A Comprehensive Literature Review with Self-Reflection

Literature Review

October 7, 2025

Abstract

This literature review provides a comprehensive analysis of recent research in the field. The review synthesizes findings from 211 research papers, identifying key themes, methodological approaches, and future research directions.

Contents

1 Introduction

1.1 The Rise of Large Language Models and Their Limitations

The advent of Large Language Models (LLMs) has marked a transformative era in artificial intelligence, showcasing unprecedented capabilities in natural language understanding and generation. These powerful generative AI systems, trained on vast corpora of text and code, have demonstrated remarkable proficiency in tasks ranging from complex question answering and summarization to creative content generation and code synthesis. However, despite their impressive performance, LLMs are inherently constrained by several critical limitations that significantly impact their reliability and trustworthiness, thereby underscoring the necessity for external knowledge augmentation mechanisms like Retrieval-Augmented Generation (RAG).

One of the most prominent limitations of LLMs is their propensity for **hallucination**, which refers to the generation of factually incorrect, nonsensical, or fabricated information presented as truth **??**. This issue arises because LLMs are trained to predict the most probable next token based on patterns in their training data, rather than possessing a true understanding of facts or the world. Consequently, when faced with queries outside their precise knowledge or when prompted ambiguously, they can confidently produce plausible-sounding but entirely false statements. For instance, an LLM might invent non-existent historical events, attribute quotes to the wrong individuals, or generate incorrect medical advice, posing significant risks in sensitive applications **?**. This inherent tendency to hallucinate undermines the factual accuracy and trustworthiness of LLM outputs, making them unreliable for knowledge-intensive tasks.

Another significant challenge is the **knowledge cutoff problem**. LLMs' knowledge is static, being confined to the information present in their training datasets up to a specific point in time **??**. They lack the ability to access or incorporate real-time, up-to-date information from the internet or proprietary databases beyond their last training update. This means that LLMs cannot provide current news, recent scientific discoveries, or evolving policy changes, rendering them obsolete for dynamic information environ-

ments. For example, an LLM trained in 2022 would be unable to answer questions about events from 2023 or 2024, leading to outdated or incomplete responses. This limitation severely restricts their utility in applications requiring contemporary or rapidly changing information.

Furthermore, LLMs often suffer from a **lack of transparency in their reasoning processes ?**. As complex neural networks, their internal mechanisms for arriving at an answer are largely opaque, making it difficult for human users to understand *how* a particular conclusion was reached or to verify the factual basis of a generated response. This "black box" nature hinders debugging, auditing, and building trust, especially in critical domains where explainability is paramount. When an LLM provides an incorrect answer, it is challenging to pinpoint whether the error stems from a misunderstanding of the query, a misinterpretation of internal knowledge, or a hallucination.

These inherent limitations of LLMs—hallucination, the knowledge cutoff, and lack of transparency—were recognized early in their development. They highlighted a critical need for mechanisms that could augment LLMs with external, up-to-date, and verifiable knowledge. This necessity directly paved the way for the development and widespread adoption of Retrieval-Augmented Generation (RAG) systems. RAG emerged as a promising paradigm to address these shortcomings by enabling LLMs to dynamically fetch relevant information from external knowledge bases during the generation process, thereby mitigating hallucinations, overcoming knowledge cutoffs, and offering a degree of verifiability by citing sources. Understanding these foundational limitations is crucial for appreciating the value and architectural evolution of RAG, as subsequent research has largely focused on refining how LLMs interact with and leverage external knowledge to overcome these challenges **???**. Despite the promise of RAG, the fundamental challenges of effectively integrating and reasoning over external knowledge, especially in the presence of noisy or irrelevant information, continue to drive ongoing research into more robust and intelligent augmentation strategies.

1.2 Introduction to Retrieval-Augmented Generation (RAG)

To address the inherent limitations of Large Language Models (LLMs), such as their propensity for factual hallucinations, reliance on static pre-training data leading to knowledge cutoffs, and a general lack of transparency in their reasoning, Retrieval-Augmented Generation (RAG) has emerged as a pivotal paradigm. RAG enhances LLMs by seamlessly integrating an information retrieval component, thereby grounding their responses in external, verifiable knowledge. This integration serves a multifaceted core purpose: to significantly mitigate LLM hallucinations, provide access to dynamic and up-to-date information, and ultimately improve the factual accuracy, reliability, and transparency of generated responses. This foundational understanding highlights RAG's critical role as a bridge between the vast, but often static and opaque, parametric knowledge encoded within LLMs and the dynamic, verifiable information available in the real world **?**.

The general mechanism of a RAG system involves two primary, synergistically operating components: a retriever and a generator. Upon receiving a user query, the retriever component first identifies and fetches relevant documents or passages from an external, non-parametric knowledge base. This knowledge base can range from a curated collection of proprietary documents indexed in a vector database to a vast corpus like Wikipedia. The selection process typically relies on semantic similarity between the query and the documents. Subsequently, these retrieved contexts are supplied to the generator component, which is typically a pre-trained LLM. The generator then synthesizes a coherent and accurate answer by leveraging both the original query and the provided external information. This process ensures that the LLM's output is not solely dependent on its internal, pre-trained knowledge, but is actively informed and constrained by external, verifiable sources.

The seminal work by **?** introduced the concept of Retrieval-Augmented Generation, proposing models that combine pre-trained parametric memory (a sequence-to-sequence model) with non-parametric memory (a dense vector index of Wikipedia). This foundational paper demonstrated that RAG models could achieve state-of-the-art results on knowledge-intensive Natural Language Processing (NLP) tasks, outperforming

parametric-only baselines by generating more specific, diverse, and factual language. This initial success underscored the transformative potential of augmenting LLMs with external knowledge, establishing RAG as a robust framework for enhancing language generation.

While the core RAG mechanism appears straightforward, its effective implementation involves a sophisticated interplay of several conceptual phases. As detailed by **?** in their comprehensive survey, the RAG paradigm can be broadly understood through four interconnected stages from an information retrieval perspective: pre-retrieval, retrieval, post-retrieval, and generation. The **pre-retrieval** phase focuses on optimizing the knowledge base and initial query, involving techniques like data indexing, chunking, and initial query manipulation to prepare for effective search. The **retrieval** phase is where the system actively searches and selects candidate documents based on the refined query. The **post-retrieval** phase then refines these initially retrieved documents, often through re-ranking, filtering, or summarization, to ensure only the most pertinent and high-quality context is passed to the LLM. Finally, the **generation** phase is where the LLM synthesizes the final response, conditioned on the original query and the carefully curated retrieved context. This structured view illustrates that RAG is not merely a simple concatenation of retrieval and generation, but a pipeline with multiple points of optimization to ensure the quality and relevance of the augmented information.

In essence, RAG provides a robust framework for overcoming the inherent limitations of standalone LLMs by dynamically integrating external knowledge. This capability is paramount for applications requiring high factual accuracy, up-to-date information, and verifiable outputs. Building on this foundational framework, subsequent research has focused on enhancing each component of the RAG pipeline, developing advanced architectures, and rigorously evaluating its performance across diverse applications. This review will systematically explore these advancements, delving into sophisticated retrieval strategies (Section 3), the evolution of RAG architectures (Section 4), critical challenges of evaluation and trustworthiness (Section 5), and its impact across various domain-specific applications (Section 6).

1.3 Scope and Organization of the Review

This literature review is meticulously structured to provide a comprehensive and pedagogical exploration of Retrieval-Augmented Generation (RAG), tracing its intellectual trajectory from foundational concepts to cutting-edge advancements and future challenges. The rapid evolution and increasing complexity of the RAG landscape, as highlighted by recent surveys such as **?**, underscore the critical need for a coherent and systematic overview. This review serves as a roadmap, guiding the reader through the interconnected developments that have shaped RAG into a pivotal paradigm for enhancing Large Language Models (LLMs).

The review commences in Section 1, "Introduction," by establishing the foundational context for RAG. It begins with an examination of the transformative capabilities of LLMs and a critical analysis of their inherent limitations, such as factual inaccuracies and knowledge cutoffs. This sets the stage for introducing RAG as a robust solution designed to mitigate these challenges by grounding LLM responses in external, verifiable knowledge.

Section 2, "Foundational Concepts, Early RAG Architectures, and Knowledge Context," delves into the bedrock of RAG. It meticulously dissects the core components—the retriever and the generator—and details their synergistic integration. This section highlights early architectural breakthroughs, including the seminal work by **?** that introduced the Retrieval-Augmented Generation model, demonstrating its transformative potential for knowledge-intensive tasks. Crucially, it also contextualizes RAG by contrasting it with methods relying solely on an LLM's internal parametric memory, thereby underscoring RAG's unique value proposition.

Building upon these foundations, Section 3, "Enhancing Retrieval: Strategies for Context Quality and Relevance," focuses on the critical advancements made in improving the quality and relevance of the retrieved context. This section explores sophisticated strategies that move beyond initial query-based retrieval, covering advanced query refinement and reformulation techniques, dynamic context ranking and reranking mechanisms, and innovative corrective and adaptive retrieval strategies. These innovations collectively aim

to provide the LLM with the most pertinent and accurate information.

Section 4, "Advanced RAG Architectures and System Optimizations," explores the evolution of RAG into more sophisticated and efficient systems. It delves into multi-stage and modular frameworks that orchestrate complex workflows, the integration of structured knowledge graphs for enhanced reasoning (GraphRAG), and the expansion of RAG to multimodal inputs. Furthermore, this section covers system-level optimizations aimed at improving the speed, scalability, and computational efficiency of RAG deployments, addressing the practical demands of real-world applications.

The critical importance of assessing RAG systems is addressed in Section 5, "Evaluation, Benchmarking, and Trustworthiness." This section examines the methodologies and challenges in systematically evaluating RAG, moving beyond anecdotal observations to rigorous assessment. It covers the development of specialized benchmarks designed to diagnose RAG's fundamental capabilities and limitations, particularly for complex reasoning tasks. The discussion also highlights innovative approaches for accurately evaluating the utility of retrieved information from the perspective of the LLM, and crucially, addresses emerging concerns surrounding privacy and security within RAG systems, emphasizing the need for trustworthy and responsible deployment, as underscored by systematic benchmarking efforts like **?**.

Section 6, "Domain-Specific Applications and Real-World Impact," showcases the practical utility and significant real-world impact of RAG across various specialized domains. It highlights how RAG is successfully applied to address complex, knowledge-intensive problems in high-stakes environments, demonstrating its ability to ground LLMs in authoritative, domain-specific knowledge, ranging from healthcare to customer service and legal applications.

Finally, Section 7, "Conclusion," and Section 8, "Future Directions and Open Challenges," synthesize the key insights from the review and project the future trajectory of RAG. These sections critically examine the evolving relationship between external retrieval and expanded LLM context windows, discuss the inherent tension in balancing increasing architectural complexity with efficiency and generalizability, and address the

paramount ethical considerations and responsible development practices for RAG systems. This concluding part outlines key areas for future research and responsible deployment to ensure RAG's continued advancement and beneficial impact.

Through this structured organization, the review aims to provide a coherent narrative that connects diverse research efforts, highlights the evolution of ideas within the field, and offers a comprehensive understanding of RAG's current state and future potential.

2 Foundational Concepts, Early RAG Architectures, and Knowledge Context

2.1 Core Components of RAG: Retriever and Generator

Retrieval-Augmented Generation (RAG) systems fundamentally address the limitations of Large Language Models (LLMs) in accessing and leveraging external, up-to-date, and factual knowledge by integrating a dynamic information retrieval mechanism. At the heart of every RAG system are two indispensable components: the retriever and the generator, working in concert to produce informed and coherent responses **?**.

The **retriever** is responsible for efficiently searching and fetching relevant documents or passages from a vast external knowledge base based on a given user query. This component acts as the system's dynamic memory, providing access to information beyond the LLM's static parametric knowledge **??**. While traditional information retrieval methods, often termed sparse retrievers, such as TF-IDF or BM25, rely on lexical matching and keyword overlap to identify relevant documents **??**, early RAG systems predominantly adopted dense passage retrievers (DPRs). DPRs map both the query and the documents into a shared high-dimensional embedding space, typically using neural networks, to capture semantic similarity **??**. By computing the similarity between the query embedding and document embeddings, the retriever can quickly identify and rank the most semantically relevant passages, even when there is no exact keyword match. This semantic understanding allows DPRs to overcome the limitations of sparse methods, which often

struggle with synonyms, polysemy, or conceptual relevance **?**. For instance, **?** introduced a neural retriever pre-trained on question-answer pairs, enabling it to access a dense vector index of Wikipedia and retrieve passages that are semantically similar to the input query, thereby dynamically augmenting the LLM's knowledge.

