can also lead to hallucinations (Gekhman et al. 2024). On the other hand, RAG is an effective way to add this knowledge. By adding additional information, questions can be answered using this data, reducing the likelihood of inaccurate answers. Thus, RAG is an effective measure to enhance *factual accuracy*.

A RAG architecture allows references to be provided to the contextual data stored in the vector database. Providing valid references to the generated result has been termed *grounding* (Magesh et al. 2024). Grounding is a significant advantage, because it gives a user additional information about where the information comes from. Therefore, a user looking for information in a particular area can follow this reference to double check the answer and get additional information.[4]

In addition, a RAG architecture can be used to limit the response spectrum to a desired knowledge domain (*knowledge-domain guardrails for LLMs*), which can be implicitly defined by providing additional knowledge. Foundation models often lack precision in responses, because they are not provided with appropriate boundary conditions and guardrails that focus the solution space of an LLM. A conversational agent deployed on a business website would be prevented from providing answers that are irrelevant to the business interests. For example, a query such as "*Tell me a joke*" should not be answered by an agent in an application (e.g., an educational platform), because it would not be in the best interest of the platform (e.g., due to cost considerations). The "*Tell me a joke*" example is simple enough, and a CA answer could therefore be prevented by an explicit guardrail in the prompt. The boundaries of what should or should not be answered by a business chatbot might be more complex and less easy to define explicitly in a simple prompt. RAG offers the possibility to do this implicitly by restricting the chatbot's answers to the domain covered by the documents that were used to create the vector database in conjunction with an appropriate prompt. Consider the augmentation template shown in Fig. 2. This template can be used to restrict the response of an LLM to a particular context specified in the template. In other words, the augmentation template defines the boundaries by reference to the database and can therefore stay the same when the database changes (e.g., when an organization adds additional data to the vector database). In our example augmentation template in Fig. 2 – "Give the context information, answer the query" – the LLM response should be generated only within the context. Assuming that the context is a database of customer information, the LLM is guided to generate answers based on these documents.

Finally, a RAG architecture also comes with reduced *initial cost of ownership*, because it is less computationally intensive to create a vector database than fine-tune a foundation model. Therefore, RAG is a potential alternative to LLM fine-tuning (Table 2).

3.3 Challenges Amplified by RAG

Along with these opportunities, RAG also presents new challenges for organizations. At the most fundamental level, most of the challenges to RAG relate to *data management* and *machine learning operations (MLOps) capabilities*. Data management has been identified as a major challenge in IS research (Abbasi et al. 2016) in general. Since RAG-based systems require additional efforts to merge data from heterogeneous data sources, these challenges are compounded. Therefore, organizations need to develop additional capabilities to address this need. Modern concepts such as data mesh structures (Dehghani 2022; Blohm et al. 2024) can also be considered as a useful approach for developing RAG-based systems. In addition, much effort is required to ensure the high quality of the additional data. In practice, there may be cases where the data sources contain counterfactual or even false information that should be eliminated. Therefore, organizations need to allocate more resources and build new capabilities, such as MLOps capabilities, to implement RAG-based systems.

In addition to data management issues, the underlying data inevitably presents new challenges in terms of unwanted *bias* effects. This is because organizations have a new way of adding data to the AI infrastructure via a vector database. This is arguably similar to LLM fine-tuning, where the organization should also be aware of unwanted new bias effects. However, it differs from the use of foundation models or Software-as-a-Service (SaaS) solutions where the organization cannot influence the data used to train or fine-tune the model. A well-known example is the use of Western documents, which are likely to reflect only a Western perspective and may be undesirable in an international context. This is part of a larger area of ongoing research related to the avoidance and correction of bias effects (e.g., Mehrabi et al. 2022; Gallegos et al. 2024), an area that is also regulated by the European Union (see, for example, the EU AI Act).

[4] Foundation models per se do not provide such references. The results are referred to as *ungrounded* (Magesh et al. 2024). It is worth noting that there are also references that do not support a generated output. These are referred to as *misgrounded*.

