A Survey on Retrieval-Augmented Text Generation for Large Language Models

Anonymous ACL submission

Abstract

Retrieval-Augmented Generation (RAG) merges retrieval methods with deep learning advancements to address the static limitations of large language models (LLMs) by enabling the dynamic integration of up-to-date external information. This methodology, focusing primarily on the text domain, provides a cost-effective solution to the generation of plausible but incorrect responses by LLMs, thereby enhancing the accuracy and reliability of their outputs through the use of real-world data. As RAG grows in complexity and incorporates multiple concepts that can influence its performance, this paper organizes the RAG paradigm into four categories: pre-retrieval, retrieval, post-retrieval, and generation, offering a detailed perspective from the retrieval viewpoint. It outlines RAG's evolution and discusses the field's progression through the analysis of significant studies. Additionally, the paper introduces evaluation methods for RAG, addressing the challenges faced and proposing future research directions. By offering an organized framework and categorization, the study aims to consolidate existing research on RAG, clarify its technological underpinnings, and highlight its potential to broaden the adaptability and applications of LLMs.

1 Introduction

The advent of ChatGPT has significantly impacted both academia and industry due to its interactive capabilities and widespread application, establishing itself as a leading artificial intelligence tool. At the core of ChatGPT is the large language model (LLM) GPT-4, as detailed by (OpenAI et al., 2023), which has seen numerous enhancements to its predecessors, showcasing exceptional abilities in a variety of Natural Language Processing (NLP) tasks. Despite these advancements, the adoption of LLMs has highlighted several critical issues primarily due to their reliance on extensive datasets. This reliance restricts their ability to incorporate new information post-training, leading to three primary challenges. First, the focus on broad and general data to maximize accessibility and applicability results in subpar performance in specialized areas. Second, the rapid creation of online data, combined with the significant resources required for data annotation and model training, hinders LLMs' ability to stay updated. Third, LLMs are susceptible to generating convincing yet inaccurate responses, known as "hallucinations", which can mislead users.

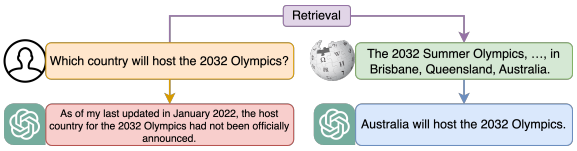

Figure 1: An example of RAG benefits ChatGPT resolves questions that cannot be answered beyond the scope of the training data and generates correct results.

Addressing these challenges is crucial for LLMs to be effectively utilized across various domains. A promising solution is the integration of Retrieval-Augmented Generation (RAG) technology, which supplements models by fetching external data in response to queries, thus ensuring more accurate and current outputs. Figure 1 illustrates how RAG can enable ChatGPT to provide precise answers beyond its initial training data.

Since its introduction by Lewis et al. (Lewis et al., 2020b) in 2020, RAG technology has undergone significant advancements, particularly influenced by ChatGPT's success. However, there is a noticeable gap in the literature regarding a thorough analysis of RAG's mechanisms and the progress made by subsequent studies. Furthermore, the field is characterized by diverse research focuses and the use of ambiguous terminology for similar methods, leading to confusion. This paper aims to clarify these aspects by offering a structured overview of RAG, categorizing various methods, and deliver-

ing an in-depth understanding of this research area. This survey will primarily focus on textual applications of RAG, reflecting the current emphasis of research efforts in this area.

RAG combines retrieval methods and advanced deep learning to address two main questions: effectively retrieving relevant information and generating accurate responses. The workflow of RAG is outlined in Section 2, categorizing the methodologies into pre-retrieval, retrieval, post-retrieval, and generation phases. These sections, from 3 to 6, provide an in-depth analysis of the technologies within these phases. Section 7 looks at future research directions, focusing on text-based studies and touching on image and multimodal data. The conclusion is presented in Section 8. Appendix A discusses the evaluation techniques for RAG. Appendices B and C provide summaries of the studies reviewed and the retrievers and generators employed, respectively.

The contributions of this paper are threefold: This paper offers a comprehensive framework for understanding the RAG domain, identifying areas for improvement and challenges for future research. It provides a detailed analysis of RAG's core technologies, examining their strengths in addressing retrieval and generation. Additionally, it introduces the evaluation methods used in RAG research, highlighting current challenges and suggesting promising directions for future studies.

2 RAG Framework

The hallucinations are largely attributed to LLMs' inability to access up-to-date information. This limitation stems from the models' reliance on their training datasets. RAG proposes a solution to this issue by supplementing the LLM's training data with current information from external sources through a retrieval model, thereby enabling the generation of accurate responses. RAG presents a more cost-effective alternative to the extensive training and fine-tuning processes typically required for LLMs. It allows for the dynamic incorporation of fresh information via traditional retrieval methods or pre-trained LMs, without the need to directly integrate this new data into the LLM. This feature makes RAG both flexible and scalable, facilitating its application across different LLMs for various purposes. The information retrieved through RAG is derived from real-world data, authored by humans, which not only simplifies the generation process but also increases the reliability of the generated responses. Figure 2 represents the unified RAG framework with basic workflow and paradigm.

Research by Khandelwal et al. (Khandelwal et al., 2020) demonstrates that accessing relevant information from the training dataset itself can significantly improve LLM performance, highlighting the effectiveness of RAG. Over time, RAG has evolved from a means of providing supplementary information to enabling multiple interactions between the retrieval and generation components. This involves conducting several rounds of retrieval to refine the accuracy of the information retrieved and iteratively improve the quality of the generated output. Platforms such as LangChain[1] and LlamaIndex[2] have modularized the RAG approach, enhancing its adaptability and expanding its range of applications. Despite these platforms employing diverse methodologies to tackle different aspects of RAG—from multiple search iterations to iterative generation—they maintain adherence to the fundamental RAG workflow. This consistency is crucial for understanding their operation and pinpointing opportunities for further development.

2.1 Basic RAG Workflow

The foundational workflow of RAG begins with the creation of an index comprising external sources. This index serves as the basis for retrieving relevant information through a retriever model based on a specific query. The final step involves a generator model, which combines the retrieved information with the query to produce the desired output.

2.1.1 Indexing

Efficient retrieval begins with comprehensive indexing, where data preparation is key. This stage involves text normalization processes such as tokenization, stemming, and the removal of stop words to enhance the text's suitability for indexing (Manning et al., 2008). Text segments are then organized into sentences or paragraphs to facilitate more focused searches, allowing for the pinpointing of segments containing pertinent keywords. The integration of deep learning has revolutionized indexing through the use of pretrained LMs for generating semantic vector representations of texts. These vectors are stored, enabling rapid and precise retrieval from extensive data collections, significantly enhancing retrieval efficiency.