Concurrently, the **generator** component synthesizes a coherent and accurate response by leveraging both the original user query and the context provided by the retrieved passages **?**. Its primary role is to ground the LLM's output in factual information, thereby mitigating hallucinations and improving the factual accuracy of its outputs. Early RAG systems commonly employed sequence-to-sequence Large Language Models (LLMs) like BART or T5 as their generators **?**. These models receive the query and the top-k retrieved documents as augmented input, learning to condition their output on this combined context. This conditioning can be applied uniformly across the entire generated sequence or dynamically for each token, demonstrating flexibility in how the generator integrates retrieved information **?**. More recently, with the advent of increasingly powerful decoder-only LLMs, these models are frequently adapted to serve as RAG generators, leveraging their advanced generative capabilities to produce nuanced and contextually rich responses based on the retrieved evidence **?**.

This foundational retriever-generator paradigm underscores RAG's ability to combine the strengths of information retrieval with the generative prowess of LLMs. The effectiveness of RAG systems critically hinges on the synergistic operation of these two components. However, the overall performance remains highly sensitive to the quality and relevance of the retrieved documents, as well as the generator's capacity to effectively discern and utilize pertinent information from potentially noisy or redundant contexts **?**. Challenges such as irrelevant or insufficient retrievals can still lead to suboptimal generations, necessitating advanced strategies for corrective retrieval and context optimization **?**. These inherent complexities drive continuous advancements aimed at enhancing both retrieval efficacy and the generator's contextual understanding, which will be explored in subsequent sections.

2.2 End-to-End Training and Integration

2.3 RAG in Context: Contrasting with LLM's Parametric Memory

Large Language Models (LLMs) inherently possess a vast repository of knowledge, implicitly encoded within their billions of parameters during extensive pre-training. This internal, or *parametric*, memory allows LLMs to 'recite' or recall information and perform foundational reasoning without explicit external aid. This paradigm of knowledge access stands in crucial contrast to, and often complements, Retrieval-Augmented Generation (RAG), which relies on *non-parametric* external knowledge bases. Understanding this fundamental distinction is essential for appreciating RAG's unique value proposition in the broader landscape of knowledge augmentation.

The ability of LLMs to leverage their internal parametric knowledge has been a significant area of research. A prime example is the Recitation-Augmented Language Models (RECITE) framework **?**. RECITE proposes a two-step closed-book paradigm where the LLM first "recites" relevant passages from its *own memory* through sampling, and then generates the final answer based on this internally retrieved information. This approach, which incorporates techniques like self-consistency and passage hint-based diversified recitation, demonstrates that LLMs can effectively unlock and utilize their "fuzzy memorization" for knowledge-intensive tasks, achieving state-of-the-art results in closed-book question answering **?**. The strength of parametric memory lies in its ability to provide broad, general knowledge and facilitate complex reasoning patterns learned during pre-training. It represents a distilled, generalized understanding of the world as captured in its training data.

However, relying solely on an LLM's parametric memory presents several inherent limitations. Firstly, this knowledge is static, reflecting a specific point in time (the knowledge cutoff of its training data). Consequently, it can become outdated, leading to factual inaccuracies or an inability to address queries about recent events or developments. Secondly, parametric memory often lacks explicit verifiability and attribution; the LLM cannot typ-

ically cite the source of its internal 'knowledge,' making it difficult to trust or audit its factual claims. Thirdly, while vast, an LLM's internal knowledge can be shallow or incomplete for highly specific, niche, or "less popular" domain knowledge. Fine-tuning an LLM to inject such specialized knowledge is often an expensive and time-consuming process, requiring substantial, high-quality training data that may be scarce **??**.

This is precisely where external Retrieval-Augmented Generation (RAG) offers critical advantages, providing a dynamic, verifiable, and up-to-date complement to the LLM's internal knowledge. The foundational RAG paradigm, introduced by lewis2020pwr, established a mechanism where a pre-trained sequence-to-sequence model (the generator) is augmented by a retriever that fetches relevant documents from a dense vector index (the non-parametric memory). This external information then conditions the generator's output. This architecture fundamentally addresses the limitations of parametric memory by:

1. **Providing Dynamic and Up-to-date Knowledge:** Unlike static parametric memory, RAG systems can access and integrate the latest information by simply updating their external knowledge base (e.g., a vector database or knowledge graph), without requiring costly re-training or fine-tuning of the LLM **?**. This is crucial for domains with rapidly evolving information.

2. **Enhancing Verifiability and Attribution:** RAG inherently provides provenance for its generated answers by presenting the retrieved documents as evidence. This transparency allows users to verify factual claims and improves the trustworthiness of the LLM's responses, a critical feature for high-stakes applications **?**.

3. **Handling Domain-Specific and Long-Tail Knowledge:** RAG excels in scenarios where an LLM's general parametric memory is insufficient or inaccurate for specialized domains. For instance, soudani20247ny conducted a comprehensive empirical comparison, demonstrating that RAG substantially outperforms fine-tuning for question answering over "less popular" factual knowledge, highlighting the difficulty of encoding such niche facts effectively into parametric memory. Similarly,

barron2024kue introduce SMART-SLIC, a domain-specific RAG framework that integrates knowledge graphs and vector stores built without LLMs for highly specialized domains like malware analysis, effectively mitigating hallucinations and lessening the need for expensive fine-tuning. In the clinical domain, RAG-based systems have been shown to greatly outperform general-purpose LLMs in producing relevant, evidence-based, and actionable answers to complex clinical questions, particularly when existing data are available **?**. This underscores RAG's superior capacity for grounding LLMs in authoritative, external knowledge that is not, or cannot be, effectively encoded in an LLM's static parameters.

4. **Mitigating Hallucination:** By grounding responses in retrieved facts, RAG significantly reduces the LLM's propensity to generate factually incorrect or fabricated information, a common challenge with parametric-only models **?**.

In conclusion, while an LLM's parametric memory provides a vast, general knowledge base and foundational reasoning abilities, external RAG offers crucial augmentation. RAG's strength lies in its capacity to provide dynamic, verifiable, domain-specific, and up-to-date information, effectively mitigating hallucination and enabling deeper factual grounding for complex queries. The most effective knowledge systems often leverage both paradigms, utilizing the LLM's internal knowledge for broad understanding and reasoning, while strategically employing RAG to access and integrate precise, current, and externally validated information. Future research continues to explore how to seamlessly integrate these two knowledge sources, dynamically determining the optimal reliance on each for superior performance across diverse tasks.

3 Enhancing Retrieval: Strategies for Context Quality and Relevance

3.1 Advanced Query Refinement and Reformulation

The effectiveness of Retrieval-Augmented Generation (RAG) systems fundamentally relies on the precision and relevance of the retrieved context. While early RAG architectures **?** demonstrated the transformative potential of grounding Large Language Models (LLMs) in external knowledge, their reliance on static user queries often proved insufficient for complex, ambiguous, or multi-hop information needs **??**. This limitation has driven significant research into sophisticated techniques where the LLM actively participates in refining or reformulating the initial user query, a critical component of the "pre-retrieval" phase as highlighted by recent surveys **??**. This proactive approach, often involving specialized instruction fine-tuning, significantly enhances the initial retrieval step, thereby improving the overall robustness and accuracy of RAG systems.

Initial advancements in query enhancement focused on expanding or augmenting the original user query, often through heuristic methods or simpler LLM prompts. One prominent technique is Hypothetical Document Embeddings (HyDE), where an LLM generates a plausible, hypothetical answer to the user's query. This synthetic document is then embedded and used as the query for retrieval, leveraging the LLM's generative capacity to create a more semantically rich search vector that often aligns better with relevant documents than the original short query. Building on this, methods like DPA-RAG **?** introduced diverse query augmentation strategies, training a retriever to align with the LLM's varied knowledge preferences, thereby alleviating preference data scarcity and improving retrieval relevance. Similarly, Telco-RAG **?**, designed for technical domains, incorporates a query enhancement stage that uses a custom glossary for lexicon-enhanced queries and an LLM to generate candidate answers from preliminary context. These candidates then help refine the user's query, clarifying intent and preventing irrelevant retrieval. Another approach, seen in the Distill-Retrieve-Read framework **?**, leverages a tool-calling mechanism to formulate keyword-based search queries, effectively translating natural language

requests into more retriever-friendly formats. These techniques underscore a foundational shift from passive query submission to active, LLM-guided query enrichment.

A more advanced paradigm involves LLMs learning to explicitly rewrite, decompose, or disambiguate queries. A foundational work in this area is Search Engine-Augmented Generation (SEA) **?**, which trained a dedicated "Search Query Generator" to formulate effective search queries from dialogue context for a real-time internet search engine. This demonstrated the feasibility of teaching LLMs to generate queries that go beyond simple keywords. Extending this, RQ-RAG **?** represents a significant leap by end-to-end training a Large Language Model to dynamically refine search queries through rewriting, decomposition, and disambiguation. Its innovation lies in a novel dataset construction pipeline that uses a powerful external LLM (ChatGPT) to craft tailored search queries for specific refinement scenarios and to regenerate contextually aligned answers. At inference, RQ-RAG employs internal trajectory selection strategies (e.g., Perplexity, Confidence) to navigate multi-path query refinement without relying on external LLMs for decision-making. This approach has shown substantial improvements on both single-hop and multi-hop QA tasks, often outperforming larger proprietary models. For multi-faceted queries, RichRAG **?** includes a sub-aspect explorer module to identify potential sub-intents, enabling a multi-faceted retriever to build a diverse candidate pool. This contrasts with RQ-RAG's more integrated decomposition, highlighting different architectural choices for handling complex queries. Furthermore, for structured knowledge bases, LLMs can be trained to translate natural language queries into specific query languages, as seen in **?**, where an LLM parses customer queries for entities and intents, then translates them into graph database language (e.g., Cypher) for precise subgraph retrieval from a Knowledge Graph. This demonstrates query reformulation tailored to data structure.

The most sophisticated query refinement techniques involve iterative and conversational approaches, where the LLM engages in multiple turns of information-seeking. Auto-RAG **?** exemplifies this by introducing an autonomous iterative retrieval model centered on the LLM's powerful decision-making. It engages in multi-turn dialogues with the retriever, systematically planning retrievals and refining queries until sufficient external in-

formation is gathered. This allows the LLM to dynamically adjust its information-seeking depth based on perceived knowledge gaps. Similarly, i-MedRAG **?** applies this iterative paradigm to the medical domain, where LLMs iteratively generate follow-up questions to search for additional information from external medical corpora. This "reason-then-query" pipeline enables LLMs to dynamically break down complex medical problems and gather context-specific information, significantly outperforming single-round retrieval for complex clinical reasoning tasks. DR-RAG **?** also contributes to this iterative refinement by dynamically assessing document relevance and improving retrieval recall by combining parts of initially retrieved documents with the query, effectively adjusting the query based on partial, even low-relevance, feedback. These iterative methods highlight a crucial shift towards LLMs managing their own information-seeking process, dynamically adapting queries based on intermediate retrieval results.

In summary, the evolution of RAG systems has progressed from passive, static queries to active, LLM-driven query refinement and reformulation. Techniques range from query expansion and augmentation **???** and the generation of hypothetical documents, to learned query rewriting and decomposition **????**, and sophisticated iterative or conversational refinement strategies **???**. The critical advancements lie in training LLMs to autonomously generate more effective search queries, often supported by specialized datasets and internal decision-making mechanisms. Future research will likely focus on making these query refinement processes even more granular, context-aware, and efficient, potentially exploring real-time adaptation to user feedback and broader generalization across diverse domains and complex reasoning tasks.

3.2 Context Ranking and Reranking Mechanisms

The effectiveness of Retrieval-Augmented Generation (RAG) systems is profoundly influenced by the quality and precise ordering of the retrieved documents presented to the Large Language Model (LLM). While foundational RAG models, such as those pioneered by **?**, established the paradigm of augmenting LLMs with external knowledge, a persistent challenge has been the LLM's inherent difficulty in effectively processing a large volume

of retrieved contexts, particularly when irrelevant or noisy information is present. This limitation often leads to degraded efficiency and accuracy, as systematically highlighted by benchmarking efforts like **?**, which revealed LLMs' struggles with noise robustness, negative rejection, and information integration. Furthermore, accurately evaluating the true utility of retrieved documents to the LLM has proven challenging, with traditional relevance metrics often showing low correlation with downstream performance, as demonstrated by **?**'s eRAG methodology. These challenges underscore the critical need for sophisticated mechanisms to optimize the order and quality of retrieved documents before LLM generation.

Initially, reranking mechanisms emerged as a crucial intermediate step to refine the output of an initial, often recall-oriented, retriever. These early approaches typically employed separate "expert ranking models," often based on smaller transformer architectures like BERT or T5, which were fine-tuned to score the relevance of individual retrieved passages to the query. These cross-encoder models, by performing full attention over the concatenated query and document, could achieve high precision in identifying relevant contexts **?**. Benchmarking efforts, such as those by **?**, have systematically evaluated the performance of various rerankers, highlighting their ability to significantly improve the quality of the top-k documents. However, these dedicated rerankers added architectural complexity, incurred additional computational overhead, and often lacked the zero-shot generalization capabilities inherent to larger LLMs, necessitating extensive fine-tuning for new domains or tasks.