Table 2 Summary of opportunities empowered by RAG-based architectures

Opportunities	Description
Enhanced contextualized understanding	RAG architectures "use the input sequence x to retrieve text documents z and use them as additional context" (Lewis et al. 2020, p. 2). Therefore, a RAG-based system does have an extended contextual understanding compared to foundation models.
Reduced hallucinations and improved factual accuracy	By including an extended context, a RAG-based system can generate outcomes based on the context which reduces hallucinations and increases factual accuracy (Lewis et al. 2020; Shuster et al. 2021)
Grounding	Generated output is combined with references to the contextual data (i.e., grounding (Magesh et al. 2024)). This allows users to double check the output and get additional information following the reference.
Knowledge domain guardrails for LLMs	RAG-based systems can be used to specify the domain that is within the interest of the provided data. Therefore, irrelevant or undesirable domains can be excluded.
Initial cost of ownership	Developing a vector database is much more cost-effective compared to fine-tuning when it comes to the total cost of ownership (Balaguer et al. 2024)

A plain vanilla RAG architecture (see Fig. 2) also suffers from what we would like to call the "*blinkered chunk effect*" (BCE). Suppose that you extract a paragraph (chunk) from a large text document such as a Harry Potter book. To what extent could one (as a human) understand that paragraph without having read the entirety of the novel? It is likely that a lack of understanding would be apparent, particularly with regard to terms that are unique to the context of the novels and the main plot of the story. Although this may be an extreme example due to the size and scope of the imaginary universe, which includes magic and fictional characters, the principle of the BCE also applies to contextualized documents in business applications. Therefore, the use of RAG with rich data still has limitations in terms of comprehensive understanding. In such cases, recent developments, including *RAPTOR* (Sarthi et al. 2024) and GraphRAG (Edge et al. 2024), should be considered.

The performance of the retriever depends on an effective ranking system, which means that the ranking mechanism is able to identify the most relevant documents. Most commonly, the ranking system is built upon the probability ranking principle (Robertson 1977), which is not always ideal. For that reason, new approaches can be considered to improve the ranking results. Current approaches include permutation-invariant ranking models (Pang et al. 2020), list-aware re-rankings, and hybrid searches (Bruch et al. 2023) (Table 3).

4 Implications for BISE Researchers

This catchword article seeks to provide a fundamental overview of RAG, highlight characteristics of RAG architectures, and outline implications of RAG. Since prior work has already identified important avenues for research with foundation models (Schneider et al. 2024; Feuerriegel et al. 2024), we illustrate some nuanced research questions that occur in combination with RAG from three different perspectives: (1) *organizational*, (2) *individual*, and (3) *economic* (see Fig. 3).

Firstly, the use of RAG-based architectures has implications for organizations. Similar to previous software architectures and paradigms, organizations need new skills to implement and leverage new AI technologies (Berente et al. 2021). This is particularly true for IT architecture and data management capabilities, which are a major challenge for organizations (Abbasi et al. 2016; Blohm et al. 2024). In addition to well-known shortcomings such as the centralization of data management (Velu et al. 2013), a RAG-based architecture brings new challenges that organizations need to address. In particular, organizations need to determine how they will use their data. For example, an organization's dataset may be used for model fine-tuning (i.e., LLM fine-tuning), vector database development (i.e., RAG), or both (e.g., using RAFT). From a theoretical perspective, appropriate configurations should be identified (Park and Mithas 2020) that guide organizations in how to organize their data for fine-tuning, RAG, or both. Furthermore, identifying trade-offs and preferred configurations is challenging, because it is highly dependent on contextual and environmental factors such as organizational size or industry. For this reason, it also raises questions about the development of internal versus external capabilities (Nevo et al. 2007). The following research questions are examples of BISE scholars conducting research at the organizational level: *Do high levels of data management capability, e.g., Data Mesh including RAG, lead to high levels of organizational performance?*, *What is the optimal balance of data going into RAG versus fine-*