[1] https://www.langchain.com
[2] https://www.llamaindex.ai

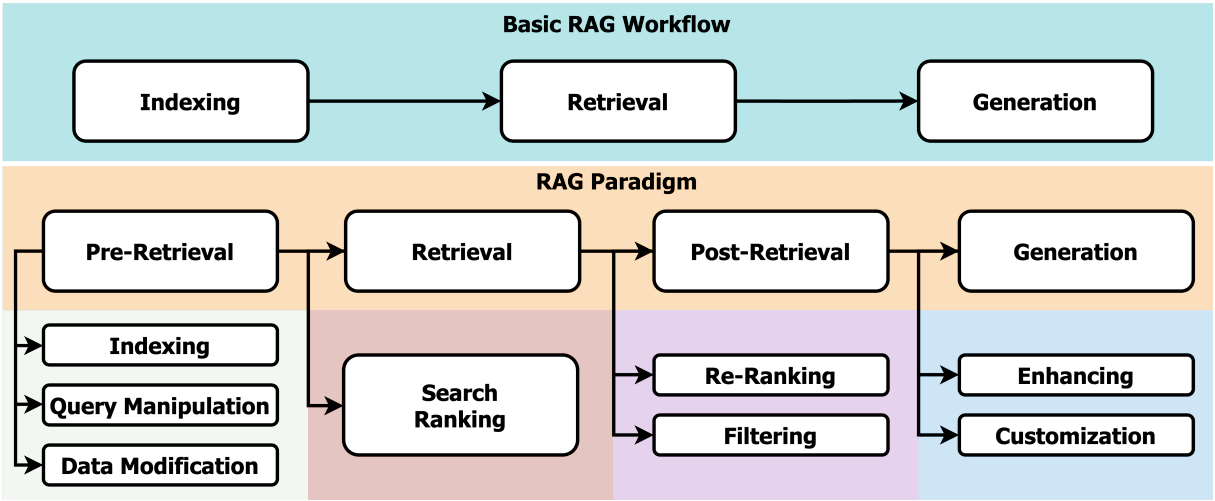

Figure 2: An unified RAG framework with basic workflow and paradigm.

2.1.2 Retrieval

While traditional retrieval methods, such as the BM25 algorithm (Robertson and Zaragoza, 2009), focus on term frequency and presence for document ranking, they often overlook the semantic information of queries. Current strategies leverage pre-trained LMs like BERT (Devlin et al., 2019), which capture the semantic essence of queries more effectively. These models improve search accuracy by considering synonyms and the structure of phrases, thereby refining document ranking through the detection of semantic similarities. This is typically achieved by measuring vector distances between documents and queries, combining traditional retrieval metrics with semantic understanding to yield search results that are both relevant and aligned with user intent.

2.1.3 Generation

The generation phase is tasked with producing text that is both relevant to the query and reflective of the information found in the retrieved documents. The usual method involves concatenating the query with the retrieved information, which is then fed into an LLM for text generation (Li et al., 2022). Although ensuring the generated text's alignment and accuracy with the retrieved content presents challenges, it is also essential to strike a balance between adhering closely to the source material and infusing the output with creativity. The generated text should accurately convey the information from the retrieved documents and align with the query's intent, while also offering the flexibility to introduce new insights or perspectives not explicitly contained within the retrieved data.

2.2 RAG Paradigm

The RAG paradigm organizes research within the domain, offering a straightforward yet robust framework to enhance LLM performance. Central to RAG is its search mechanism, crucial for generating high-quality outcomes. Therefore, this paradigm is structured into four main phases from a retrieval perspective: pre-retrieval, retrieval, post-retrieval, and generation. Both single-hop and multi-hop retrieval approaches, encompassing iterative retrieve-generate cycles, follow this four-phase structure. Figure 3 is the taxonomy tree of RAG's core techniques.

2.2.1 Pre-Retrieval

The pre-retrieval phase of retrieval-augmented generation lays the foundation for successful data and query preparation, ensuring efficient information retrieval. This phase includes essential tasks to prepare for effective data access.

Indexing The process starts with indexing, which establishes an organized system to enable fast and accurate retrieval of information. The specificity of indexing depends on the task and data type. For example, sentence-level indexing is beneficial for question-answering systems to precisely locate answers, while document-level indexing is more appropriate for summarizing documents to understand their main concepts and ideas.

Query Manipulation After indexing, query manipulation is performed to adjust user queries for a better match with the indexed data. This involves query reformulation, which rewrites the query to align more closely with the user's intention; query

expansion, which extends the query to capture more relevant results through synonyms or related terms; and query normalization, which resolves differences in spelling or terminology for consistent query matching.

Data Modification Data modification is also critical in enhancing retrieval efficiency. This step includes preprocessing techniques like removing irrelevant or redundant information to improve the quality of results and enriching the data with additional information such as metadata to boost the relevance and diversity of the retrieved content.

2.2.2 Retrieval

Search & Ranking The retrieval stage is the combination of search and ranking. It focuses on selecting and prioritizing documents from a dataset to enhance the quality of the generation model's outputs. This stage employs search algorithms to navigate through the indexed data, finding documents that match a user's query. After identifying relevant documents, the process of initially ranking these documents starts to sort them according to their relevance to the query.

2.2.3 Post-Retrieval

The post-retrieval phase serves to refine the initially retrieved documents to improve the quality of text generation. This phase consists of re-ranking and filtering, each aimed at optimizing the document selection for the final generation task.

Re-Ranking In the re-ranking step, the documents previously retrieved are reassessed, scored, and reorganized. The objective is to more accurately highlight the documents most relevant to the query and diminish the importance of the less relevant ones. This step involves incorporating additional metrics and external knowledge sources to enhance precision. In this context, pre-trained models with superior accuracy but lower efficiency can be effectively employed due to the limited set of candidate documents available.

Filtering Filtering aims to remove documents that fail to meet specified quality or relevance standards. This can be done through several approaches, such as establishing a minimum relevance score threshold to exclude documents below a certain relevance level. Furthermore, the use of feedback from users or prior relevance evaluations assists in adjusting the filtering process, guaranteeing that only the most relevant documents are retained for text generation.

2.2.4 Generation

The generation stage is a crucial component of the RAG process, responsible for leveraging retrieved information to enhance the quality of the generated response. This stage encompasses several sub-steps aimed at producing content that is readable, engaging, and informative.

Enhancing At the heart of the generation phase is the enhancement step, where the objective is to merge the retrieved information with the user's query to create a coherent and relevant response. This includes the process of elaboration, adding extra details to the retrieved content to enrich it. Efforts are focused on improving the output's quality by increasing its clarity, coherence, and stylistic appeal through methods such as rephrasing and restructuring. Information from various sources is combined to offer a comprehensive perspective, and verification is conducted to ensure the accuracy and relevance of the content.

Customization Customization is an optional step, involving the adjustment of content to align with the user's specific preferences or the context of the request. This tailoring includes adapting the content to meet the needs of the target audience or the format in which it will be presented and condensing the information to succinctly convey the essence of the content. The process also entails creating summaries or abstracts that emphasize the key points or arguments, ensuring the output is both informative and concise.

3 Pre-Retrieval

3.1 Indexing

The integration of the k-nearest neighbor (kNN) algorithm with pre-trained neural LMs, as demonstrated in kNN-LMs (Khandelwal et al., 2020), represents significant progress in language modeling. This method employs a datastore created from collections of texts, enabling the dynamic retrieval of contextually relevant examples to improve perplexity without necessitating additional training.