The field has since evolved to leverage the powerful natural language understanding and reasoning capabilities of LLMs themselves for reranking. This shift is motivated by the observation that LLMs, especially when instruction-tuned, can discern nuanced relevance and contextual relationships more effectively than smaller, specialized models. One direction involves training LLMs to align their retrieval preferences with their generation capabilities. For instance, **?**'s DPA-RAG proposes a dual preference alignment framework that integrates pairwise, pointwise, and contrastive preference alignment into the reranker. This external alignment, combined with an internal alignment stage for

the LLM, helps the reranker better anticipate what knowledge the LLM will find most useful for generation, thereby improving the reliability of the RAG system. Similarly, **?** introduced an RAG framework that uses "reflective tags" to enable adaptive control of retrieval, where the LLM implicitly performs a form of reranking by evaluating documents in parallel and selecting the highest quality content for generation, reducing reliance on irrelevant data. Expanding this to multimodal contexts, **?** demonstrated that Multimodal Large Language Models (MLLMs) can serve as strong rerankers, effectively filtering top-k retrieved images in multimodal RAG systems, showcasing the versatility of LLM-based reranking across modalities.

A significant architectural evolution in this domain is the unification of context ranking and answer generation within a single instruction-tuned LLM, as exemplified by **?**'s RankRAG. This approach directly addresses the limitations of separate expert rankers and the added complexity of multi-component pipelines. RankRAG proposes a novel two-stage instruction fine-tuning framework that trains a single LLM for the dual purpose of context ranking and answer generation. It integrates a specialized instruction-tuning task for context ranking, framed as a simple question-answering problem where the LLM learns to identify context relevance (e.g., generating "True" or "False"). This task is seamlessly blended with context-rich and retrieval-augmented QA datasets. Remarkably, **?** observed that incorporating even a small fraction of this specialized ranking data into the instruction-tuning blend yields superior ranking performance, often outperforming LLMs exclusively fine-tuned on significantly larger ranking datasets. This effectiveness stems from the LLM's inherent ability to transfer its general reasoning and language understanding capabilities to the ranking task, leading to a more robust and generalized understanding of relevance. This unification simplifies the RAG pipeline, reduces architectural complexity, and leverages the LLM's inherent capabilities to discern context relevance, leading to superior zero-shot generation performance and strong generalization across diverse tasks, including biomedical RAG benchmarks without domain-specific tuning.

While RankRAG represents a substantial step towards streamlining RAG by integrat-

ing ranking into the LLM, the field continues to explore how ranking mechanisms can handle increasingly complex scenarios and user needs. A key challenge lies in developing ranking mechanisms capable of identifying and prioritizing interconnected contexts for multi-hop queries, which require reasoning over multiple disparate pieces of evidence. Traditional rerankers often struggle with this, as they typically score documents independently. To address this, **?**'s HippoRAG, inspired by neurobiology, employs a knowledge graph and Personalized PageRank algorithm to perform efficient, single-step multi-hop reasoning and ranking, demonstrating how structural awareness can enhance context selection for complex queries. Furthermore, beyond mere relevance, there is a growing need for ranking mechanisms that can ensure diversity and comprehensiveness in retrieved contexts, especially for broad, multi-faceted queries. **?**'s RichRAG introduces a generative list-wise ranker that not only identifies relevant documents but also ensures they collectively cover various query aspects, aligning with the generator's preference for producing rich, long-form answers. This highlights a shift towards listwise ranking, where the utility of a *set* of documents is optimized, rather than just individual documents. Future research in context ranking must continue to address these complexities, focusing on developing adaptive, diverse, and collectively optimal ranking strategies that can truly empower LLMs to synthesize comprehensive and accurate responses from vast and varied knowledge bases.

3.3 Corrective and Adaptive Retrieval Strategies

Traditional Retrieval-Augmented Generation (RAG) systems, while effective at grounding Large Language Models (LLMs) with external knowledge **?**, often operate under the implicit assumption of perfect initial retrieval. However, real-world information retrieval is inherently noisy, prone to irrelevance, and can suffer from incompleteness, leading to issues like hallucination, factual inaccuracies, and limited coverage in generated responses **?**. This fundamental challenge has spurred the development of advanced RAG architectures that move beyond static, one-shot retrieval by dynamically assessing the quality and sufficiency of retrieved documents and taking proactive or corrective actions. These

strategies empower LLMs to exhibit meta-cognition over their knowledge acquisition process, leading to more robust and intelligent responses.

A prominent paradigm in this area involves enabling LLMs to self-reflect on the relevance and sufficiency of retrieved information, dynamically triggering subsequent steps. The *Self-RAG* framework **?**, for instance, empowers LLMs to dynamically decide when to retrieve additional information and, crucially, to critique their own generations. This is achieved by training the LLM to generate special "reflection tokens" that indicate the quality of retrieved passages and the faithfulness/helpfulness of its own generated text. Based on these self-critiques, the LLM can then decide to re-retrieve, refine its generation, or even abstain from answering if the information is insufficient. This integrated, LLM-centric approach enhances robustness against retrieval failures by allowing the model to actively manage its knowledge acquisition and output quality, making the LLM a more autonomous agent in the RAG pipeline.

Complementing this LLM-driven self-reflection are frameworks that introduce explicit, modular mechanisms for evaluating retrieval quality and initiating corrective actions. Corrective Retrieval Augmented Generation (CRAG) **?** introduces a pioneering strategy that employs a lightweight, external retrieval evaluator to assess the confidence in the initial set of retrieved documents. Based on this assessment, CRAG dynamically triggers one of three distinct corrective actions: "Correct" (if relevant documents are found, leading to knowledge refinement), "Incorrect" (if documents are largely irrelevant, prompting a large-scale web search for external correction), or "Ambiguous" (a soft strategy combining refinement of initial documents with web search results). Furthermore, CRAG refines relevant documents using a "decompose-then-recompose" algorithm, segmenting them into fine-grained "knowledge strips" and filtering out irrelevant parts to optimize information utilization. This dynamic, multi-action approach significantly mitigates the impact of poor initial retrieval, a critical vulnerability in traditional RAG systems.

Another approach to adaptive retrieval is seen in Active Retrieval Augmented Generation (ARAG) **?**. Similar to Self-RAG in its LLM-driven decision-making, ARAG focuses on the LLM actively deciding *when* to retrieve and *what* to retrieve next based on its

confidence in generating an answer. If the LLM's internal confidence score is low, indicating uncertainty or insufficient information, ARAG triggers further retrieval steps, potentially with refined queries. This proactive adaptation allows the system to actively seek out necessary information rather than passively accepting initial retrieval results, thereby improving the accuracy and completeness of responses, especially for complex or knowledge-intensive queries.

The concept of iterative and adaptive information seeking is further explored in multi-round frameworks. For example, IM-RAG **?** (Inner Monologue RAG) leverages an LLM's "inner monologue" to generate and refine plans for complex decision-making, which in turn guides flexible, multi-round retrieval and generation. While primarily an architectural framework for complex tasks, its multi-round nature implies an adaptive loop where the LLM's internal reasoning (monologue) can implicitly assess the sufficiency of previous retrieval and adjust its subsequent information-seeking strategy, effectively correcting its path towards a better answer.

Comparing these approaches reveals distinct philosophies in achieving robustness. Self-RAG and ARAG represent LLM-centric, integrated self-correction, where the LLM itself is endowed with meta-cognitive abilities to assess and adapt. This offers high flexibility and potentially more nuanced adaptation, but relies heavily on the LLM's fine-tuning and inherent capabilities to self-critique effectively. In contrast, CRAG adopts a more modular approach, employing a separate, lightweight evaluator and explicit, pre-defined corrective paths, including a robust web search fallback for severe retrieval failures. This modularity can offer greater reliability and control, especially for out-of-domain queries or when the initial knowledge base is truly insufficient, but might be less flexible than an LLM's integrated self-reflection.

In conclusion, the evolution of RAG systems is marked by a clear trajectory towards greater intelligence and robustness, moving from passive information consumption to active, adaptive knowledge seeking. By integrating LLM-driven self-reflection (Self-RAG, ARAG), dynamic corrective actions via external evaluators (CRAG), and multi-round adaptive strategies (IM-RAG), these advanced frameworks enable LLMs to navigate the

complexities of real-world information retrieval more effectively. However, these advancements often introduce increased computational overhead and architectural complexity, necessitating ongoing research into balancing efficiency, generalizability, and the continued development of sophisticated evaluation metrics for these dynamic systems. The ability to dynamically assess and correct retrieval failures is paramount for deploying RAG in critical, real-world applications where accuracy and reliability are non-negotiable.

4 Advanced RAG Architectures and System Optimizations

4.1 Multi-stage and Modular RAG Frameworks

The foundational paradigm of Retrieval-Augmented Generation (RAG) typically operates on a straightforward "retrieve-then-generate" sequence ?. However, as Large Language Models (LLMs) are increasingly tasked with complex, multi-faceted queries and dynamic information needs, this simple pipeline proves insufficient ??. This has spurred the evolution of RAG into more sophisticated, multi-stage, and modular architectures, where the LLM transcends a passive role to become an intelligent agent capable of proactive planning, dynamic decision-making, and the orchestration of various sub-tasks ?. This section focuses on frameworks that empower LLMs to actively manage the information-seeking process through iterative planning, query decomposition, and the dynamic assembly of specialized modules. It is crucial to distinguish these proactive, agentic approaches from reactive or corrective mechanisms (e.g., self-correction, re-ranking) that primarily refine retrieval quality, which are discussed in detail in Section 3.

A significant advancement in modular RAG involves empowering LLMs to act as sophisticated planning agents, iteratively refining their information-seeking process and orchestrating multi-round interactions. ? introduced PlanRAG, which extends the popular ReAct framework by incorporating explicit "Plan" and "Re-plan" steps. This allows LLMs to dynamically generate and iteratively refine analytical approaches based

on intermediate retrieval results, effectively acting as decision-makers for complex data analysis tasks. Similarly, ? presented IM-RAG, a multi-round RAG system that leverages learned inner monologues and a multi-agent reinforcement learning approach. In IM-RAG, an LLM-based "Reasoner" dynamically switches between a "Questioner" role (crafting queries) and an "Answerer" role, guided by mid-step rewards from a "Progress Tracker," leading to flexible and interpretable multi-round information gathering. Building on the concept of autonomous interaction, ?'s Auto-RAG enables LLMs to engage in multi-turn dialogues with the retriever, systematically planning retrievals and refining queries until sufficient external information is gathered. This framework highlights the LLM's powerful decision-making capabilities, autonomously adjusting iterations based on query difficulty and knowledge utility. Another approach, ?'s M-RAG, proposes a multi-partition paradigm for external memories, employing a multi-agent reinforcement learning framework with an "Agent-S" for dynamic partition selection and an "Agent-R" for memory refinement. This enables more fine-grained and focused retrieval by orchestrating memory access across different knowledge partitions. To further optimize the interaction between these modular components, ?'s RAG-DDR (Differentiable Data Rewards) offers an end-to-end training method that aligns data preferences between different RAG modules (agents). By collecting rewards and evaluating the impact of perturbations on the entire system, RAG-DDR optimizes agents to produce outputs that enhance overall RAG performance, particularly for smaller LLMs. These agentic frameworks collectively transform RAG into a dynamic, adaptive system capable of tackling complex, multi-hop queries that require sophisticated reasoning and iterative information synthesis.

Beyond specific agentic planning algorithms, other modular architectures focus on meta-frameworks and system-level optimizations for orchestrating and deploying complex RAG pipelines. Given the proliferation of RAG modules and techniques, ?'s AutoRAG proposes an automated framework to identify optimal combinations of RAG modules for specific datasets. This meta-level modularity simplifies the complex task of RAG pipeline optimization, making it more accessible and efficient for researchers and practitioners. ?'s FlashRAG provides a comprehensive, modular toolkit specifically designed for efficient

RAG research. It supports various complex RAG process flows, including sequential, branching, conditional, and loop-based pipelines, by offering fine-grained modularity at both component and pipeline levels. This enables researchers to easily swap, combine, and customize RAG workflows, accelerating the development and benchmarking of novel multi-stage RAG architectures. In a different vein, ?'s uRAG introduces a unified retrieval engine designed to serve multiple downstream RAG systems, each with unique purposes like question answering or fact verification. This framework exemplifies modularity at a broader system level, standardizing communication and enabling a shared retrieval infrastructure, akin to a "search engine for machines" ?. Similarly, ?'s Ragnarök provides a reusable RAG framework and baselines for evaluating RAG systems, contributing to the standardization and systematic assessment of these increasingly complex architectures.