Table 3 Summary of important challenges exacerbated by RAG

Challenges	Description
Additional data management/MLOps capabilities required	RAG-based systems require the inclusion of heterogeneous and potentially dynamically changing data sources. Therefore, new data management capabilities are required to meet this need. New concepts such as data mesh structures are potentially useful for RAG-based systems (Dehghani 2022; Blohm et al. 2024).
Potential new biases	With an extended contextual understanding by means of new data, there is also a risk of introducing new bias effects requiring additional efforts to prevent the propagation of bias effects in LLMs and RAG-based systems. For a current overview, see for instance Mehrabi et al. (2022) or Gallegos et al. (2024).
Blinkered chunk effect (BCE)	A plain vanilla RAG implementation is limited in terms of a comprehensive understanding of extensive data. New approaches such as *RAPTOR* (Sarthi et al. 2024) are required to reduce this issue.
Retrieval effectiveness	The effectiveness of a RAG architecture depends on how effectively the retrieval mechanism works. This includes the effectiveness of the document ranking (i.e., are the most relevant documents ranked first?) and how well the retrieval process performs.

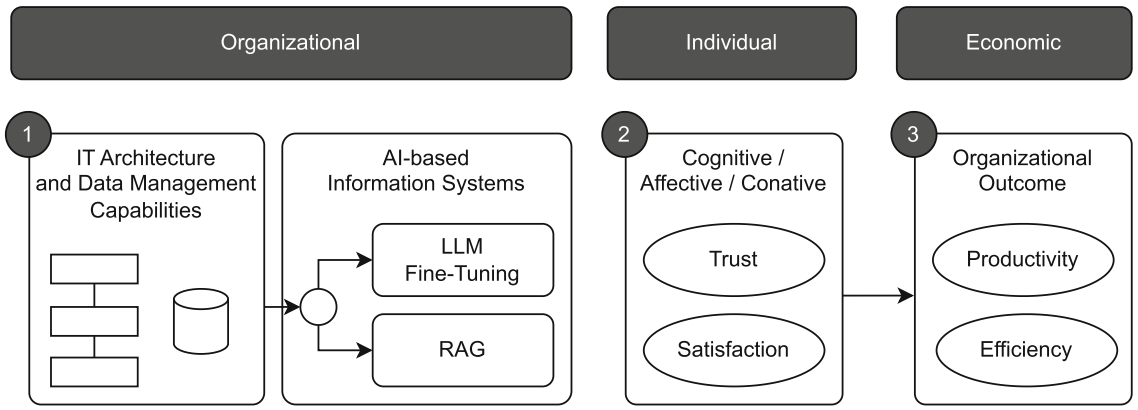

Fig. 3 Research questions related to RAG

tuning that leads to superior organizational performance?, or *To what extent does a RAG-based architecture contribute to better IT business alignment?*

Secondly, individuals interacting with RAG-based systems (e.g., using a CA) will experience changes in how results are presented. Most importantly, RAG offers the ability to add references to contextual data, which has been coined "grounding" (Magesh et al. 2024). Providing references to the contextual data is closely related to the concept of Explainable AI (XAI) (Schneider 2024; Longo et al. 2024), because users get additional information about the results of an LLM. So far, prior literature has used different approaches to provide post-hoc explanations including LIME (Ribeiro et al. 2016), Shapley values (SHAP) (Lundberg and Lee 2017), or gradient-weighted class activation mapping (Grad-CAM) (Selvaraju et al. 2017). These approaches add an extra layer that provides visual or textual explanations. For example, SHAP can be used to highlight parts of an image that have an important influence on the AI prediction. Providing references – such as citations in a book – to the contextual data can be considered as an alternative and complementary approach to provide an additional explanation layer to users. Various determinants of behavior, including cognitive, affective, and conative (CAC) constructs (Bagozzi 1992), may be influenced. This is similar to previous research that has investigated the relationship between XAI and latent constructs such as trust (Hamm et al. 2023) or intention to use (Meske and Bunde 2022). We argue that the impact of grounding is still under-researched, and more empirical data is required to investigate if grounding is a valuable addition to XAI. Moreover, it needs to be explored to what extent grounding influences perceived constructs such as perceived explainability or perceived trusting intentions. In addition to exploring latent constructs, RAG-based systems also have the potential to reduce the actual retrieval time, which in turn can increase user performance. Since RAG-based systems are commonly used as a foundation for a CA, they are a state-of-the-art alternative to more traditional knowledge-based systems such as an intranet or