Known for its efficiency, FAISS (Johnson et al., 2021) has been adopted in many studies for indexing purposes (Khandelwal et al., 2020; Lewis et al., 2020b; Khattab et al., 2022). Some research integrates enhancements like the Hierarchical Naviga-

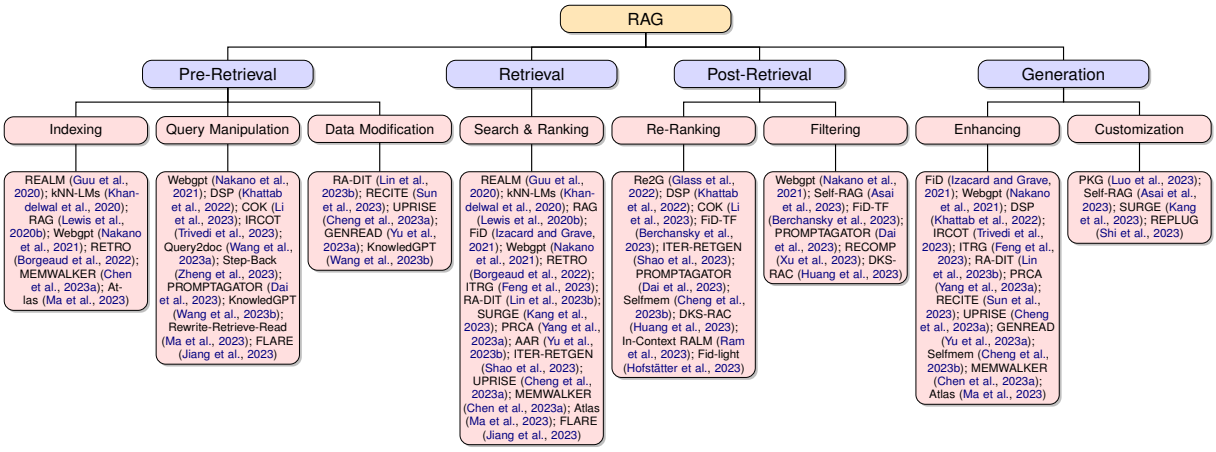

Figure 3: Taxonomy tree of RAG's core techniques

ble Small World (HNSW) approximation (Malkov and Yashunin, 2020) to achieve faster retrieval (Lewis et al., 2020b). In addition, alternative tools like utilizing the Bing API [3] for indexing based on actual user search histories as outlined in Webgpt (Nakano et al., 2021), illustrate the variety of indexing techniques under investigation.

Furthermore, MEMWALKER (Chen et al., 2023a) introduces an innovative method to overcome the limitations of context window size in LLMs by creating a memory tree from the input text. This tree is formed by initially segmenting the text into smaller pieces and then summarizing these segments into a hierarchical structure of summary nodes, facilitating efficient indexing and management of large volumes of information.

3.2 Query Manipulation

Studies such as FiD (Izacard and Grave, 2021), COK(Li et al., 2023), and Query2doc (Wang et al., 2023a) emphasize the significance of creating new queries or refining existing ones to achieve more pertinent retrieval results. These research efforts highlight the necessity of efficiently gathering evidence from multiple passages and tailoring queries to suit various knowledge sources, whether structured or unstructured. Techniques ranging from the creation of pseudo-documents to enhance queries have shown to bolster retrieval performance across diverse information retrieval datasets.

Further exploration into query manipulation has been conducted by Step-Back (Zheng et al., 2023) and PROMPTAGATOR (Dai et al., 2023), which focus on abstracting high-level concepts or utilizing LLMs for prompt-based query generation. These strategies strive to better align queries with the retrieval system's functionality by rephrasing tasks into more generalized versions or crafting task-specific queries from limited examples. Such methodologies enhance the consistency between queries and indexed data, facilitating the retrieval of more pertinent and insightful information.

Moreover, KnowledGPT (Wang et al., 2023b) and Rewrite-Retrieve-Read (Ma et al., 2023) introduce approaches for query manipulation through "program of thought" prompting and innovative query rewriting techniques. KnowledGPT innovates by generating code to interface with knowledge bases, converting user queries into structured search commands. In contrast, Rewrite-Retrieve-Read utilizes a trainable compact LM for query reformulation, adjusting them to more effectively reflect the user's intent and context.

Lastly, FLARE (Jiang et al., 2023) presents a strategy based on confidence for query formulation, which focuses on crafting queries that precisely reflect the information needs. This method incorporates the use of generated sentences or fragments thereof as a foundation for search queries. By opting to directly use sentences, obscuring tokens of low confidence, or formulating explicit questions, this approach aims to boost the efficiency of the retrieval process, ensuring that the retrieved information faithfully satisfies the requirements of the generation process.

3.3 Data Modification

RA-DIT (Lin et al., 2023b) and RECITE (Sun et al., 2023) emphasize enhancements through internal data modifications. RA-DIT distinguishes between

[3] https://www.microsoft.com/en-us/bing/apis/bing-web-search-api

fine-tuning datasets for LLMs and retrievers, aiming to bolster the LLM's contextual comprehension and the retriever's ability to align with queries. RECITE, on the other hand, utilizes passage hints and synthetic question-passage pairs to increase the variety and relevance of its generated recitations and responses. This approach seeks to broaden the model's knowledge base and improve its response accuracy.

UPRISE (Cheng et al., 2023a) and GENREAD (Yu et al., 2023a) target the refinement of external data. UPRISE converts raw task data into a structured format and refines the selection of prompts to enhance retrieval outcomes. In contrast, the Clustering-Based Prompts method employed by GENREAD generates documents from questions and clusters them to eliminate irrelevant data, enriching the input with varied contextual insights. This technique aims to improve the performance of the generative model by providing it with a richer set of information.

Furthermore, KnowledGPT (Wang et al., 2023b) is dedicated to augmenting raw text data with structured, semantically rich information through entity linking. This enrichment process not only structures the data more cohesively and makes it more amenable to queries but also boosts the model's retrieval efficiency. It leverages precise, linked knowledge to enhance the model's understanding and its ability to generate relevant responses, thereby improving its overall performance.

4 Retrieval

4.1 Search & Ranking

Atlas (Izacard et al., 2023) investigates few-shot learning approaches, including Attention Distillation and Perplexity Distillation, to steer the retriever toward retrieving more relevant documents. IRCOT (Trivedi et al., 2023) integrates retrieval with reasoning to improve the effectiveness of retrieval. SURGE (Kang et al., 2023) employs a subgraph retriever to extract relevant subgraphs from a knowledge graph, while AAR (Yu et al., 2023b) modifies search preferences to help LLMs in fetching pertinent documents.

PRCA (Yang et al., 2023a) focuses on employing domain-specific abstractive summarization to extract relevant and context-rich information from documents, using a supervised learning strategy to prioritize content crucial for accurate query responses. Meanwhile, MEMWALKER (Chen et al., 2023a) leverages an internal search and ranking mechanism in the constructed memory tree to identify pertinent information for long-context question answering. Additionally, the Confidence-based Active Retrieval approach of FLARE (Jiang et al., 2023) dynamically triggers information retrieval based on the confidence levels of generated sentences, utilizing the insight that low-confidence tokens signal a need for external knowledge.

5 Post-Retrieval

5.1 Re-Ranking

Re2G (Glass et al., 2022) introduces a sequence-pair classification approach for re-ranking, utilizing a BERT transformer to simultaneously analyze the query and passage. This interaction model, employing cross-attention between sequences, offers a contrast to the representation model typically used in initial retrieval phases. PROMPTAGATOR (Dai et al., 2023) also employs a cross-attention model for re-scoring. Its "Lift Yourself Up" strategy iteratively selects the best candidate from a pool for further generation rounds, progressively improving content quality via self-generated content.

Re-ranking is also a significant focus of In-Context RALM (Ram et al., 2023). Two approaches to reranking are explored: zero-shot reranking using language models and predictive reranking through trained models. This step is aimed at refining the selection of documents based on their expected utility for improving language model performance. ITER-RETGEN (Shao et al., 2023), in particular, leverages knowledge distillation from the re-ranker to the dense retriever, fine-tuning retrieval efforts based on relevance signals from LLM outputs. This optimization of the retrieval model aims to more accurately capture query nuances, thereby improving document selection.