It is also worth noting that Graph-Augmented RAG (GraphRAG), discussed in detail in Section 4.2, inherently represents a multi-stage and modular paradigm, necessitating specialized processing for structured knowledge before integration with LLMs.

In conclusion, the evolution towards multi-stage and modular RAG frameworks marks a significant advancement, transforming RAG from a simple pipeline into an intelligent, adaptive system. By enabling LLMs to engage in iterative refinement, agentic planning, and dynamic orchestration of sub-tasks, these architectures enhance robustness, reduce hallucinations, and improve the depth and faithfulness of generated responses, particularly for complex, multi-hop queries ?. However, this sophistication often introduces challenges related to increased computational overhead, the complexity of orchestrating multiple modules, and the need for robust evaluation methodologies that can accurately assess the contributions of each stage and the overall system performance. Benchmarks like ?'s RAGBench, ?'s FRAMES, and ?'s MultiHop-RAG highlight these challenges, emphasizing the need for explainable metrics and unified frameworks to evaluate the intricate interplay of retrieval, reasoning, and generation in these advanced systems. Future research will likely focus on optimizing the efficiency of these multi-stage processes, developing more autonomous and self-correcting agents, and creating more generalized frameworks that can seamlessly integrate diverse knowledge sources and reasoning paradigms while addressing

the inherent trade-offs between complexity and efficiency.

4.2 Graph-Augmented Retrieval-Augmented Generation (GraphRAG)

Large Language Models (LLMs) often struggle with factual accuracy, outdated knowledge, and complex, multi-hop reasoning, leading to issues like hallucination **?**. While Retrieval-Augmented Generation (RAG) offers a powerful paradigm to ground LLMs with external knowledge **?**, traditional RAG systems primarily rely on semantic similarity over unstructured text chunks, often failing to capture the explicit structural and relational information critical for intricate queries **?**. Graph-Augmented RAG (GraphRAG) emerges as a specialized solution, integrating structured knowledge, particularly Knowledge Graphs (KGs) or textual graphs, to enhance reasoning, factual accuracy, and context awareness by leveraging explicit relational information **??**.

Early GraphRAG research began to address the limitations of conventional RAG when confronted with complex, structured data. A pioneering effort is **?**'s **G-Retriever**, which introduces the first RAG approach specifically designed for *general textual graphs*. G-Retriever tackles the challenges of hallucination and scalability inherent in processing complex graph structures by formulating subgraph retrieval as a Prize-Collecting Steiner Tree (PCST) optimization problem, enabling the precise extraction of contextually and structurally relevant graph portions. Building on this, **?** demonstrates the practical benefits of integrating KGs for customer service question answering. Their approach constructs a novel dual-level KG that preserves both intra-issue structure and inter-issue relations from support tickets, employing an LLM-driven mechanism to translate natural language queries into graph database languages (e.g., Cypher) for highly precise subgraph retrieval. This significantly improved Mean Reciprocal Rank by 77.6

Further advancements in graph-aware retrieval and integration techniques have refined how LLMs interact with structured knowledge. **?**'s **GRAG** extends RAG for *networked documents* by integrating joint textual and topological information. GRAG employs a divide-and-conquer strategy with soft pruning for efficient textual subgraph retrieval and

a dual-view prompting mechanism that converts subgraphs into hierarchical text descriptions (hard prompts) and uses relevance-guided Graph Neural Networks (GNNs) for soft prompts. Complementing this, **?**'s **GNN-RAG** repurposes GNNs as powerful "dense subgraph reasoners" for precise retrieval of multi-hop answer candidates and their reasoning paths from KGs. These verbalized paths are then fed to an LLM, achieving state-of-the-art performance on KGQA benchmarks like WebQSP and CWQ with smaller LLMs, often outperforming larger models like GPT-4. Emphasizing efficiency, **?**'s **SubgraphRAG** proposes a lightweight MLP with Directional Distance Encoding (DDE) for scalable subgraph extraction, formulating retrieval as a triple factorization problem. This "simple is effective" approach allows unfine-tuned LLMs to achieve competitive accuracy on multi-hop KGQA tasks while significantly reducing hallucinations and improving explainability.

The field has also seen the emergence of sophisticated hybrid approaches and iterative reasoning paradigms. **?**'s **HybridRAG** combines the strengths of traditional Vector-RAG and GraphRAG to overcome their individual limitations, particularly for complex, domain-specific texts like financial earnings call transcripts. This hybrid model leverages a two-tiered LLM chain for robust KG construction and amalgamates context from both retrieval mechanisms, demonstrating superior performance in information extraction. Taking iterative reasoning a step further, **?**'s **Think-on-Graph 2.0 (ToG-2)** introduces a *tight-coupling* iterative exploration between KGs and unstructured text. ToG-2 alternates between knowledge-guided graph search and context retrieval, using LLMs for dynamic relation and entity pruning, enabling deeper and more faithful multi-step reasoning trajectories. Furthermore, **?**'s **HippoRAG** offers a neurobiologically inspired framework for efficient *single-step multi-hop reasoning*. By extracting a schemaless KG and applying Personalized PageRank (PPR), HippoRAG achieves significant speed and cost advantages over iterative methods while outperforming single-step baselines on challenging multi-hop QA benchmarks.

In conclusion, GraphRAG represents a critical evolution in RAG, moving beyond semantic similarity to explicitly leverage the rich structural and relational information within knowledge graphs and textual graphs. These approaches significantly enhance

LLM reasoning capabilities, improve factual accuracy, and mitigate hallucination, especially for complex, multi-hop queries. However, challenges remain in the automated construction and dynamic updating of high-quality knowledge graphs, optimizing the efficiency of subgraph extraction from massive graphs, and effectively balancing the depth of graph-based reasoning with the computational overhead it introduces **?**. Future research will likely focus on more adaptive and autonomous graph construction, real-time graph updates, and the seamless integration of diverse graph-aware retrieval and reasoning modules within increasingly intelligent RAG architectures.

4.3 Multimodal RAG: Integrating Diverse Knowledge Sources

The landscape of Retrieval-Augmented Generation (RAG) is rapidly evolving beyond its foundational text-centric paradigm, moving towards the integration of diverse knowledge modalities to foster more comprehensive and contextually rich responses from Large Language Models (LLMs). This expansion is crucial for enabling LLMs to interact with and understand the real world, which inherently comprises visual, auditory, and other forms of information alongside text. The goal is to create more versatile LLMs capable of understanding and generating responses based on a richer, real-world context, thereby mitigating the limitations of purely textual knowledge bases.

A pivotal step in this direction was the introduction of MuRAG (Multimodal Retrieval-Augmented Generator) by **?**, which pioneered multimodal retrieval-augmented generation for open question answering over images and text. Prior RAG systems were predominantly limited to retrieving textual knowledge, posing a significant challenge for queries requiring visual grounding or multimodal reasoning **?**. MuRAG addresses this by proposing a novel architecture that leverages a unified multimodal encoder, combining pre-trained T5 and ViT models, to process queries and memory candidates across both image and text modalities. Its methodology involves a retriever stage utilizing Maximum Inner Product Search (MIPS) to fetch relevant Top-K multimodal items, which are then fed to a reader stage for text generation **?**. A key innovation lies in its joint pre-training objective, which integrates a contrastive loss for effective retrieval with a generative loss for leveraging

multimodal knowledge, alongside an efficient two-stage fine-tuning pipeline designed to manage the computational complexities of large external multimodal memories ?. While MuRAG demonstrated the substantial benefits of incorporating visual knowledge into the generation process, its monolithic design and joint optimization posed challenges in terms of scalability and adaptability to dynamic, noisy multimodal inputs.

Building upon this foundation, subsequent research has focused on refining the retrieval and integration processes to enhance robustness and accuracy, particularly in the face of real-world complexities. For instance, the challenge of multi-granularity noisy correspondence (MNC) and the static nature of Multimodal Large Language Model (MLLM) training data can hinder accurate retrieval and generation in dynamic contexts. To address these limitations, ? introduced RagVL, a novel framework featuring knowledge-enhanced reranking and noise-injected training. RagVL instruction-tunes an MLLM to serve as a powerful reranker, precisely filtering the top-k retrieved images to improve the quality of augmented information ?. Furthermore, it enhances the generator's robustness by injecting visual noise during training at both data and token levels, thereby making the system more resilient to variations and imperfections in multimodal inputs ?. This approach directly improves upon the concept of multimodal retrieval by ensuring that the retrieved information is not only relevant but also of high quality and effectively utilized by the generator, offering a more modular and robust alternative to MuRAG's end-to-end joint training.

Beyond specific architectural designs, the broader integration of multimodal capabilities into RAG systems is gaining traction. Some comprehensive RAG optimization frameworks, while primarily focused on text, also explore the incorporation of multimodal retrieval. For example, ? investigates best practices across the entire RAG workflow and highlights the significant enhancement of question-answering capabilities on visual inputs, and the acceleration of multimodal content generation through multimodal retrieval techniques, including a "retrieval as generation" strategy. This suggests that the principles of efficient RAG design, such as optimal chunking, embedding, and reranking, are being extended to encompass multimodal data, indicating a convergence of general RAG

advancements with multimodal requirements.

Despite these advancements, a critical challenge in multimodal RAG lies in the effective evaluation and utilization of non-textual evidence. Benchmarking efforts have revealed that even state-of-the-art MLLMs struggle to efficiently extract and utilize visual knowledge. **?** introduced Visual-RAG, a question-answering benchmark specifically designed for visually grounded, knowledge-intensive queries that require text-to-image retrieval and the integration of retrieved clue images to extract visual evidence. Their findings underscore the persistent need for improved visual retrieval, grounding, and attribution mechanisms within multimodal RAG systems, highlighting a gap in current models' ability to fully leverage visual context. This points to a deeper issue beyond mere retrieval accuracy: the capacity of the MLLM to *reason* effectively with the retrieved visual information.

The practical impact of multimodal RAG is particularly evident in high-stakes domains where factual accuracy and hallucination reduction are paramount. In healthcare, for instance, Multimodal Large Language Models (MLLMs) face significant challenges with hallucination, especially when generating medical reports from images. To address this, **?** demonstrated how Visual RAG (V-RAG), incorporating both text and visual data from retrieved images, can significantly improve the accuracy of entity probing in medical image caption generation and chest X-ray report generation. By grounding medical entities in visual evidence, V-RAG enhances clinical accuracy and reduces hallucinations, showcasing the transformative potential of multimodal RAG in critical applications. This work highlights that multimodal RAG is not just about expanding input modalities, but about enhancing trustworthiness and reliability in sensitive contexts.

The progression from pioneering multimodal retrieval to refining its components and addressing its evaluation challenges highlights a critical trajectory in RAG research. While significant strides have been made in enabling LLMs to integrate diverse knowledge sources, challenges persist in scaling these systems to even larger and more heterogeneous multimodal knowledge bases. Future directions include developing more sophisticated cross-modal reasoning capabilities that go beyond simple concatenation of modalities,

improving the efficiency of multimodal indexing and retrieval for real-time applications involving massive datasets (e.g., millions of video or audio segments), and exploring novel ways to synthesize information from an ever-increasing array of modalities beyond just images and text, such as video, audio, and sensor data. Furthermore, the development of robust evaluation metrics for visual grounding and attribution, as highlighted by **?**, remains a critical need. The ultimate goal remains the creation of truly versatile LLMs capable of understanding and generating responses based on a richer, real-world context, while ensuring faithfulness and interpretability across all modalities.

4.4 System-Level Optimizations and Efficiency

The successful deployment of Retrieval-Augmented Generation (RAG) systems in real-world scenarios hinges critically on their efficiency, speed, and scalability. As RAG architectures grow in complexity, integrating external knowledge often leads to increased latency, higher computational overhead, and significant memory demands, necessitating advanced system-level optimizations.

A primary bottleneck in RAG is the computational and memory cost associated with processing long input sequences, particularly the Key-Value (KV) caches generated during the prefill phase of Large Language Model (LLM) inference. To address this, **?** introduced *RAGCache*, a novel multilevel dynamic caching system tailored for RAG. RAGCache caches the intermediate states (KV tensors) of retrieved documents in a prefix tree structure, called the Knowledge Tree, allowing for efficient sharing across multiple requests while respecting the LLM's position sensitivity. This system also employs a Prefix-aware Greedy-Dual-Size-Frequency (PGDSF) replacement policy for cache eviction and dynamic speculative pipelining to overlap CPU-bound retrieval with GPU-bound LLM inference, demonstrating up to a 4x reduction in Time to First Token (TTFT) and a 2.1x increase in throughput. Complementing this, **?** proposed *TurboRAG*, which further accelerates RAG by pre-computing and storing KV caches of documents offline. This approach eliminates online KV cache computation during inference, leading to an average 8.6x reduction in TTFT while maintaining comparable performance to standard RAG systems.