Wikipedia and may be more effective in finding relevant information. Example research questions that emerge with RAG-based systems relevant for individual-level BISE research are *Can grounding-based explanations outperform traditional XAI approaches in improving user trust?* or *To what extent does RAG enhance individuals' workplace performance?*

Finally, RAG also invites more research that investigates the economic value of new information systems architectures. More generally, research is needed that investigates what outcomes can be expected from the evolution and advancements of new architectures (Haki et al. 2020). Ultimately, organizations seek opportunities to increase efficiency and productivity. RAG has the potential to improve business processes and enhance organizational decision making. Nevertheless, it remains unclear to what extent organizations can benefit from using RAG. In addition, RAG can also help organizations meet regulatory requirements. For example, the European AI Act asks for more transparency when AI is used. Again, the concept of grounding has the potential to contribute to this requirement and offer a potential pathway for organizations. For these reasons, the following research questions are examples for business-oriented *BISE* researchers: *How can organizations achieve a competitive advantage with RAG-based systems?* and *To what extent can RAG-based systems enhance the fulfillment of regulatory requirements?*

Acknowledgements The authors would like to express sincere gratitude to the Department Editor, Christine Legner, for her constructive guidance throughout the review process, and to the two anonymous reviewers for their thoughtful feedback that significantly strengthened this work.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Bran M, Cox AS, Schilter O, Baldassari C, White AD, Schwaller P (2024) Augmenting large language models with chemistry tools. Nat Mach Intell 6:525–535. https://doi.org/10.1038/s42256-024-00832-8

Ding N, Qin Y, Yang G, Wei F, Yang Z, Su Y, Sun M (2023) Parameter-efficient fine-tuning of large-scale pre-trained language models. Nat Mach Intell 5(3):220–235. https://doi.org/10.1038/s42256-023-00626-4

Abbasi A, Sarker S, Chiang R (2016) Big data research in information systems: Toward an inclusive research agenda. J Assoc Inf Syst 17(2):3. https://doi.org/10.17705/1jais.00423

Alavi M, Leidner DE (2001) Knowledge management and knowledge management systems: Conceptual foundations and research issues. MIS Q 25(1):107–136. https://doi.org/10.2307/3250961

Alavi M, Leidner DE, Mousavi R (2024) Knowledge management perspective of generative artificial intelligence. J Assoc Inf Syst 25(1):1–12. https://doi.org/10.17705/1jais.00859

Amri S, Bani R, Bani S (2024) An approach to the analysis of financial documents using generative AI. In: Proceedings of the 7th international conference on networking, intelligent systems and security, ACM, Meknes, pp 1–5. https://doi.org/10.1145/3659677.3659736

Bagozzi RP (1992) The self-regulation of attitudes, intentions, and behavior. Soc Psychol Q 55(2):178. https://doi.org/10.2307/2786945

Balaguer A, Benara V, Cunha RLdF, Filho RdME, Hendry T, Holstein D, Marsman J, Mecklenburg N, Malvar S, Nunes LO, Padilha R, Sharp M, Silva B, Sharma S, Aski V, Chandra R (2024) RAG vs fine-tuning: pipelines, tradeoffs, and a case study on agriculture. https://doi.org/10.48550/ARXIV.2401.08406

Berente N, Gu B, Recker J, Santhanam R (2021) Managing artificial intelligence. MIS Q 45(3):1433–1450. https://doi.org/10.25300/MISQ/2021/16274

Blohm I, Wortmann F, Legner C, Köbler F (2024) Data products, data mesh, and data fabric: New paradigm(s) for data and analytics? Bus Inf Syst Eng 66:643–652. https://doi.org/10.1007/s12599-024-00876-5

Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D (2020) Language models are few-shot learners. In: Proceedings of the 34th international conference on neural information processing systems. Curran Associates Inc., Red Hook, NY, USA, NIPS '20, pp 1877–1901. https://doi.org/10.5555/3495724.3495883

Bruch S, Gai S, Ingber A (2023) An analysis of fusion functions for hybrid retrieval. ACM Trans Inf Syst 42(1):1–35. https://doi.org/10.1145/3596512

Cleverley PH, Burnett S (2019) Enterprise search and discovery capability: The factors and generative mechanisms for user satisfaction. J Inf Sci 45(1):29–52. https://doi.org/10.1177/0165551518770969

Dehghani Z (2022) Data mesh delivering data-driven value at scale. O'Reilly, Sebastopol

Edge D, Trinh H, Cheng N, Bradley J, Chao A, Mody A, Truitt S, Larson J (2024) From local to global: A graph RAG approach to query-focused summarization. https://doi.org/10.48550/arXiv.2404.16130

Feuerriegel S, Dolata M, Schwabe G (2020) Fair AI: Challenges and opportunities. Bus Inf Syst Eng 62(4):379–384. https://doi.org/10.1007/s12599-020-00650-3

Feuerriegel S, Hartmann J, Janiesch C, Zschech P (2024) Generative AI. Bus Inf Syst Eng 66(1):111–126. https://doi.org/10.1007/s12599-023-00834-7

Gallegos IO, Rossi RA, Barrow J, Tanjim MM, Kim S, Dernoncourt F, Yu T, Zhang R, Ahmed NK (2024) Bias and fairness in large language models: A survey. Comp Linguist 50(3):1–79. https://doi.org/10.1162/coli_a_00524

Gao L, Biderman S, Black S, Golding L, Hoppe T, Foster C, Phang J, He H, Thite A, Nabeshima N, Presser S, Leahy C (2020) The pile: An 800GB dataset of diverse text for language modeling. https://doi.org/10.48550/arXiv.2101.00027

Gao Y, Xiong Y, Gao X, Jia K, Pan J, Bi Y, Dai Y, Sun J, Wang M, Wang H (2023) Retrieval-augmented generation for large language models: A survey. https://doi.org/10.48550/ARXIV.2312.10997

Gekhman Z, Yona G, Aharoni R, Eyal M, Feder A, Reichart R, Herzig J (2024) Does fine-tuning LLMs on new knowledge encourage hallucinations? https://doi.org/10.48550/arXiv.2405.05904

Guerreiro NM, Alves DM, Waldendorf J, Haddow B, Birch A, Colombo P, Martins AFT (2023) Hallucinations in large multilingual translation models. Trans Assoc Comput Linguist 11:1500–1517. https://doi.org/10.1162/tacl_a_00615

Guu K, Lee K, Tung Z, Pasupat P, Chang MW (2020) REALM: Retrieval-augmented language model pre-training. In: Proceedings of the 37th international conference on machine learning, JMLR.org, ICML'20. https://doi.org/10.5555/3524938.3525306

Haki K, Beese J, Aier S, Winter R (2020) The evolution of information systems architecture: An agent-based simulation model. MIS Q 44(1):155–184. https://doi.org/10.25300/MISQ/2020/14494

Hamm P, Klesel M, Coberger P, Wittmann HF (2023) Explanation matters: An experimental study on explainable AI. Electron Mark 33(1):17. https://doi.org/10.1007/s12525-023-00640-9

Hicks MT, Humphries J, Slater J (2024) ChatGPT is bullshit. Ethics Inf Technol 26(2):38. https://doi.org/10.1007/s10676-024-09775-5

Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L, Chen W (2022) LoRA: Low-rank adaptation of large language models. In: International conference on learning representations, virtual conference