DKS-RAC (Huang et al., 2023) presents the Dense Knowledge Similarity (DKS) for aligning the knowledge between answers and retrieved passages at the sequence level. This approach is categorized under re-ranking due to its direct impact on passage selection based on knowledge similarity, refining the match between queries and documents.

FiD-light (Hofstätter et al., 2023) introduces a listwise autoregressive re-ranking method that employs source pointers to optimize the ranking order. This method maintains a link between the generated text and source passages, enabling a more structured generation process. By incorporating

textual citations within the model's output as pointers to relevant information sources, this approach facilitates an organized retrieval and generation process, enhancing the overall coherence and relevance of the generated content.

5.2 Filtering

COK (Li et al., 2023) presents the Progressive Rationale Correction technique, aimed at iteratively refining rationales with retrieved knowledge. This method constitutes a continuous optimization process, significantly enhancing the relevance and quality of information used in content generation. Self-RAG (Asai et al., 2023) introduces a self-reflection mechanism to efficiently filter out irrelevant content. By employing critique tokens, this approach evaluates the relevance, supportiveness, and utility of retrieved passages, ensuring the integration of only high-quality information into the content generation process.

Additionally, FiD-TF (Berchansky et al., 2023) and RECOMP (Xu et al., 2023) are dedicated to the removal of irrelevant or redundant tokens and information from retrieved documents. FiD-TF employs a dynamic mechanism to identify and eliminate unnecessary tokens, enhancing the efficiency of information processing. RECOMP, on the other hand, compresses documents into concise summaries, focusing on selecting only the most pertinent content for the generation process. These methods streamline the content generation workflow by ensuring that only relevant and supportive information is utilized, thereby improving the overall quality and relevance of the generated content.

6 Generation

6.1 Enhancing

DSP (Khattab et al., 2022) introduces a framework designed to generate multiple retrieval queries to summarize and answer questions, drawing upon information aggregated from various passages. This framework employs CombSUM (Fox and Shaw, 1994) to calculate a cumulative probability score for passages across different retrieval lists, facilitating the compilation of a comprehensive response from multiple sources.

PRCA (Yang et al., 2023a) outlines a Reward-Driven Stage, wherein the distilled context is refined based on feedback from the generator. Utilizing reinforcement learning, this stage adjusts the parameters of PRCA according to the rewards received for providing relevant context. The objective is to fine-tune the extracted context to meet the specific requirements of the generator, thereby optimizing the generation process.

REPLUG (Shi et al., 2023) proposes a method for prepending retrieved documents to the input context before the final prediction by the black-box LM. It introduces an ensemble strategy to encode retrieved documents in parallel, overcoming the limitations of LM context length and enhancing accuracy through the allocation of increased computational resources. This approach improves the generation process by ensuring that the LM has access to a broader range of relevant information.

RECITE (Sun et al., 2023) implements a self-consistency technique, which involves generating multiple recitations independently and employing a plurality/majority vote system to determine the most appropriate answer. This method is designed to increase the reliability and accuracy of the answers, thereby improving the quality and credibility of the output.

6.2 Customization

The PKG framework, introduced by (Luo et al., 2023), represents an approach to customizing the output of LMs. By generating background knowledge internally using a pre-trained model, PKG eliminates the need for traditional external retrieval processes. This method directly integrates domain- or task-specific knowledge into the generation step, significantly enhancing the LM's capacity to produce responses that are specifically tailored to the given context or requirements.

Self-RAG (Asai et al., 2023) offers a strategy that incorporates reflection tokens within a customizable decoding algorithm. This technique permits dynamic adjustment of the model's retrieval and generation behaviors based on the specific task, facilitating more versatile response generation. Depending on the requirements, this approach can be tuned for accuracy or creativity, providing flexibility in generating outputs that meet diverse needs.

SURGE (Kang et al., 2023) achieves customization through the application of graph-text contrastive learning. This method ensures that the generated dialogue responses are in tight alignment with the knowledge contained in the retrieved subgraph, yielding responses that are specific, relevant, and deeply rooted in the dialogue context. By maintaining consistency between the retrieved knowledge and the generated text, SURGE is capable

of producing outputs that precisely reflect the detailed knowledge of the subgraph, enhancing the relevance and specificity of the responses.

7 Future Directions

7.1 Retrieval Quality

The integration of RAG into LLMs faces significant hurdles due to the vast amounts of unreliable information on the internet, including fake news. This presents a challenge for accurately retrieving useful knowledge, leading to the unreliable generation of responses by LLMs. As a result, LLMs may generate content based on incorrect information, undermining their reliability. Recent research efforts are directed towards enhancing retrieval methods to improve the efficiency, scalability, and effectiveness of LLMs in generating accurate and reliable responses.

Differentiable Search Indices (Tay et al., 2022) and (Bevilacqua et al., 2022) developed differentiable search indices that integrate the retrieval process within a Transformer model, enabling direct mapping of text queries to document identifiers. These approaches offer superior performance and potential for more efficient and scalable retrieval.

Generative Models for Search GERE (Chen et al., 2022a) can directly generate document titles and evidence sentences for fact-verification tasks. PARADE (Li et al., 2024) is a method for document reranking that aggregates passage representations into a unified document relevance score. Both of them demonstrate significant improvements in retrieval quality over traditional methods.

Fine-tuning Pre-trained Language Models RankT5 (Zhuang et al., 2023) is a model that fine-tunes the T5 framework specifically for text ranking. It leverages ranking losses to optimize performance metrics and exhibits promising zero-shot performance on out-of-domain data.

Noise Power (Cuconasu et al., 2024) provide a comprehensive analysis of the impact of IR components on RAG systems, revealing that the inclusion of irrelevant documents can significantly improve accuracy. It challenges conventional retrieval strategies and underscores the potential for developing specialized approaches that integrate retrieval with language generation models.

7.2 Multimodal RAG

The multimodal RAG domain has experienced significant growth, highlighting a pivotal advancement at the confluence of text and visual comprehension. The introduction of MuRAG (Chen et al., 2022b) marked a breakthrough by amalgamating textual and visual information for language generation, establishing a new standard for multimodal datasets. This model showcased the efficacy of utilizing a multimodal memory system to boost the accuracy in question-answering and reasoning tasks.

After MuRAG, studies such as REVEAL (Hu et al., 2023) and Re-Imagen (Chen et al., 2023c) have focused on enhancing visual question answering and text-to-image generation. They achieved this through the incorporation of dynamic retrieval mechanisms and the improvement of image fidelity, respectively. These advancements laid the groundwork for further models by researchers like Sarto et al. (Sarto et al., 2022) for image captioning, and Yuan et al. (Yuan et al., 2023) for text-to-audio generation, broadening the scope of RAG's application across different modalities and improving the quality and realism of the generated outputs. Furthermore, Re-ViLM (Yang et al., 2023b) refined image captioning capabilities through a retrieval-augmented visual language model. By fine-tuning model parameters and implementing innovative filtering strategies, it has made strides in producing more precise and contextually appropriate captions. By tapping into external resources, these models have provided significant enhancements over traditional benchmarks, highlighting the advantage of integrating diverse sources of knowledge.

8 Conclusions

In this paper, we have presented a comprehensive framework for understanding the RAG domain, highlighting its significance in enhancing the capabilities of LLMs. Through a structured overview of RAG, categorizing various methods, and an in-depth analysis of its core technologies and evaluation methods, this study illuminates the path for future research. It identifies crucial areas for improvement and outlines potential directions for advancing RAG applications, especially in textual contexts. This survey aims to elucidate the core concepts of the RAG field from a retrieval perspective, and it is intended to facilitate further exploration and innovation in the accurate retrieval and generation of information.