Beyond caching, algorithm-system co-design approaches are crucial for enhancing RAG performance. ? presented *PipeRAG*, an innovative framework that co-designs the RAG algorithm with the underlying retrieval system to reduce generation latency, especially during periodic retrievals. PipeRAG introduces pipeline parallelism by using a "stale" query window to prefetch content, enabling concurrent execution of retrieval and inference. It also supports flexible retrieval intervals and employs performance-model-driven retrievals to dynamically adjust the Approximate Nearest Neighbor (ANN) search space, balancing retrieval quality and latency. This co-design achieved up to a 2.6x speedup in end-to-end generation latency and improved generation quality.

Other architectural and algorithmic strategies also contribute to system efficiency. ? developed *Telco-RAG* for the telecommunications domain, which includes a Neural Network (NN) router to predict relevant document sub-sections. This intelligent routing significantly reduces RAM consumption by 45

The management of large knowledge bases is another area for system-level optimization. ? proposed *M-RAG*, a multiple partition paradigm that organizes external memories into distinct partitions. This allows for fine-grained retrieval by selecting the most suitable partition for a given query, which not only enhances retrieval precision but also offers benefits for index management, privacy, and distributed processing, thereby improving overall system scalability.

To facilitate efficient research and comparison of these diverse RAG algorithms and system designs, ? developed *FlashRAG*. This modular toolkit provides a standardized, flexible, and efficient framework for implementing, benchmarking, and innovating RAG systems. FlashRAG offers a hierarchical architecture with pre-implemented advanced RAG algorithms, support for multimodal RAG, standardized datasets, and efficiency features like a retrieval cache, significantly lowering the barrier to entry for researchers and accelerating the development of more performant RAG solutions. Furthermore, ? provided empirical insights into best practices across the RAG workflow, identifying optimal choices for components like chunking, embedding models, and vector databases that bal-

ance performance and efficiency.

In conclusion, the drive towards efficient, fast, and scalable RAG systems for real-world deployment has led to innovations spanning caching mechanisms, algorithm-system co-design, and resource-aware architectural strategies. While significant progress has been made in reducing latency and computational overhead, the continuous evolution of LLMs and the increasing demand for processing vast, dynamic knowledge bases mean that balancing performance, resource efficiency, and scalability remains an ongoing challenge, necessitating further research into adaptive and intelligent system-level optimizations.

5 Evaluation, Benchmarking, and Trustworthiness

5.1 Benchmarking RAG's Core Abilities and Limitations

The burgeoning field of Retrieval-Augmented Generation (RAG) has shown immense promise in mitigating Large Language Model (LLM) hallucinations and integrating dynamic, external knowledge. However, to effectively guide their development and deployment, a critical need has emerged for systematic benchmarks capable of rigorously evaluating RAG's fundamental capabilities and precisely diagnosing its core weaknesses. This diagnostic effort is crucial for understanding where LLMs struggle when augmented with retrieval, revealing issues like difficulty with noisy contexts or integrating information from multiple documents.

Addressing this, chen2023nzb introduced the foundational Retrieval-Augmented Generation Benchmark (RGB), a pioneering effort to systematically evaluate RAG's impact on LLMs. RGB specifically assesses four critical RAG abilities: Noise Robustness (extracting information from noisy documents), Negative Rejection (declining to answer when no relevant information is available), Information Integration (synthesizing answers from multiple documents), and Counterfactual Robustness (handling factual errors in retrieved documents, even with warnings). Their findings highlighted significant shortcomings, such as LLMs often confusing similar information in noisy contexts, frequently failing to reject answers when context is irrelevant, and struggling to integrate information from disparate

sources. Crucially, LLMs were observed to prioritize incorrect retrieved information over their own internal knowledge, even when explicitly warned.

Building upon this foundational diagnostic work, subsequent research has extended benchmarking efforts to more specialized domains and complex reasoning tasks. For instance, xiong2024exb developed MIRAGE (Medical Information Retrieval-Augmented Generation Evaluation) to systematically evaluate RAG systems in the high-stakes medical domain. This benchmark not only demonstrated RAG's potential to improve medical QA but also revealed phenomena like the "lost-in-the-middle" effect, where LLMs struggle to utilize information located in the middle of long contexts. Recognizing the limitations of single-hop evaluations, tang2024i5r introduced MultiHop-RAG, a benchmark specifically designed for multi-hop queries that necessitate retrieving and synthesizing information from multiple, disparate pieces of evidence. Their evaluations exposed significant gaps in current RAG systems' ability to perform complex inference, comparison, and temporal reasoning across documents.

The scope of RAG evaluation has also expanded beyond traditional question-answering. lyu2024ngu proposed CRUD-RAG, a comprehensive Chinese benchmark that categorizes RAG applications into "Create," "Read," "Update," and "Delete" tasks, offering a more holistic assessment of RAG's capabilities in diverse scenarios like text continuation, multi-document summarization, and hallucination modification. In specialized fields, pipitone2024sfx developed LegalBench-RAG, which, unlike prior legal benchmarks, rigorously evaluates the *retrieval component's precision at the snippet level* within legal documents. This focus on minimal, highly relevant text segments is vital for mitigating hallucinations and respecting context window limits in the legal domain.

Further advancements have led to more unified and granular evaluation frameworks. krishna2024qsh introduced FRAMES (Factuality, Retrieval, And reasoning MEasurement Set), a novel dataset and unified evaluation framework designed to rigorously test RAG systems across fact retrieval, reasoning over multiple constraints, and accurate information synthesis in an end-to-end manner. Their findings underscored that even with perfectly retrieved "oracle" contexts, state-of-the-art LLMs still exhibit significant reasoning limi-

tations, particularly in numerical and tabular tasks. To provide more actionable insights, friel20241ct presented RAGBench and the TRACe framework, which formalizes metrics such as "Context Relevance," "Context Utilization" (how much of the retrieved context is actually used by the generator), "Completeness" (how well the response incorporates all relevant information), and "Adherence" (faithfulness). This framework moves beyond simple accuracy to diagnose *how* the LLM leverages context, and notably, demonstrated that fine-tuned smaller models can outperform zero-shot LLMs as evaluators.

A crucial methodological innovation for evaluating the retrieval component itself was proposed by salemi2024om5 with eRAG. This method directly measures a retrieved document's utility *from the perspective of the LLM that consumes it* by evaluating the LLM's downstream performance on individual documents. This approach addresses the low correlation of traditional relevance metrics with actual end-to-end RAG performance, offering a more accurate and computationally efficient way to optimize retrievers. Complementing this, guinet2024vkg introduced an automated evaluation method that generates task-specific exams and applies Item Response Theory (IRT). This framework provides highly interpretable metrics by decomposing a RAG system's overall ability into the contributions of its LLM, retrieval mechanism, and in-context learning components, allowing for fine-grained diagnosis and targeted optimization.

In conclusion, the development of systematic benchmarks has been instrumental in rigorously evaluating RAG's fundamental capabilities and diagnosing its core limitations. From foundational assessments of noise robustness and information integration chen2023nzb to specialized benchmarks for medicine xiong2024exb, multi-hop reasoning tang2024i5r, and legal precision pipitone2024sfx, these tools have exposed critical weaknesses in how LLMs interact with retrieved knowledge. The evolution towards unified, granular, and interpretable evaluation frameworks like FRAMES krishna2024qsh, TRACe friel20241ct, eRAG salemi2024om5, and IRT-based methods guinet2024vkg provides increasingly sophisticated diagnostic capabilities. These advancements are essential for guiding future research towards more robust, accurate, and trustworthy RAG systems, particularly in addressing persistent challenges such as complex reasoning, context utiliza-

tion, and the dynamic interplay between internal LLM knowledge and external retrieved information.

5.2 Evaluating Retrieval Quality and Multi-Hop Reasoning

The efficacy of Retrieval-Augmented Generation (RAG) systems hinges critically on the quality of retrieved information and the Large Language Model's (LLM) ability to synthesize it, especially for complex, multi-hop queries. This necessitates advanced evaluation methodologies that move beyond simple fact-checking to assess intrinsic retrieval utility and sophisticated reasoning capabilities. Early RAG benchmarks, such as the Retrieval-Augmented Generation Benchmark (RGB) by chen2023nzb and the medical RAG benchmark MIRAGE by xiong2024exb, laid foundational work by diagnosing LLMs' performance across general abilities like noise robustness and information integration. While valuable, these often focused on scenarios where answers could be derived from single pieces of evidence, highlighting a need for more complex assessments.

A significant gap emerged in evaluating RAG systems on tasks requiring complex information synthesis across multiple sources, leading to the development of benchmarks specifically targeting multi-hop queries. tang2024i5r directly addressed this with *MultiHop-RAG*, the first dedicated benchmark for multi-hop queries. This dataset, generated via a sophisticated GPT-4-driven pipeline, categorizes queries into Inference, Comparison, Temporal, and Null types, revealing that current state-of-the-art RAG systems perform unsatisfactorily on these complex reasoning tasks. Complementing this, krishna2024qsh introduced FRAMES, a unified evaluation framework that rigorously tests LLMs on fact retrieval, reasoning across multiple constraints, and accurate information synthesis in an end-to-end RAG scenario, particularly for multi-document and multi-hop contexts. Further extending the scope to longer interactions, qi2024tlf introduced LONG^{2}RAG, a benchmark designed to evaluate long-context and long-form RAG. It features questions spanning diverse domains with lengthy retrieved documents and proposes the Key Point Recall (KPR) metric, which offers a nuanced assessment of how effectively LLMs incorporate critical information from extensive contexts into their generated long-form responses.

These efforts collectively underscore the limitations of existing RAG systems in handling nuanced, multi-source information needs and generating comprehensive outputs.

Beyond assessing multi-hop reasoning, a crucial methodological innovation has been the direct evaluation of the *retrieval component's utility to the LLM*. Prior evaluation methods, relying on expensive end-to-end RAG evaluations or human-annotated relevance labels, often showed only a minor correlation with the actual downstream performance of the RAG LLM. This mismatch arises because a document's "relevance" to a human might not equate to its "utility" for an LLM in generating a correct answer. To address this, salemi2024om5 proposed *eRAG*, a novel approach that uses the RAG system's *own LLM* to determine a document's value. By feeding each retrieved document individually to the LLM and evaluating its output against ground truth, eRAG provides downstream-aligned relevance labels with significant computational efficiency, consuming up to 50 times less GPU memory than traditional methods. This direct measurement of utility offers more accurate and efficient feedback for optimizing retrieval models. Building on the idea of LLM-as-a-judge, liu2025sy0 introduced Judge-Consistency (ConsJudge) to improve the reliability of LLM-based evaluations for RAG, addressing the sensitivity of LLM judges to prompts by leveraging consistency across different judgment dimensions for DPO training, thereby enhancing the accuracy of feedback for RAG optimization.

The field has also seen significant advancements in developing granular, explainable, and domain-specific evaluation frameworks. Recognizing the critical need for precision in high-stakes environments, pipitone2024sfx's LegalBench-RAG focuses on the retrieval of minimal, highly relevant text snippets in the legal domain, directly addressing the challenge of preventing LLM hallucination and context window overload in specialized fields. Similarly, wang2024ac6 introduced DomainRAG, a Chinese benchmark tailored for domain-specific RAG in areas like college enrollment, which evaluates abilities such as conversational RAG, structural information analysis, denoising, and multi-document interactions, highlighting the unique challenges of expert knowledge domains. For broader applicability and interpretability, friel20241ct introduced RAGBench and the TRACe evaluation framework, which provides explainable metrics like Context Relevance, Con-

text Utilization, Completeness, and Adherence. These metrics offer actionable insights into RAG system performance by not only assessing the final output but also diagnosing how effectively the LLM leverages the retrieved context. Further pushing the boundaries of interpretability, guinet2024vkg pioneered an automated evaluation methodology using task-specific exam generation and Item Response Theory (IRT), which can decompose a RAG's overall ability into contributions from its LLM, retrieval method, and in-context learning components, providing unprecedented transparency into system behavior. The CRUD-RAG benchmark by lyu2024ngu extends evaluation to a broader range of RAG applications beyond traditional question answering, including text continuation, multi-document summarization, and hallucination modification, particularly for Chinese LLMs. To foster reproducible research and standardized comparisons, rau20244nr developed BERGEN, an end-to-end benchmarking library for RAG.