Jaber R, Zhong S, Kuoppamäki S, Hosseini A, Gessinger I, Brumby DP, Cowan BR, Mcmillan D (2024) Cooking with agents: Designing context-aware voice interaction. In: Proceedings of the CHI conference on human factors in computing systems, ACM, Honolulu, pp 1–13. https://doi.org/10.1145/3613904.3642183

Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, Ishii E, Bang YJ, Madotto A, Fung P (2023) Survey of hallucination in natural language generation. ACM Comput Surv 55(12):1–38. https://doi.org/10.1145/3571730

Jiang AQ, Sablayrolles A, Mensch A, Bamford C, Chaplot DS, Casas Ddl, Bressand F, Lengyel G, Lample G, Saulnier L, Lavaud LR, Lachaux MA, Stock P, Scao TL, Lavril T, Wang T, Lacroix T, Sayed WE (2023) Mistral 7B. https://doi.org/10.48550/arXiv.2310.06825

Karpukhin V, Oguz B, Min S, Lewis P, Wu L, Edunov S, Chen D, Yih Wt (2020) Dense passage retrieval for open-domain question answering. In: Webber B (ed) Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), Association for Computational Linguistics, Online, pp 6769–6781. https://doi.org/10.18653/v1/2020.emnlp-main.550

Kazemitabaar M, Ye R, Wang X, Henley AZ, Denny P, Craig M, Grossman T (2024) CodeAid: Evaluating a classroom deployment of an LLM-based programming assistant that balances student and educator needs. In: Proceedings of the CHI conference on human factors in computing systems, ACM, Honolulu, pp 1–20. https://doi.org/10.1145/3613904.3642773

Khan AA, Hasan MT, Kemell KK, Rasku J, Abrahamsson P (2024) Developing retrieval augmented generation (RAG) based LLM systems from PDFs: an experience report. https://doi.org/10.48550/ARXIV.2410.15944

Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih Wt, Rocktäschel T, Riedel S, Kiela D (2020) Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Proceedings of the 34th international conference on neural information processing systems, Curran, Red Hook, NIPS '20, pp 9459–9474. https://doi.org/10.5555/3495724.3496517

Lins S, Pandl KD, Teigeler H, Thiebes S, Bayer C, Sunyaev A (2021) Artificial intelligence as a service. Bus Inf Syst Eng 63(4):441–456. https://doi.org/10.1007/s12599-021-00708-w

Liu Y, Peng X, Zhang X, Liu W, Yin J, Cao J, Du T (2024) RA-ISF: Learning to answer and understand from retrieval augmentation via iterative self-feedback. https://doi.org/10.18653/v1/2024.findings-acl.281

Longo L, Brcic M, Cabitza F, Choi J, Confalonieri R, Ser JD, Guidotti R, Hayashi Y, Herrera F, Holzinger A, Jiang R, Khosravi H, Lecue F, Malgieri G, Páez A, Samek W, Schneider J, Speith T, Stumpf S (2024) Explainable artificial intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. Inf Fusion 106:102301. https://doi.org/10.1016/j.inffus.2024.102301

Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. In: Proceedings of the 31st international conference on neural information processing systems, Curran, Red Hook, NIPS'17, pp 4768–4777. https://doi.org/10.5555/3295222.3295230

Magesh V, Surani F, Dahl M, Suzgun M, Manning CD, Ho DE (2024) Hallucination-free? Assessing the reliability of leading (AI) legal research tools. https://doi.org/10.48550/arXiv.2405.20362

Maynez J, Narayan S, Bohnet B, McDonald R (2020) On faithfulness and factuality in abstractive summarization. In: Proceedings of the 58th annual meeting of the association for computational linguistics, Association for Computational Linguistics, Online, pp 1906–1919. https://doi.org/10.18653/v1/2020.acl-main.173

Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A (2022) A survey on bias and fairness in machine learning. ACM Comput Surv 54(6):1–35. https://doi.org/10.1145/3457607

Meske C, Bunde E (2022) Design principles for user interfaces in AI-based decision support systems: the case of explainable hate speech detection. Inf Syst Front. https://doi.org/10.1007/s10796-021-10234-5