9 Limitations

This survey comprehensively examines existing RAG models, summarizing their core techniques into four main steps from a retrieval perspective. It recognizes that some methods may encompass multiple steps and that decoupling these steps could potentially obscure their intrinsic connections. Nevertheless, the primary objective is to simplify the complexity of the approach, clearly delineating the specific problems it addresses. This allows for a clearer identification of areas ripe for further optimization and improvement. Despite the thorough investigation, the rapid evolution of the field and page limits mean that certain aspects might not have been fully analyzed and explored, or recent developments could have been missed. While the paper references evaluation methods that can aid in the development of RAG, it also acknowledges mature tools like LangChain and LlamaIndex as useful resources. However, the focus of this survey is not on detailing the evaluation pipeline or how these tools are specifically used, but rather on illustrating how evaluation aspects can support the advancement of RAG. This choice highlights an area for future work, emphasizing the importance of methodological clarity and the application of evaluation tools in refining and enhancing RAG models.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. *arXiv*, abs/2310.11511.

Moshe Berchansky, Peter Izsak, Avi Caciularu, Ido Dagan, and Moshe Wasserblat. 2023. Optimizing Retrieval-augmented Reader Models via Token Elimination. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1506–1524. Association for Computational Linguistics.

Michele Bevilacqua, Giuseppe Ottaviano, Patrick S. H. Lewis, Scott Yih, Sebastian Riedel, and Fabio Petroni. 2022. Autoregressive Search Engines: Generating Substrings as Document Identifiers. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2022. Improving Language Models by Retrieving from Trillions of Tokens. In *International Conference on Machine Learning (ICML)*, pages 2206–2240.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Conference on Neural Information Processing Systems (NeurIPS)*, volume abs/2005.14165.

Howard Chen, Ramakanth Pasunuru, Jason Weston, and Asli Celikyilmaz. 2023a. Walking Down the Memory Maze: Beyond Context Limit through Interactive Reading. *arXiv*, abs/2310.05029.

Jiangui Chen, Ruqing Zhang, Jiafeng Guo, Yixing Fan, and Xueqi Cheng. 2022a. Gere: Generative Evidence Retrieval for Fact Verification. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2023b. Benchmarking large language models in retrieval-augmented generation. *arXiv*, abs/2309.01431.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, and others. 2021. Evaluating large language models trained on code. *arXiv*, abs/2107.03374.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. 2022b. Murag: Multimodal Retrieval-Augmented Generator for Open Question Answering over Images and Text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W. Cohen. 2023c. Re-Imagen: Retrieval-Augmented Text-to-Image Generator. In *International Conference on Learning Representations (ICLR)*.

Zhihong Chen, Feng Jiang, Junying Chen, Tiannan Wang, Fei Yu, Guiming Chen, Hongbo Zhang, Juhao Liang, Chen Zhang, Zhiyi Zhang, and others. 2023d. Phoenix: Democratizing chatgpt across languages. *arXiv*, abs/2304.10453.

Daixuan Cheng, Shaohan Huang, Junyu Bi, Yuefeng Zhan, Jianfeng Liu, Yujing Wang, Hao Sun, Furu Wei, Weiwei Deng, and Qi Zhang. 2023a. Uprise: Universal Prompt Retrieval for Improving Zero-Shot Evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12318–12337. Association for Computational Linguistics.

Xin Cheng, Di Luo, Xiuying Chen, Lemao Liu, Dongyan Zhao, and Rui Yan. 2023b. Lift Yourself Up: Retrieval-augmented Text Generation with Self-Memory. In *Thirty-seventh Conference on Neural Information Processing Systems*, volume abs/2305.02437.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. Palm: Scaling Language Modeling with Pathways. *Journal of Machine Learning Research (JMLR)*, 24:240:1–240:113.

Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, W. Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, S. Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, E. Chi, J. Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling Instruction-Finetuned Language Models. *arXiv*, abs/2210.11416.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The Power of Noise: Redefining Retrieval for RAG Systems. *arXiv*, abs/2401.14887.

Zhuyun Dai, Vincent Y. Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B. Hall, and Ming-Wei Chang. 2023. Promptagator: Few-shot Dense Retrieval From 8 Examples. In *International Conference on Learning Representations (ICLR)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North*, pages 4171–4186. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-Powered Conversational Agents. In *International Conference on Learning Representations (ICLR)*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General Language Model Pretraining with Autoregressive Blank Infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Hady ElSahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon S. Hare, Frédérique Laforest, and Elena Simperl. 2018. T-REx: A Large Scale Alignment of Natural Language with Knowledge Base Triples. In *International Conference on Language Resources and Evaluation (LREC)*.

Zhangyin Feng, Xiaocheng Feng, Dezhi Zhao, Maojin Yang, and Bing Qin. 2023. Retrieval-generation synergy augmented large language models. *arXiv*, abs/2310.05149.

Edward A. Fox and Joseph A. Shaw. 1994. Combination of multiple searches. In *TREC-2: Text retrieval conference*, 500215, pages 105–108.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple Contrastive Learning of Sentence Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910. Association for Computational Linguistics.

Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Naik, Pengshan Cai, and Alfio Gliozzo. 2022. Re2g: Retrieve, Rerank, Generate. In *Proceedings of the 2022 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2701–2715. Association for Computational Linguistics.

Simon Gottschalk and Elena Demidova. 2018. *EventKG: A Multilingual Event-Centric Temporal Knowledge Graph*. Springer International Publishing.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Retrieval Augmented Language Model Pre-Training. In *International Conference on Machine Learning (ICML)*, pages 3929–3938.

William L. Hamilton. 2020. *Graph representation learning*. Springer International Publishing.

Sebastian Hofstätter, Jiecao Chen, Karthik Raman, and Hamed Zamani. 2023. Fid-light: Efficient and effective retrieval-augmented text generation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1437–1447.

Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A. Ross, and Alireza Fathi. 2023. Reveal: Retrieval-Augmented Visual-Language Pre-Training with Multi-Source Multimodal Knowledge Memory. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23369–23379. IEEE.

Jie Huang, Hanyin Shao, Kevin Chen-Chuan Chang, Jinjun Xiong, and Wen-mei Hwu. 2022. Understanding Jargon: Combining Extraction and Generation for Definition Modeling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Wenyu Huang, Mirella Lapata, Pavlos Vougiouklis, Nikos Papasarantopoulos, and Jeff Z Pan. 2023. Retrieval Augmented Generation with Rich Answer Encoding. *Proc. of IJCNLP-AACL*, 2023.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised Dense Information Retrieval with Contrastive Learning. *Transactions on Machine Learning Research (TMLR)*, 2022.

Gautier Izacard and Edouard Grave. 2021. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880. Association for Computational Linguistics.

Gautier Izacard, Patrick S. H. Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot Learning with Retrieval Augmented Language Models. *Journal of Machine Learning Research (JMLR)*, 24:251:1–251:43.

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active Retrieval Augmented Generation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7969–7992.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. Triviaqa: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611. Association for Computational Linguistics.

Minki Kang, Jin Myung Kwak, Jinheon Baek, and Sung Ju Hwang. 2023. Knowledge Graph-Augmented Language Models for Knowledge-Grounded Dialogue Generation. *arXiv*, abs/2305.18846.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through Memorization: Nearest Neighbor Language Models. In *International Conference on Learning Representations (ICLR)*.