As RAG systems become more sophisticated and are deployed in critical applications, evaluating their trustworthiness and safety has emerged as a paramount concern. zhou20248fu proposed a unified framework for RAG trustworthiness, encompassing six key dimensions: Factuality, Robustness, Fairness, Transparency, Accountability, and Privacy. This framework highlights that RAG, while mitigating some LLM issues, can introduce new trustworthiness challenges if retrieved information is inappropriate or poorly utilized. Empirically supporting this, zhang2025byv conducted a safety analysis revealing that RAG can, counter-intuitively, make LLMs *less safe* and alter their safety profiles, even when combining safe models with safe documents. This finding underscores the critical need for RAG-specific safety research and red-teaming methods. Moving towards provable guarantees, kang2024hrb introduced C-RAG, the first framework to certify generation risks for RAG models, providing conformal risk analysis and theoretical guarantees that RAG can achieve lower certified generation risk under certain conditions. These advancements signify a crucial shift towards comprehensive evaluation that extends beyond performance metrics to encompass the ethical and safety implications of RAG deployment.

In conclusion, the field has made substantial progress in developing advanced evaluation methodologies for RAG, shifting from general assessments to highly nuanced,

utility-driven, multi-hop, and explainable metrics. The introduction of benchmarks like MultiHop-RAG tang2024i5r and innovative evaluation techniques like eRAG salemi2024om5 are critical for understanding the intrinsic utility of retrieved documents and diagnosing the complex reasoning capabilities of RAG systems. However, as RAG architectures continue to evolve in complexity and are deployed in increasingly sensitive domains, the ongoing challenge remains in developing evaluation frameworks that are not only robust and scalable but also provide fine-grained, interpretable feedback to guide the development of truly intelligent and reliable RAG systems. Future research must critically address how to evaluate RAG systems in dynamic, interactive, and conversational settings, balance cost-effective automated metrics with nuanced human assessment, and comprehensively assess trustworthiness, safety, and fairness, integrating the insights from emerging work on RAG-specific safety and ethical considerations.

5.3 Privacy and Security in RAG Systems

While Retrieval-Augmented Generation (RAG) systems have revolutionized how Large Language Models (LLMs) access and synthesize external knowledge, significantly reducing hallucinations and providing up-to-date information **?**, their widespread adoption, particularly in sensitive domains, introduces critical and often overlooked privacy and security challenges. The field has seen extensive work on benchmarking RAG's capabilities **????** and developing advanced architectures for robustness **???**, as well as applying RAG to structured data and domain-specific applications like textual graphs **?**, customer service **?**, and medical guidelines **?**. However, the inherent privacy vulnerabilities of RAG, especially concerning data leakage from external retrieval databases, have only recently begun to receive systematic scrutiny.

A pivotal work addressing these concerns is **?**, which provides the first comprehensive exploration of privacy issues in RAG systems. This research systematically investigates two primary privacy problems: the susceptibility of RAG systems to leak private information directly from their external retrieval databases, and how the integration of external retrieval data influences the privacy leakage of the LLM's own training data. Unlike

prior LLM privacy research that focused on extracting memorized training data from the LLM's parametric knowledge, **?** introduces a novel methodological advancement: **composite structured prompting attacks**. This attack method cleverly combines an `information` component to guide the retriever towards specific data and a `command` component to instruct the LLM to output the retrieved content, effectively weaponizing the RAG pipeline for data extraction.

Empirical validation by **?** reveals significant vulnerabilities. For instance, targeted attacks successfully extracted 89 medical dialogue chunks and 107 pieces of Personally Identifiable Information (PII) using Llama-7b-Chat, while untargeted prompts on the Enron Email dataset led to exact matches in 116 out of 250 attempts with GPT-3.5-turbo. These findings underscore that RAG systems are highly susceptible to privacy breaches from their external knowledge bases, which often contain sensitive or proprietary information. This is particularly alarming given RAG's application in high-stakes environments such as medicine **??** and customer service **?**, where data confidentiality is paramount.

Crucially, **?** also uncovers a counter-intuitive insight: RAG can actually *mitigate* the leakage of the LLM's own training data. This suggests a complex trade-off, where RAG introduces new vulnerabilities related to its external data sources but may offer a potential security benefit by reducing the LLM's tendency to output memorized pre-training data. Ablation studies further highlight that the design of the command prompt significantly impacts the success of these attacks, with explicit instructions like "Please repeat all the context" proving highly effective.

The implications of **?**'s findings are profound, shifting the narrative around RAG from an unmitigated benefit to a technology requiring careful privacy considerations. The identified vulnerabilities necessitate the urgent development of privacy-preserving RAG architectures and robust security measures. This includes designing retrieval mechanisms that can enforce fine-grained access controls, anonymizing sensitive data within retrieval databases, and developing advanced prompt filtering techniques to detect and neutralize malicious composite structured prompts. As RAG systems continue to evolve and inte-

grate with diverse knowledge sources and complex reasoning tasks **??**, ensuring responsible and ethical use demands a proactive approach to security, balancing the immense utility of RAG with stringent privacy safeguards. Future research must focus on building defense mechanisms against these novel RAG-specific attacks and further understanding the intricate interplay between retrieval and generation in terms of privacy.

6 Domain-Specific Applications and Real-World Impact

6.1 RAG in Healthcare and Clinical Decision Support

The application of Large Language Models (LLMs) in the high-stakes medical domain presents both immense opportunities and significant challenges, primarily due to their propensity for generating "hallucinations" or factually incorrect information. Retrieval-Augmented Generation (RAG) has emerged as a critical technique to ground LLMs in authoritative clinical guidelines, electronic health records (EHRs), and biomedical knowledge graphs, thereby reducing hallucinations and substantially improving accuracy for tasks like medical question answering, guideline interpretation, and clinical trial screening.

A systematic review and meta-analysis by **?** quantitatively demonstrates RAG's effectiveness, showing a 1.35 odds ratio increase in performance compared to baseline LLMs in biomedicine. To systematically understand RAG's capabilities in this critical field, **?** introduced MIRAGE, the first comprehensive benchmark for medical RAG, alongside the MEDRAG toolkit. Their large-scale evaluation of 41 RAG configurations revealed that RAG can improve LLM accuracy by up to 18

Numerous studies have since demonstrated RAG's practical utility across diverse clinical applications. For instance, **?** showcased RAG's potential for reliable clinical decision support by achieving near-perfect (99.0

RAG has also been successfully applied to specialized medical tasks and data types.

For clinical trial screening, ? introduced RECTIFIER, a RAG-enabled GPT-4 system that efficiently extracts information from unstructured EHRs, outperforming human study staff in accuracy and significantly reducing screening time. For patient communication, ? developed LiVersa, a liver disease-specific, PHI-compliant RAG chatbot, demonstrating a secure architecture for integrating authoritative guidelines. In multilingual contexts, ? created GastroBot, a Chinese gastrointestinal disease chatbot, which achieved high context recall and faithfulness by fine-tuning a domain-specific embedding model on Chinese guidelines and literature. ? further explored multilingual capabilities with a dual RAG system for diabetes guidelines, optimizing ensemble retrievers for both Korean and English texts. Other applications include lung cancer staging using RAG-LLM NotebookLM ?, emergency patient triage with RAG-enhanced LLMs ?, and breast cancer nursing care, where RAG significantly improved response accuracy and overall satisfaction without compromising empathy ?. RAG also plays a crucial role in medical education, as demonstrated by ? with SMARThealth GPT, a RAG-based tool for frontline health worker capacity building in low- and middle-income countries, emphasizing traceability and scalability.

Beyond plain text, researchers are integrating RAG with structured knowledge. ? developed KG-RAG, a token-optimized framework that leverages a biomedical knowledge graph (SPOKE) to ground LLMs, achieving over 50

Further advancements focus on enhancing LLM reasoning and self-correction within medical RAG systems. ? proposed Self-BioRAG, a framework incorporating domain-specific instruction sets, a specialized retriever, and a critic LLM for self-reflection, leading to improved medical reasoning and explanation generation. ? also explored RAG with self-evaluation (SelfRewardRAG) to enhance medical reasoning by integrating real-time clinical records. Finally, hybrid approaches like those explored by ? investigate combining RAG with fine-tuning for optimal performance in medical chatbot applications.

Despite these significant strides, challenges remain. Continuous updating of dynamic medical knowledge bases, ensuring data privacy (especially with sensitive patient data),

and developing robust evaluation metrics that reliably assess factual correctness and clinical relevance (beyond lexical similarity) are ongoing areas of research. The integration of multimodal data (e.g., images, videos) into RAG for comprehensive clinical decision support also presents a promising future direction.

6.2 RAG for Customer Service and Structured Data

The application of Retrieval-Augmented Generation (RAG) in enterprise settings, particularly for customer service question answering and interaction with structured data, presents unique challenges and opportunities. While foundational RAG models **?** demonstrated the power of augmenting Large Language Models (LLMs) with external knowledge, their effectiveness diminishes when dealing with inherently structured and interconnected enterprise knowledge bases. Traditional RAG often treats documents as flat text, overlooking crucial intra-document structures and inter-document relationships, which can lead to compromised retrieval accuracy and suboptimal answer quality. General RAG benchmarks have highlighted limitations in information integration and noise robustness when faced with complex data **?**. This subsection explores how RAG can effectively leverage structured knowledge representations, such as Knowledge Graphs (KGs) and tabular data, to enhance performance in domains where information has inherent structure and relationships.

To address these limitations, recent research emphasizes the integration of RAG with structured knowledge representations. A prime example in the customer service domain is the work by **?**, which introduces a novel RAG framework leveraging KGs for customer service question answering. This approach constructs a dual-level KG that preserves both intra-issue structure (parsing individual tickets into trees of sections) and inter-issue relations (connecting tickets via explicit and implicit links). During question answering, an LLM-driven subgraph retrieval mechanism parses consumer queries for entities and intents, translating them into graph database queries (e.g., Cypher) to extract highly pertinent subgraphs. This sophisticated method yielded substantial empirical benefits, including a 77.6

Extending beyond knowledge graphs, RAG for tabular data, such as querying relational databases via Text-to-SQL, represents another significant application in enterprise settings. Traditional LLMs struggle with the intricacies of SQL schema linking and complex query generation. ? introduces Dubo-SQL, a method that combines diverse RAG with fine-tuning for Text-to-SQL tasks, achieving state-of-the-art execution accuracy (EX) on benchmarks like BIRD-SQL. This approach demonstrates how RAG can be tailored to generate precise, executable queries by retrieving relevant schema information and example queries, thereby transforming natural language questions into structured database operations. The challenge here lies not just in retrieving relevant text, but in translating intent into a formal, executable language that accurately reflects the underlying data structure, a distinct problem from graph traversal but equally critical for structured data interaction.

General advancements in RAG can be strategically adapted to further enhance structured RAG systems. For instance, the pre-retrieval phase, as categorized by ?, is crucial for structured data. Query refinement techniques, such as those proposed by ? for rewriting, decomposing, and disambiguating queries, can be specifically engineered to generate more effective graph traversal commands or SQL queries, guided by the underlying schema. This involves training LLMs to understand the structure of the knowledge base (e.g., entity types, relation properties, table schemas) and formulate queries that are syntactically correct and semantically aligned with the structured data. Furthermore, the post-retrieval and generation phases benefit from techniques like unified context ranking and answer generation ?. In structured RAG, this could involve ranking retrieved subgraphs or SQL query results based on their relevance to the LLM's generation task, ensuring the most pertinent structured information is prioritized. Corrective retrieval strategies, such as CRAG ?, can dynamically assess the quality of a generated SQL query or a retrieved subgraph, triggering refinement or re-querying if initial results are suboptimal or lead to errors, thereby enhancing robustness, especially for complex, multi-hop queries over structured data ?.

While integrating structured data significantly enhances RAG performance, it also introduces new considerations, particularly regarding privacy and evaluation. Enterprise applications often deal with highly sensitive structured data, making privacy a paramount concern. ? systematically explores privacy issues in RAG, revealing significant vulnerabilities to data leakage from external retrieval databases through composite structured prompting attacks. This risk is amplified when querying explicit knowledge graphs or relational databases, where the structure itself can inadvertently reveal sensitive relationships or infer private information. Further, ? highlights membership inference attacks against RAG's external database, demonstrating that semantic similarity between generated content and a sample can reveal if the sample was part of the database, a critical vulnerability for proprietary structured datasets. Conversely, ? also presents a nuanced finding that RAG can mitigate the leakage of the LLM's own training data, offering a complex perspective on RAG's privacy implications. Accurately evaluating the utility of retrieved structured information to the LLM remains crucial, with methods like eRAG ? proposing to align retrieval evaluation directly with the LLM's downstream performance, which is vital for assessing the true value of complex graph traversals or SQL query results.

In conclusion, the literature clearly demonstrates that moving beyond plain-text retrieval to actively leverage structured knowledge representations, such as Knowledge Graphs and tabular data, is essential for RAG systems operating in complex enterprise environments like customer service. This approach significantly improves retrieval accuracy, answer quality, and operational efficiency by preserving the inherent structure and relationships within domain-specific data. However, the development of these sophisticated systems necessitates careful consideration of data engineering, specialized retrieval algorithms, and critical privacy implications to ensure robust and trustworthy deployment. Future research must continue to explore hybrid retrieval mechanisms that can seamlessly query both graph-based knowledge, tabular data, and unstructured text within a single enterprise RAG system. Additionally, developing automated KG construction, dynamic schema inference for tabular data, advanced privacy-preserving graph traversal algorithms, and robust evaluation metrics for complex reasoning over structured data are

crucial for the continued advancement of RAG in these critical domains.