Nevo S, Wade MR, Cook WD (2007) An examination of the trade-off between internal and external IT capabilities. J Strat Inf Syst 16(1):5–23. https://doi.org/10.1016/j.jsis.2006.10.002

Pal A, Umapathi LK, Sankarasubbu M (2023) Med-HALT: medical domain hallucination test for large language models. https://doi.org/10.48550/arXiv.2307.15343

Pang L, Xu J, Ai Q, Lan Y, Cheng X, Wen J (2020) SetRank: learning a permutation-invariant ranking model for information retrieval. In: Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval, ACM, Virtual Event China, pp 499–508. https://doi.org/10.1145/3397271.3401104

Park Y, Mithas S (2020) Organized complexity of digital business strategy: a configurational perspective. MIS Q 44(1):85–127. https://doi.org/10.25300/MISQ/2020/14477

Perez E, Karamcheti S, Fergus R, Weston J, Kiela D, Cho K (2019) Finding generalizable evidence by learning to convince Q &A models. In: Inui K (ed) Conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing, Hong Kong. https://doi.org/10.18653/v1/D19-1244

Rai A, Chen L, Breazeal C, Ramesh B, Long Y, Aria A (2024) Design and evaluation attributes for scalable, cost-effective personalization of LLM tutors in programming education. In: ICIS 2024 proceedings

Ribeiro MT, Singh S, Guestrin C (2016) "Why should I trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. Association for Computing Machinery, New York, KDD '16, pp 1135–1144. https://doi.org/10.1145/2939672.2939778

Robertson S (1977) The probability ranking principle in IR. J Doc 33(4):294–304. https://doi.org/10.1108/eb026647

Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. Nature 323(6088):533–536. https://doi.org/10.1038/323533a0

Sarthi P, Abdullah S, Tuli A, Khanna S, Goldie A, Manning CD (2024) RAPTOR: Recursive abstractive processing for tree-organized retrieval. https://doi.org/10.48550/arXiv.2401.18059

Schneider J (2024) Explainable generative artificial intelligence (GenXAI): A survey, conceptualization, and research agenda. Artif Intell Rev 57(11):289. https://doi.org/10.1007/s10462-024-10916-x

Schneider J, Meske C, Kuss P (2024) Foundation models: A new paradigm for artificial intelligence. Bus Inf Syst Eng 66(2):221–231. https://doi.org/10.1007/s12599-024-00851-0

Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: 2017 IEEE international conference on computer vision (ICCV), pp 618–626, https://doi.org/10.1109/ICCV.2017.74

Shuster K, Poff S, Chen M, Kiela D, Weston J (2021) Retrieval augmentation reduces hallucination in conversation. https://doi.org/10.48550/arXiv.2104.07567

Steck H, Ekanadham C, Kallus N (2024) Is cosine-similarity of embeddings really about similarity? In: Companion proceedings of the ACM web conference 2024, ACM, Singapore, pp 887–890. https://doi.org/10.1145/3589335.3651526

Strobel G, Banh L (2024) What did the doctor say? Empowering patient comprehension with generative artificial intelligence. In: ECIS 2024 proceedings, Paphos

Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, Bashlykov N, Batra S, Bhargava P, Bhosale S, Bikel D, Blecher L, Ferrer CC, Chen M, Cucurull G, Esiobu D, Fernandes J, Fu J, Fu W, Fuller B, Gao C, Goswami V, Goyal N, Hartshorn A, Hosseini S, Hou R, Inan H, Kardas M, Kerkez V, Khabsa M, Kloumann I, Korenev A, Koura PS, Lachaux MA, Lavril T, Lee J, Liskovich D, Lu Y, Mao Y, Martinet X, Mihaylov T, Mishra P, Molybog I, Nie Y, Poulton A, Reizenstein J, Rungta R, Saladi K, Schelten A, Silva R, Smith EM, Subramanian R, Tan XE, Tang B, Taylor R, Williams A, Kuan JX, Xu P, Yan Z, Zarov I, Zhang Y, Fan A, Kambadur M, Narang S, Rodriguez A, Stojnic R, Edunov S, Scialom T (2023) Llama 2: Open foundation and fine-tuned chat models. https://doi.org/10.48550/arXiv.2307.09288