O. Khattab, Keshav Santhanam, Xiang Lisa Li, David Leo Wright Hall, Percy Liang, Christopher Potts, and M. Zaharia. 2022. Demonstrate-Search-Predict: Composing retrieval and language models for knowledge-intensive NLP. *arXiv*, abs/2212.14024.

Omar Khattab and Matei Zaharia. 2020. Colbert - Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 39–48. ACM.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy,

Veselin Stoyanov, and Luke Zettlemoyer. 2020a. Bart: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880. Association for Computational Linguistics.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Canjia Li, Andrew Yates, Sean MacAvaney, Ben He, and Yingfei Sun. 2024. Parade: Passage Representation Aggregation forDocument Reranking. *ACM Transactions on Information Systems*, 42(2):1–26.

Huayang Li, Yixuan Su, Deng Cai, Yan Wang, and Lemao Liu. 2022. A Survey on Retrieval-Augmented Text Generation. *arXiv*, abs/2202.01110.

Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Shafiq R. Joty, Soujanya Poria, and Lidong Bing. 2023. Chain-of-Knowledge: Grounding Large Language Models via Dynamic Knowledge Adapting over Heterogeneous Sources. *arXiv*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Sheng-Chieh Lin, Akari Asai, Minghan Li, Barlas Oguz, Jimmy Lin, Yashar Mehdad, Wen-tau Yih, and Xilun Chen. 2023a. How to Train Your Dragon: Diverse Augmentation Towards Generalizable Dense Retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6385–6400. Association for Computational Linguistics.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Rich James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, and others. 2023b. Ra-dit: Retrieval-augmented dual instruction tuning. *arXiv*, abs/2310.01352.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot Learning with Multilingual Generative Language Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Yi Liu, Lianzhe Huang, Shicheng Li, Sishuo Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023. Recall: A Benchmark for LLMs Robustness against External Counterfactual Knowledge. *arXiv*, abs/2311.08147.

Ziyang Luo, Can Xu, Pu Zhao, Xiubo Geng, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2023. Augmented Large Language Models with Parametric Knowledge Guiding. *arXiv*, abs/2305.04757.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query Rewriting in Retrieval-Augmented Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5303–5315. Association for Computational Linguistics.

Yu A. Malkov and D. A. Yashunin. 2020. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):824–836.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, and others. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv*, abs/2112.09332.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, and Yinfei Yang. 2022. Large Dual Encoders Are Generalizable Retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9844–9855. Association for Computational Linguistics.

Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. 2019. Multi-stage document ranking with BERT. *CoRR*, abs/1910.14424.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott

Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, and others. 2023. Gpt-4 Technical Report. *PREPRINT*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv*.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. Kilt: a Benchmark for Knowledge Intensive Language Tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544. Association for Computational Linguistics.

Filip Radlinski and Nick Craswell. 2010. Comparing the sensitivity of information retrieval metrics. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, page 667–674, New York, NY, USA. Association for Computing Machinery.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research (JMLR)*, 21:140:1–140:67.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-Context Retrieval-Augmented Language Models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Ori Ram, Gal Shachaf, Omer Levy, Jonathan Berant, and Amir Globerson. 2022. Learning to Retrieve Passages without Supervision. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3980–3990. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: Bm25 and Beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Jon Saad-Falcon, O. Khattab, Christopher Potts, and Matei Zaharia. 2023. Ares: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems. *arXiv*, abs/2311.09476.

Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2022. Retrieval-Augmented Transformer for Image Captioning. In *International Conference on Content-based Multimedia Indexing*. ACM.

ES Shahul, Jithin James, Luis Espinosa Anke, and S. Schockaert. 2023. Ragas: Automated Evaluation of Retrieval Augmented Generation. *arXiv*, abs/2309.15217.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Enhancing Retrieval-Augmented Large Language Models with Iterative Retrieval-Generation Synergy. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9248–9274. Association for Computational Linguistics.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *arXiv*, abs/2301.12652.

Zhiqing Sun, Xuezhi Wang, Yi Tay, Yiming Yang, and Denny Zhou. 2023. Recitation-Augmented Language Models. In *International Conference on Learning Representations (ICLR)*.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. 2023. Ul2: Unifying Language Learning Paradigms. In *International Conference on Learning Representations (ICLR)*.

Yi Tay, Vinh Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Prakash Gupta, Tal Schuster, William W. Cohen, and Donald Metzler. 2022. Transformer Memory as a Differentiable Search Index. In *Conference on Neural Information Processing Systems (NeurIPS)*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothee Lacroix, Baptiste Roziere, Naman Goyal, Eric Hambro, Faisal Azhar, and others. 2023a. Llama: Open and efficient foundation language models. *arXiv*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and others. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arxiv*, abs/2307.09288.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037. Association for Computational Linguistics.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Neural Information Processing Systems*, pages 5998–6008.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A Stickier Benchmark for General-Purpose Language Understanding Systems. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 3261–3275.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

Liang Wang, Nan Yang, and Furu Wei. 2023a. Query2doc: Query Expansion with Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9414–9423. Association for Computational Linguistics.

Xintao Wang, Qian Yang, Yongting Qiu, Jiaqing Liang, Qi He, Zhouhong Gu, Yanghua Xiao, and W. Wang. 2023b. Knowledgpt: Enhancing Large Language Models with Retrieval and Storage Access on Knowledge Bases. *arXiv*, abs/2308.11761.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, and others. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv*, abs/2211.05100.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. In *International Conference on Learning Representations (ICLR)*.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. 2023. Recomp: Improving Retrieval-Augmented LMs with Compression and Selective Augmentation. *arXiv*, abs/2310.04408.

Haoyan Yang, Zhitao Li, Yong Zhang, Jianzong Wang, Ning Cheng, Ming Li, and Jing Xiao. 2023a. Prca: Fitting Black-Box Large Language Models for Retrieval Question Answering via Pluggable Reward-Driven Contextual Adapter. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5364–5375. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380. Association for Computational Linguistics.

Zhuolin Yang, Wei Ping, Zihan Liu, Vijay Korthikanti, Weili Nie, De-An Huang, Linxi Fan, Zhiding Yu, Shiyi Lan, Bo Li, Mohammad Shoeybi, Ming-Yu Liu, Yuke Zhu, Bryan Catanzaro, Chaowei Xiao, and Anima Anandkumar. 2023b. Re-ViLM: Retrieval-Augmented Visual Language Model for Zero and Few-Shot Image Captioning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11844–11857. Association for Computational Linguistics.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023a. Generate rather than Retrieve: Large Language Models are

Strong Context Generators. In *International Conference on Learning Representations (ICLR)*.

Zichun Yu, Chenyan Xiong, Shi Yu, and Zhiyuan Liu. 2023b. Augmentation-Adapted Retriever Improves Generalization of Language Models as Generic Plug-In. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2421–2436. Association for Computational Linguistics.

Yi Yuan, Haohe Liu, Xubo Liu, Qiushi Huang, Mark D Plumbley, and Wenwu Wang. 2023. Retrieval-Augmented Text-to-Audio Generation. *arXiv*, abs/2309.08051.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, and others. 2022. Opt: Open pre-trained transformer language models. *arXiv*, abs/2205.01068.

Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, E. Chi, Quoc V. Le, and Denny Zhou. 2023. Take a Step Back: Evoking Reasoning via Abstraction in Large Language Models. *arXiv*, abs/2310.06117.