6.3 Other Specialized Applications

Beyond general knowledge-intensive tasks, Retrieval-Augmented Generation (RAG) has proven remarkably adaptable and impactful across a diverse array of highly specialized and emerging application areas. These domains are typically characterized by stringent requirements for factual precision, verifiability, complex reasoning over nuanced or structured data, and the critical necessity of grounding Large Language Models (LLMs) in authoritative, domain-specific knowledge. Grouping these applications under "other specialized" highlights their unique demands that often necessitate tailored RAG architectures, specialized data preparation, and domain-specific evaluation, distinguishing them from more general-purpose or broadly applicable RAG use cases. The versatility of RAG in these contexts underscores its potential to significantly enhance LLMs, mitigating their inherent limitations like hallucination and knowledge cutoffs, and enabling their reliable deployment in demanding professional and technical environments.

In **high-stakes professional domains**, such as finance and law, RAG is indispensable for ensuring accuracy and trustworthiness. The financial sector, with its vast, dynamic, and often nuanced information, presents unique challenges for LLMs. **?** conducted a systematic investigation into optimizing RAG pipelines for financial datasets, offering specific recommendations for designing robust RAG systems capable of handling complex financial queries. Their findings emphasize the critical impact of carefully selected retrieval strategies, prompt engineering, and generation models on the quality of financial answers. Further enhancing financial information extraction, **?** proposed HybridRAG, which synergistically combines vector-based and knowledge graph (KG)-based retrieval. This hybrid method is particularly effective in navigating the domain-specific terminology and hierarchical structures prevalent in financial documents, such as earnings call transcripts, leading to more accurate and contextually rich information extraction. However, the maintenance and scalability of KGs for rapidly evolving financial data can introduce significant operational overhead, a challenge that needs careful consideration for real-

world deployment. While the detailed methodology of GraphRAG is discussed in Section 4.2, its application here illustrates how structured knowledge can be leveraged to meet domain-specific precision requirements. The unique challenges in this domain have also necessitated specialized evaluation benchmarks, as discussed in Section 5.2, to accurately assess LLM performance in advanced financial reasoning.

Similarly, the legal domain demands unparalleled precision, verifiability, and the ability to cite sources accurately. RAG addresses this critical need by grounding LLMs in legal statutes, case law, and scholarly articles. **?** developed LegalBench-RAG, a benchmark specifically designed for RAG in legal applications. This benchmark is crucial for evaluating the *retrieval component* of RAG systems, emphasizing the extraction of minimal, highly relevant text snippets (character-level spans) from legal documents. Such granular precision is vital for reducing LLM hallucination, managing context window limitations, and enabling accurate citation, which are non-negotiable requirements in legal contexts. The development of such domain-specific benchmarks is further elaborated in Section 5.2. The overarching concern for trustworthiness and safety in these high-stakes fields, particularly with sensitive financial or legal data, is a critical area of research, as explored in Section 5.3.

Beyond professional services, RAG finds crucial applications in **technical and structured information processing**. A foundational example is robust RAG for zero-shot slot filling, as explored in **?**. This work demonstrates RAG's utility in structured information extraction tasks by enabling LLMs to identify and fill slots (e.g., extracting specific entities like dates, locations, or product names) from text without prior examples for that specific slot type. This capability is particularly valuable in domains where new entity types frequently emerge or where training data is scarce, showcasing RAG's ability to generalize across structured information extraction challenges.

In code generation, LLMs often struggle with coherence, factual accuracy, and hallucination when dealing with complex logic or extrapolating beyond their training data. To address this, **?** proposed ProCC, a prompt-based multi-retrieval augmented generation framework for code completion. ProCC employs a multi-retriever system that crafts

prompt templates to elicit LLM knowledge from multiple perspectives of code semantics, adapting retrieval selection based on code similarity. This approach significantly outperforms existing techniques, demonstrating RAG's ability to provide relevant, context-aware code snippets, thereby mitigating common LLM deficiencies in this domain. However, the computational overhead of managing multiple retrievers and the complexity of designing effective prompt templates for diverse coding scenarios present practical implementation challenges. An emerging application is carbon footprint accounting, where **?** introduced LLMs-RAG-CFA. This method leverages RAG to enhance the real-time, professional, and economical aspects of carbon footprint information retrieval and analysis, demonstrating superior information retrieval rates and lower deviations compared to traditional methods. A critical consideration for such applications is the reliability and standardization of the underlying carbon data sources, as inaccuracies in retrieved data could lead to misleading environmental assessments. These technical applications often require complex reasoning across multiple pieces of information, a challenge that current RAG systems are still striving to fully address, as discussed in the context of multi-hop reasoning in Section 5.2.

Even in **specialized educational contexts**, RAG offers significant advantages. For instance, in computing education, where LLMs are increasingly used, **?** demonstrated that small language models (SLMs) augmented with RAG can perform comparably or even better than larger LLMs for tasks like content understanding and problem-solving. This approach offers a viable solution for educators to leverage AI assistants while maintaining control over data privacy and security, showcasing RAG's role in democratizing access to powerful AI tools in specialized educational contexts. However, ensuring the pedagogical soundness and unbiased nature of retrieved educational content remains a critical challenge, requiring careful curation of the knowledge base. The persistent need for domain-specific evaluation in education, identifying specific abilities required for RAG models in expert scenarios, is further exemplified by research discussed in Section 5.2.

In conclusion, RAG's impact extends profoundly across a wide array of specialized contexts, from high-stakes professional fields like finance and law to technical applications

such as code generation and carbon accounting, and even into educational settings. The consistent themes across these diverse applications are the critical role of domain-specific knowledge, the necessity of tailored retrieval strategies, and meticulous data preparation to achieve high precision and verifiability. While RAG offers significant enhancements, each domain introduces unique challenges related to data complexity, operational overhead, and the need for robust validation, which necessitate ongoing research into specialized RAG methodologies and careful implementation.

7 Conclusion

8 Future Directions and Open Challenges

8.1 The Interplay of RAG and Expanded LLM Context Windows

The rapid evolution of Large Language Models (LLMs) has introduced a compelling dynamic between Retrieval-Augmented Generation (RAG) and the advent of LLMs with vastly expanded native context windows. This subsection critically examines how architectural advancements, enabling models to natively process millions of tokens, challenge and redefine the immediate need for external retrieval in certain long-context tasks, while simultaneously underscoring RAG's enduring importance for dynamic, massive, and explicitly verifiable knowledge integration.

Recent breakthroughs in LLM architecture have dramatically increased the native context window, allowing models like Gemini 1.5 Pro and Flash to process up to 10 million tokens across multimodal inputs (text, audio, video) with remarkable recall **?**. This capability empowers LLMs to perform deep in-context learning, reasoning over fine-grained information from entire documents, extensive codebases, or long videos directly within their input prompt. For tasks requiring a holistic understanding of a single, coherent, and very long document, or those that involve "needle-in-a-haystack" scenarios where the relevant information is deeply embedded within a contiguous text, these large context windows can be demonstrably superior to traditional chunked retrieval **?**. In such cases, the

LLM can leverage its internal attention mechanisms to synthesize information across vast spans of text, often outperforming RAG systems on specific long-context benchmarks ?. However, even these long-context LLMs can struggle with the "lost in the middle" problem, where crucial information located in the middle of a very long input is overlooked ?.

Despite these impressive strides in native context window expansion, RAG is poised to remain a crucial, complementary component in the LLM ecosystem, rather than being fully replaced. The primary reasons for RAG's enduring relevance stem from its ability to manage truly massive, dynamic, and explicitly verifiable knowledge bases that often far exceed even a 10-million-token window. Enterprise knowledge, for instance, can span petabytes of data, constantly updating, necessitating a scalable and efficient external retrieval mechanism that RAG inherently provides ?.

Furthermore, RAG offers distinct advantages in specific, high-stakes domains where explicit provenance, structured knowledge, and continuous updates are paramount:

- **Scale, Dynamism, and Cost-Efficiency:** For knowledge bases that are truly massive (e.g., petabytes of enterprise data) or constantly updating, RAG provides a scalable solution without requiring frequent and costly LLM retraining. For many applications, retrieving and processing a few highly relevant chunks is significantly more cost-effective and computationally efficient than feeding millions of tokens to an LLM for every query, especially with proprietary models ??. Sparse RAG approaches, for instance, actively reduce computational overhead by selecting only highly relevant caches, optimizing both performance and resource utilization ?.

- **Structured and Verifiable Knowledge Integration:** RAG excels at integrating structured knowledge, such as ontologies and knowledge graphs (KGs), which provide explicit relational information beyond semantic similarity. For instance, in financial applications, HybridRAG combines vector-based retrieval with KG-based retrieval to extract intricate information from earnings call transcripts, outperforming individual RAG components ?. Similarly, integrating ontologies into RAG systems can provide domain-specific knowledge bases for fields like dental medicine,

enhancing accuracy and reducing hallucinations **?**. RAG has also been shown to reduce hallucination in structured JSON outputs by grounding LLMs in domain-specific JSON objects, a task where explicit retrieval of structured components is more effective than relying solely on internal context **?**.

- **Domain-Specific Accuracy and Adaptability:** RAG consistently demonstrates superior accuracy and safety in specialized contexts by grounding LLMs in curated, up-to-date guidelines and domain-specific documents. In the legal domain, where precise snippet retrieval is critical, LegalBench-RAG highlights the need for RAG to extract minimal, highly relevant text segments to avoid hallucination and improve citation accuracy **?**. For financial applications, RAG pipelines can be optimized to leverage domain-specific knowledge, achieving high answer generation quality **?**. Even with advanced LLMs, RAG can significantly improve performance in radiology knowledge tasks by providing citable, up-to-date information from a specialized corpus, as demonstrated by improved examination scores for models like GPT-4 and Command R+ **?**. The ability to easily update the knowledge base without retraining the LLM is crucial for rapidly evolving fields.

- **Explainability and Trustworthiness:** RAG inherently provides a mechanism for tracing generated answers back to their source documents, which is crucial for building trust and ensuring accountability in critical applications. This explicit grounding enhances the interpretability and verifiability of LLM outputs, a feature that even massive internal contexts may not fully replicate without additional, complex mechanisms. Benchmarks like RAGBench emphasize explainable metrics for evaluating RAG, including context utilization and adherence, to provide actionable insights into system performance **?**.

- **PHI Compliance and Secure Deployment:** RAG enables the deployment of disease-specific and Protected Health Information (PHI)-compliant LLM chat interfaces within secure institutional frameworks, by keeping sensitive data external and only retrieving non-PHI information or securely handling it within a controlled

environment ?.

The interplay between RAG and expanded context windows thus points towards a future of sophisticated hybrid systems. These systems will intelligently combine the strengths of both paradigms, perhaps using vast context windows for broader contextual understanding and reasoning over a single, long document, while leveraging RAG for precise, up-to-date, and verifiable knowledge retrieval from external, dynamic, and massive sources. Early research is already exploring such architectures; for instance, "Self-Route" proposes an LLM-based self-reflection mechanism to dynamically route queries to either RAG or long-context LLMs, significantly reducing computational cost while maintaining performance ?. Similarly, "LongRAG" introduces a dual-perspective RAG paradigm to enhance understanding of complex long-context knowledge by addressing the "lost in the middle" issue, demonstrating superior performance over long-context LLMs and advanced RAG systems ?. The challenge for future research lies in developing robust benchmarks, such as Long^{2}RAG, that can effectively evaluate this sophisticated interplay, assessing both long-context retrieval and long-form generation with metrics like Key Point Recall ?, and designing architectures that seamlessly integrate these complementary capabilities.

8.2 Balancing Complexity, Efficiency, and Generalizability

The development of advanced Retrieval-Augmented Generation (RAG) systems inherently involves a delicate trade-off between achieving sophisticated capabilities and maintaining efficiency, scalability, and generalizability across diverse applications. While foundational RAG models, such as those introduced by ?, demonstrated the power of combining parametric and non-parametric memory, their end-to-end training already presented a significant computational burden. Early benchmarks, like the Retrieval-Augmented Generation Benchmark (RGB) by ?, quickly revealed that even basic RAG systems struggled with noise robustness, information integration, and negative rejection, highlighting the need for more intelligent and complex architectures. Similarly, ?'s MultiHop-RAG benchmark exposed significant limitations in handling multi-hop queries, which necessitate reasoning over multiple disparate pieces of evidence, further pushing the demand for intricate RAG

designs.