Trinh TH, Wu Y, Le QV, He H, Luong T (2024) Solving olympiad geometry without human demonstrations. Nature 625(7995):476–482. https://doi.org/10.1038/s41586-023-06747-5

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: Proceedings of the 31st international conference on neural information processing systems, Curran, Red Hook, NIPS'17, pp 6000—6010. https://doi.org/10.5555/3295222.3295349

Velu CK, Madnick SE, Van Alstyne MW (2013) Centralizing data management with considerations of uncertainty and information-based flexibility. J Manag Inf Syst 30(3):179–212. https://doi.org/10.2753/MIS0742-1222300307

Veturi S, Vaichal S, Jagadheesh RL, Tripto NI, Yan N (2024) RAG based question-answering for contextual response prediction system. https://doi.org/10.48550/ARXIV.2409.03708

Wang S, Khramtsova E, Zhuang S, Zuccon G (2024a) FeB4RAG: Evaluating federated search in the context of retrieval augmented generation. In: Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval, ACM, Washington DC USA, pp 763–773. https://doi.org/10.1145/3626772.3657853

Wang Y, Lipka N, Zhang R, Siu A, Zhao Y, Ni B, Wang X, Rossi R, Derr T (2024b) Topology-aware retrieval augmentation for text generation. In: Proceedings of the 33rd ACM international conference on information and knowledge management, ACM, Boise, pp 2442–2452. https://doi.org/10.1145/3627673.3679746

Wang Z, Liu A, Lin H, Li J, Ma X, Liang Y (2024c) RAT: Retrieval augmented thoughts elicit context-aware reasoning in long-horizon generation. https://doi.org/10.48550/ARXIV.2403.05313

Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, Chi EH, Le QV, Zhou D (2022) Chain-of-thought prompting elicits reasoning in large language models. In: Proceedings of the 36th international conference on neural information processing systems, Curran, Red Hook, NY, USA, NIPS '22, pp 24824–24837. https://doi.org/10.5555/3600270.3602070

Wei Z, Huang D, Zhang J, Song C, Zhang S, Zhang J, Li Z, Jiang K, Li R, Duan Q (2024) GARAG: A general adaptive question-answering system based on RAG. In: Proceedings of the 2024 international conference on cloud computing and big data, Association for Computing Machinery, New York, ICCBD '24, pp 442–447. https://doi.org/10.1145/3695080.3695156

White RW (2024) Advancing the search frontier with AI agents. Commun ACM 67(9):54–65. https://doi.org/10.1145/3655615

Wu Y, Tang B, Xi C, Yu Y, Wang P, Liu Y, Kuang K, Deng H, Li Z, Xiong F, Hu J, Cheng P, Wang Z, Wang Y, Luo Y, Yang M (2024) Xinyu: An efficient (LLM)-based system for commentary generation. In: Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining. Association for Computing Machinery, New York, KDD '24, pp 6003–6014. https://doi.org/10.1145/3637528.3671537

Yu H, Gan A, Zhang K, Tong S, Liu Q, Liu Z (2024) Evaluation of retrieval-augmented generation: A survey. In: Zhu W (ed) Proceedings of the 2024 international conference on cloud computing and big data, New York, ICCBD '24, pp 442–447. https://doi.org/10.1145/3695080.3695156

Zhang T, Patil SG, Jain N, Shen S, Zaharia M, Stoica I, Gonzalez JE (2024) RAFT: Adapting language model to domain specific RAG. https://doi.org/10.48550/arXiv.2403.10131

Zhao P, Zhang H, Yu Q, Wang Z, Geng Y, Fu F, Yang L, Zhang W, Jiang J, Cui B (2024) Retrieval-augmented generation for AI-generated content: A survey. https://doi.org/10.48550/ARXIV.2402.19473