Honglei Zhuang, Zhen Qin, Rolf Jagerman, Kai Hui, Ji Ma, Jing Lu, Jianmo Ni, Xuanhui Wang, and Michael Bendersky. 2023. Rankt5: Fine-Tuning T5 for Text Ranking with Ranking Losses. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.

A Evaluation in RAG

To understand the effectiveness of LMs in generating more accurate, relevant, and robust responses by leveraging external knowledge, the evaluation of RAG systems has become a significant research area. With the popularity of dialogue-based interactions, recent works have been focused on assessing the performance of RAG models on such downstream tasks using established metrics like Exact Match (EM) and F1 scores. Furthermore, a wide array of datasets has been utilized for this purpose, including TriviaQA (Joshi et al., 2017), HotpotQA (Yang et al., 2018), FEVER (Thorne et al., 2018), Natural Questions (Kwiatkowski et al., 2019), Wizard of Wikipedia (Dinan et al., 2019), and T-REX (ElSahar et al., 2018).

However, evaluation solely from the perspective of downstream tasks falls short in addressing the evolving needs of RAG development. Recent research has introduced various frameworks and benchmarks that aim to evaluate these systems across multiple dimensions, including the quality of the generated text, the relevance of retrieved documents, and the model's resilience to misinformation, as shown in Table 1. These evaluations focus on assessing specific capabilities such as noise robustness, negative prompting, information integration, and counterfactual robustness, highlighting the complex challenges faced by RAG systems in practical applications. The continuous development of evaluation frameworks and metrics is crucial for advancing the field, broadening the applicability of RAG systems, and ensuring they meet the demands of a complex and evolving information landscape.

A.1 Retrieval-based Aspect

In information retrieval, the quality of search results is typically evaluated using standard metrics such as Mean Average Precision (MAP), Precision, Reciprocal Rank, and Normalized Discounted Cumulative Gain (NDCG) (Radlinski and Craswell, 2010; Reimers and Gurevych, 2019; Nogueira et al., 2019). These metrics primarily assess the relevance of retrieved documents to a given query.

Retrieval-based Metrics in RAG focus on the effectiveness of retrieving relevant information to support generation tasks. These include Accuracy, which measures the precision of retrieved documents in providing correct information for answering queries, and Rejection Rate (Chen et al., 2023b), assessing a system's ability to decline answering when no relevant information is found. Additionally, Error Detection Rate (Chen et al., 2023b) evaluates the model's capability to identify and disregard incorrect or misleading information from retrieved documents. Context Relevance is another essential metric, assessing the pertinence of the retrieved documents to the query. It's vital to ensure the information used to generate responses is directly related to the query's context. Faithfulness (Shahul et al., 2023) measures the accuracy with which the generated content reflects the information in the retrieved documents, ensuring that the generation process with no misinformation.

A.2 Generation-based Aspect

Evaluating the quality of text produced by LLMs involves analyzing their performance on various downstream tasks using standard metrics. These metrics assess linguistic quality, coherence, accuracy, and the extent to which the generated text reflects ground-truth data. Linguistic quality and coherence are evaluated through metrics such as

Evaluation Framework	Aspects	Methods	Metrics	Datasets
RAGAS (Shahul et al., 2023)	Quality of RAG Systems	Context Relevance	Extracted Sentences / Total Sentences	WikiEval [4]
		Answer Relevance	Average Cosine Similarity	
		Faithfulness	Supported Statements / Total Statements	
ARES (Saad-Falcon et al., 2023)	Improving RAGAS	Context Relevance	Confidence Intervals	KILT (Petroni et al., 2021) SuperGLUE (Wang et al., 2019)
		Answer Relevance		
		Answer Faithfulness		
RECALL (Liu et al., 2023)	Counterfactual Robustness	Response Quality	Accuracy (QA) BLEU, ROUGE-L (Generation)	EventKG (Gottschalk and Demidova, 2018) UJ (Huang et al., 2022)
		Robustness	Misleading Rate (QA) Mistake Reappearance Rate (Generation)	
RGB (Chen et al., 2023b)	Impact of RAG on LLMs	Noise Robustness	Accuracy	Synthetic Dataset including English and Chinese
		Negative Rejection	Rejection Rate	
		Information Integration	Accuracy	
		Counterfactual Robustness	Error Detection Rate Error Correction Rate	

Table 1: The Comparison of Different RAG Evaluation Frameworks

BLEU (Papineni et al., 2002), which measures fluency and similarity to human-produced text, and ROUGE-L (Lin, 2004), which quantifies the overlap with reference summaries to gauge the text's capacity to encapsulate main ideas and phrases. Accuracy and overlap with ground-truth data are gauged using metrics like EM and F1 Score, which respectively determine the percentage of answers that are entirely correct and offer a balanced assessment of precision and recall in retrieving relevant answers while minimizing inaccuracies.

Beyond these standard metrics, the evaluation may also incorporate task-specific criteria and novel metrics tailored to particular applications. For instance, in dialogue generation, perplexity and entropy are used to evaluate response diversity and naturalness. Additionally, metrics such as Misleading Rate and Mistake Reappearance Rate (Liu et al., 2023) gauge a model's ability to avoid misinformation and inaccuracies. Other specialized metrics include Answer Relevance (Shahul et al., 2023), assessing the precision of responses to queries; Kendall's tau (Saad-Falcon et al., 2023), for evaluating the accuracy of RAG system rankings; Micro-F1 (Saad-Falcon et al., 2023), which fine-tunes accuracy evaluation in tasks with multiple correct answers; and Prediction Accuracy, directly measuring the alignment of generated answers with expected responses, thereby offering a direct insight into a system's effectiveness in generating accurate content.

B The Comprehensive Summary of RAG

Table 2 offers a detailed examination of the RAG studies discussed in this paper. The analysis reveals that the majority of these studies have employed external data sources to enhance the information of LLMs. A preference for multiple-hop over single-hop retrieval was observed, indicating that iterative search rounds yield significantly better results. Furthermore, there is a notable focus on optimizing the retrieval phase, underscoring its critical role in the research. However, there appears to be a lack of studies focusing on customization in the generation stage, suggesting this area as a potential direction for future exploration.

C Retriever and Generator

In RAG, the primary components are the retriever and the generator. Table 3 provides a summary of the retrievers and generators utilized in the studies discussed in this paper. It is evident from the table that, although most generators employ advanced language models, a significant number of retrievers use the traditional BM25 for its efficiency. The method of retrieval is a fundamental issue in RAG, making it crucial to explore ways to enhance retrieval performance without sacrificing efficiency.

Research	Year	Retrieval Source		Multi-hop	Training	Pre-Retrieval			Retrieval	Post-Retrieval		Generation	
		Internal	External			Indexing	Query Manipulation	Data Modification	Search & Ranking	Re-Ranking	Filtering	Enhancing	Customization
REALM (Guu et al., 2020)	2020		✓		✓	✓			✓				
kNN-LMs (Khandelwal et al., 2020)	2020	✓	✓			✓			✓			✓	
RAG (Lewis et al., 2020b)	2020		✓		✓	✓			✓				
FiD (Izacard and Grave, 2021)	2021		✓						✓			✓	
Webgpt (Nakano et al., 2021)	2021		✓	✓	✓	✓	✓		✓		✓	✓	
Re2G (Glass et al., 2022)	2022	✓		✓	✓					✓			
RETRO (Borgeaud et al., 2022)	2022		✓	✓	✓	✓			✓				
DSP (Khattab et al., 2022)	2022		✓	✓			✓			✓		✓	
COK (Li et al., 2023)	2023		✓	✓			✓			✓			
IRCOT (Trivedi et al., 2023)	2023		✓	✓			✓					✓	
ITRG (Feng et al., 2023)	2023	✓	✓	✓					✓			✓	
PKG (Luo et al., 2023)	2023	✓											✓
RA-DIT (Lin et al., 2023b)	2023		✓	✓	✓			✓	✓			✓	
Self-RAG (Asai et al., 2023)	2023		✓		✓						✓		✓
SURGE (Kang et al., 2023)	2023	✓							✓				✓
FiD-TF (Berchansky et al., 2023)	2023		✓							✓	✓		
PRCA (Yang et al., 2023a)	2023		✓		✓				✓			✓	
REPLUG (Shi et al., 2023)	2023		✓		✓								✓
AAR (Yu et al., 2023b)	2023		✓		✓				✓				
Query2doc (Wang et al., 2023a)	2023	✓					✓						
Step-Back (Zheng et al., 2023)	2023		✓	✓			✓						
ITER-RETGEN (Shao et al., 2023)	2023		✓	✓					✓	✓			
RECITE (Sun et al., 2023)	2023	✓		✓	✓			✓				✓	
PROMPTAGATOR (Dai et al., 2023)	2023	✓		✓			✓			✓	✓		
UPRISE (Cheng et al., 2023a)	2023	✓		✓	✓			✓	✓			✓	
GENREAD (Yu et al., 2023a)	2023	✓						✓				✓	
KnowledGPT (Wang et al., 2023b)	2023		✓	✓			✓	✓					
Selfmem (Cheng et al., 2023b)	2023		✓	✓	✓					✓		✓	
MEMWALKER (Chen et al., 2023a)	2023		✓			✓			✓			✓	
RECOMP (Xu et al., 2023)	2023		✓		✓						✓		
Rewrite-Retrieve-Read (Ma et al., 2023)	2023		✓		✓		✓						
Atlas (Ma et al., 2023)	2023		✓	✓	✓	✓			✓	✓			
DKS-RAC (Huang et al., 2023)	2023		✓	✓	✓					✓	✓		
In-Context RALM (Ram et al., 2023)	2023		✓							✓			
Fid-light (Hofstätter et al., 2023)	2023		✓	✓						✓			
FLARE (Jiang et al., 2023)	2023		✓				✓		✓				

Table 2: The comprehensive summary of RAG studies. A ✓in the "Multi-hop" column signifies that the research involves multiple search rounds. Similarly, a ✓in the "Training" column indicates that the study included training phases. It is important to note that in this context, "Training" encompasses both initial model training and fine-tuning processes.

Research	Year	Retriever	Generator
REALM (Guu et al., 2020)	2020	BERT (Devlin et al., 2019)	Transformers (Vaswani et al., 2017)
kNN-LMs (Khandelwal et al., 2020)	2020	FAISS (Johnson et al., 2021)	Transformers
RAG (Lewis et al., 2020b)	2020	DPR (Karpukhin et al., 2020)	BART-Large (Lewis et al., 2020a)
FiD (Izacard and Grave, 2021)	2021	BM25 (Robertson and Zaragoza, 2009), DPR	T5 (Raffel et al., 2020)
Webgpt (Nakano et al., 2021)	2021	Bing	GPT-3 (Brown et al., 2020)
Re2G (Glass et al., 2022)	2022	BM25, DPR	BART
RETRO (Borgeaud et al., 2022)	2022	BERT	Transformer
DSP (Khattab et al., 2022)	2022	ColBERTv2 (Khattab and Zaharia, 2020)	GPT-3.5 (text-davinci-002)
COK (Li et al., 2023)	2023	LLaMA2-7B (Touvron et al., 2023b), ChatGPT (gpt-3.5-turbo-0613)	ChatGPT (gpt-3.5-turbo-0613)
IRCOT (Trivedi et al., 2023)	2023	BM25	GPT-3 (code-davinci-002), Flan-T5 (Chung et al., 2022)
ITRG (Feng et al., 2023)	2023	Atlas (Ma et al., 2023)	LLaMA 33B (Touvron et al., 2023a)
PKG (Luo et al., 2023)	2023	LLaMa-7B	InstructGPT-3.5 (text-davinic-002) (Ouyang et al., 2022)
RA-DIT (Lin et al., 2023b)	2023	DRAGON+ (Lin et al., 2023a)	LLama
Self-RAG (Asai et al., 2023)	2023	Contriever (Izacard et al., 2022)	Llama2 (7B and 13B) , GPT-4 (OpenAI et al., 2023)
SURGE (Kang et al., 2023)	2023	Graph Neural Networks (GNN) (Hamilton, 2020)	Transformers
FiD-TF (Berchansky et al., 2023)	2023	BM25, Sentence Transformers	T5
PRCA (Yang et al., 2023a)	2023	BM25, DPR, Contriver, SimCSE (Gao et al., 2021), SBERT (Reimers and Gurevych, 2019)	T5-large, Phoenix-7B (Chen et al., 2023d), Vicuna-7B (Peng et al., 2023), ChatGLM (Du et al., 2022), GPT-3.5
REPLUG (Shi et al., 2023)	2023	Contriever	GPT-3
AAR (Yu et al., 2023b)	2023	ANCE (Xiong et al., 2021), Contriever	Flan-T5, InstructGPT
Query2doc (Wang et al., 2023a)	2023	BM25, DPR	GPT-3 (text-davinci-003)
Step-Back (Zheng et al., 2023)	2023	PaLM-2L (Chowdhery et al., 2023)	PaLM-2L, GPT-4
ITER-RETGEN (Shao et al., 2023)	2023	Contriever	InstructGPT (text-davinci-003), Llama-2
RECITE (Sun et al., 2023)	2023		PaLM, UL2 (Tay et al., 2023), OPT (Zhang et al., 2022), Codex (Chen et al., 2021)
PROMPTAGATOR (Dai et al., 2023)	2023	T5	FLAN
UPRISE (Cheng et al., 2023a)	2023	GPT-Neo-2.7B (Black et al., 2021)	BLOOM-7.1B (Workshop et al., 2022), OPT-66B, GPT-3-175B
GENREAD (Yu et al., 2023a)	2023		InstructGPT
KnowledGPT (Wang et al., 2023b)	2023		GPT-4
Selfmem (Cheng et al., 2023b)	2023	BM25	XGLM (Lin et al., 2022), XLM-Rbase (Conneau et al., 2020)
MEMWALKER (Chen et al., 2023a)	2023	LLaMA-2	LLaMA-2
RECOMP (Xu et al., 2023)	2023	BM25	T5-Large
Rewrite-Retrieve-Read (Ma et al., 2023)	2023	Bing	T5-Large, ChatGPT(gpt-3.5-turbo), Vicuna-13B
Atlas (Ma et al., 2023)	2023	Contriever	T5
DKS-RAC (Huang et al., 2023)	2023	DPR	BART
In-Context RALM (Ram et al., 2023)	2023	BM25, BERT-base, Contriever, Spider (Ram et al., 2022)	GPT-2, GPT-Neo, GPT-J (Wang and Komatsuzaki, 2021), OPT, and LLaMA
Fid-light (Hofstätter et al., 2023)	2023	GTR-Base (Ni et al., 2022)	T5
FLARE (Jiang et al., 2023)	2023	BM25, Bing	GPT-3.5 (text-davinci-003)

Table 3: The summary of Retrievers and Generators. The retrieval models and pre-trained language models explicitly mentioned in these studies have been recorded.