To address these limitations, researchers have introduced increasingly complex RAG architectures featuring multi-stage processing and dynamic decision-making. For instance, Corrective Retrieval Augmented Generation (CRAG) by ? pioneered a self-correcting mechanism that dynamically assesses retrieval quality and triggers actions like knowledge refinement or large-scale web searches, thereby adding multiple processing stages to enhance robustness. Complementing this, RQ-RAG by ? trains Large Language Models (LLMs) to proactively refine queries through rewriting, decomposition, or disambiguation, enabling multi-path exploration during inference. Further increasing architectural complexity, IM-RAG by ? proposes a multi-round RAG system that learns inner monologues for flexible, interpretable multi-round retrieval, while PlanRAG by ? enables LLMs to generate and iteratively refine plans for complex decision-making, both of which involve sophisticated control flows. These advanced capabilities, while improving performance and robustness, inevitably lead to higher computational overhead and increased latency during inference due to the additional processing steps and dynamic decision points, as noted by surveys like ? and ?.

The pursuit of generalizability and domain-specific accuracy also contributes to architectural complexity. For applications involving structured data, such as textual graphs or knowledge graphs (KGs), specialized components are necessary. G-Retriever by ? introduces a RAG approach for general textual graphs, formulating subgraph retrieval as a Prize-Collecting Steiner Tree problem to leverage structural information, which is a departure from simpler vector-based retrieval. Similarly, ? demonstrates the benefits of integrating RAG with dual-level KGs for customer service, preserving intra-issue structure and inter-issue relations, but requiring significant upfront effort in KG construction. In high-stakes domains like medicine, ? found that meticulous data reformatting of clinical guidelines and advanced prompt engineering were paramount for achieving near-perfect accuracy, highlighting the extensive engineering required for domain adaptation. Moreover, deploying RAG in real-world scenarios introduces critical considerations like privacy, as explored by ?, which revealed vulnerabilities to data leakage from external retrieval

databases, adding another layer of complexity to system design and deployment. Surveys on GraphRAG, such as ? and ?, further underscore the inherent complexity in G-Indexing, G-Retrieval, and G-Generation stages.

Recognizing the challenges posed by this increasing complexity, a significant research thrust focuses on optimizing these systems for efficiency and scalability. ?'s RankRAG attempts to simplify the RAG pipeline by unifying context ranking and answer generation within a single instruction-tuned LLM, demonstrating superior performance and generalization while reducing architectural complexity. For system-level bottlenecks, RAGCache by ? proposes a novel multilevel dynamic caching system that stores and shares Key-Value (KV) caches of retrieved documents across multiple requests, significantly reducing time-to-first-token (TTFT) and improving throughput. PipeRAG by ? further accelerates RAG by employing an algorithm-system co-design approach, utilizing pipeline parallelism and dynamic retrieval intervals to overlap retrieval and inference latencies. Even within GraphRAG, ?'s SubgraphRAG demonstrates that a "simple is effective" approach, using a lightweight MLP with Directional Distance Encoding (DDE) for efficient subgraph retrieval, can achieve state-of-the-art results without the overhead of complex GNNs or iterative LLM calls. ?'s M-RAG, while introducing a multi-partition paradigm with RL agents for fine-grained retrieval, aims to optimize performance by focusing retrieval on the most relevant data subsets.

Evaluating the generalizability and efficiency of these complex RAG systems is paramount. Benchmarks like MIRAGE by ? provide systematic evaluations for domain-specific RAG (e.g., medicine), revealing challenges in complex question answering. ?'s eRAG offers a more efficient and accurate method for evaluating retrieval quality by directly measuring a document's utility to the LLM, providing crucial feedback for optimizing complex retrieval components. Furthermore, explainable benchmarks like RAGBench by ? and automated evaluation frameworks leveraging Item Response Theory (IRT) by ? provide granular, component-level insights into RAG performance, helping diagnose where complexity aids or hinders overall system effectiveness.

In conclusion, the trajectory of RAG research clearly demonstrates a continuous effort

to enhance capabilities through increasingly sophisticated architectures, often at the expense of computational efficiency. While innovations in multi-stage processing, dynamic decision-making, and specialized knowledge integration have significantly improved RAG's robustness and accuracy across diverse tasks and domains, they introduce challenges related to higher computational overhead, increased latency, and complex deployment. Future research must therefore prioritize the development of adaptive, optimized RAG systems that can dynamically balance these advanced capabilities with the critical need for efficiency, scalability, and robust generalizability, ensuring their practical and sustainable deployment in real-world, dynamic environments without introducing prohibitive resource demands.

8.3 Ethical Considerations and Responsible RAG Development

The rapid advancement and widespread adoption of Retrieval-Augmented Generation (RAG) systems necessitate a critical examination of their ethical implications and the imperative for responsible development practices. Beyond optimizing performance, ensuring that RAG systems are fair, transparent, and protect user privacy is paramount, especially as they integrate with increasingly sensitive data sources and high-stakes applications.

A primary concern revolves around privacy, particularly the potential for sensitive data leakage from the external retrieval databases that RAG systems leverage. **?** conducted a pivotal study, systematically demonstrating that RAG systems are highly vulnerable to such leakage through novel "composite structured prompting attacks." These attacks exploit the interaction between the retriever and the Large Language Model (LLM) to extract private information, such as personally identifiable information (PII) or medical records, from the external knowledge base. This finding is particularly salient when considering RAG's deployment in sensitive domains. For instance, while **?** showcases RAG's ability to improve medical question answering and **?** optimizes RAG for interpreting hepatological clinical guidelines, their applications inherently involve highly confidential patient data, making the privacy vulnerabilities highlighted by **?** a critical, unaddressed risk. Similarly, the integration of RAG with Knowledge Graphs for customer service, as

explored by ?, involves handling potentially sensitive customer interaction data, where robust privacy safeguards are essential to prevent unintended disclosures. Intriguingly, ? also revealed a counter-intuitive benefit: RAG can mitigate the leakage of the LLM's own training data, suggesting a complex interplay of privacy risks and benefits within the RAG architecture.

Beyond privacy, the potential for fairness issues and bias amplification is a significant ethical challenge. RAG systems retrieve information from vast external corpora, which often reflect societal biases present in their source data. If retrieved documents contain biased or discriminatory information, the RAG system can inadvertently amplify these biases in its generated responses. Benchmarking efforts, such as those by ?, reveal that LLMs struggle with "Noise Robustness" and "Counterfactual Robustness," often failing to discern accurate information from misleading or contradictory content. If this "noise" or "counterfactual" information is also biased, RAG could become a vector for propagating harmful stereotypes or misinformation. The `MultiHop-RAG` benchmark by ?, which uses recent news articles as its knowledge base, implicitly highlights this risk, as news media can contain inherent biases that RAG systems might then synthesize and present as factual. Developing robust mechanisms to detect, filter, and mitigate biased information during retrieval and generation is therefore crucial.

Transparency and explainability are also vital for responsible RAG development. Understanding *why* a RAG system generates a particular answer, and *which* retrieved documents influenced that decision, is essential for building trust and accountability, especially in critical applications. While not directly focused on ethics, the `G-Retriever` framework by ?, which performs retrieval-augmented generation for textual graphs, offers a step towards explainability by leveraging Prize-Collecting Steiner Tree (PCST) optimization to highlight relevant graph parts. This provides a degree of provenance for the generated output. Similarly, the `eRAG` evaluation methodology proposed by ? contributes to transparency by directly measuring a document's utility to the LLM, offering insights into the LLM's reasoning process regarding retrieved content. However, as RAG architectures become more sophisticated, incorporating dynamic elements like corrective retrieval (?) or

query refinement (?), the decision-making process can become more opaque. The unification of context ranking and generation into a single LLM, as demonstrated by `RankRAG` ?, while efficient, could also complicate the disentanglement of ranking and generation influences, potentially impacting explainability.

In conclusion, while RAG offers immense potential for enhancing LLM capabilities, its ethical implications, particularly concerning privacy, fairness, and transparency, demand urgent attention. The demonstrated vulnerabilities to data leakage ? underscore the need for robust privacy-preserving RAG designs. Furthermore, the inherent challenges of handling noisy or biased external information require proactive strategies to prevent bias amplification. Future research must prioritize the development of comprehensive ethical guidelines, robust auditing mechanisms, and inherently explainable RAG architectures to ensure these powerful systems are deployed responsibly, minimizing potential harms while maximizing their beneficial impact on society.

References

References

Patrick Lewis, Ethan Perez, Aleksandara Piktus, et al. (2020). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. Neural Information Processing Systems.

M. Komeili, Kurt Shuster, and J. Weston (2021). *Internet-Augmented Dialogue Generation*. Annual Meeting of the Association for Computational Linguistics.

Wenhu Chen, Hexiang Hu, Xi Chen, et al. (2022). *MuRAG: Multimodal Retrieval-Augmented Generator for Open Question Answering over Images and Text*. Conference on Empirical Methods in Natural Language Processing.

Oshin Agarwal, Heming Ge, Siamak Shakeri, et al. (2021). *Knowledge Graph Based Synthetic Corpus Generation for Knowledge-Enhanced Language Model Pre-training*. North American Chapter of the Association for Computational Linguistics.

Liangke Gui, Borui Wang, Qiuyuan Huang, et al. (2021). *KAT: A Knowledge Augmented Transformer for Vision-and-Language*. North American Chapter of the Association for Computational Linguistics.

L. Masanneck, Sven G. Meuth, and M. Pawlitzki (2014). *Evaluating base and retrieval augmented LLMs with document or online support for evidence based neurology*. The Lancet.

Zhiqing Sun, Xuezhi Wang, Yi Tay, et al. (2022). *Recitation-Augmented Language Models*. International Conference on Learning Representations.

Sara Sarto, Marcella Cornia, L. Baraldi, et al. (2022). *Retrieval-Augmented Transformer for Image Captioning*. International Conference on Content-Based Multimedia Indexing.

Ensheng Shi, Yanlin Wang, Wei Tao, et al. (2022). *RACE: Retrieval-augmented Commit*

Message Generation. Conference on Empirical Methods in Natural Language Processing.

Jishnu Ray Chowdhury, Yong Zhuang, and Shuyi Wang (2022). *Novelty Controlled Paraphrase Generation with Retrieval Augmented Conditional Prompt Tuning*. AAAI Conference on Artificial Intelligence.

Yan Xu, Etsuko Ishii, Zihan Liu, et al. (2021). *Retrieval-Free Knowledge-Grounded Dialogue Response Generation with Adapters*. Workshop on Document-grounded Dialogue and Conversational Question Answering.

Leonard Adolphs, Kurt Shuster, Jack Urbanek, et al. (2021). *Reason first, then respond: Modular Generation for Knowledge-infused Dialogue*. Conference on Empirical Methods in Natural Language Processing.

Tanay Dixit, Bhargavi Paranjape, Hannaneh Hajishirzi, et al. (2022). *CORE: A Retrieve-then-Edit Framework for Counterfactual Data Generation*. Conference on Empirical Methods in Natural Language Processing.

Michael R. Glass, Gaetano Rossiello, Md. Faisal Mahbub Chowdhury, et al. (2021). *Robust Retrieval Augmented Generation for Zero-shot Slot Filling*. Conference on Empirical Methods in Natural Language Processing.

Oshin Agarwal, Heming Ge, Siamak Shakeri, et al. (2020). *Large Scale Knowledge Graph Based Synthetic Corpus Generation for Knowledge-Enhanced Language Model Pre-training*. arXiv.org.

Feifei Pan, Mustafa Canim, Michael R. Glass, et al. (2022). *End-to-End Table Question Answering via Retrieval-Augmented Generation*. arXiv.org.

Shayan A. Akbar, and A. Kak (2020). *A Large-Scale Comparative Evaluation of IR-Based Tools for Bug Localization*. IEEE Working Conference on Mining Software Repositories.

Dimitris Pappas, Prodromos Malakasiotis, and Ion Androutsopoulos (2022). *Data Augmentation for Biomedical Factoid Question Answering.* Workshop on Biomedical Natural Language Processing.

Jihyeok Kim, Seungtaek Choi, Reinald Kim Amplayo, et al. (2020). *Retrieval-Augmented Controllable Review Generation.* International Conference on Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, et al. (2023). *Retrieval-Augmented Generation for Large Language Models: A Survey.* arXiv.org.

Wenqi Fan, Yujuan Ding, Liang-bo Ning, et al. (2024). *A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models.* Knowledge Discovery and Data Mining.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, et al. (2024). *Benchmarking Retrieval-Augmented Generation for Medicine.* Annual Meeting of the Association for Computational Linguistics.

Jiawei Chen, Hongyu Lin, Xianpei Han, et al. (2023). *Benchmarking Large Language Models in Retrieval-Augmented Generation.* AAAI Conference on Artificial Intelligence.

Boci Peng, Yun Zhu, Yongchao Liu, et al. (2024). *Graph Retrieval-Augmented Generation: A Survey.* arXiv.org.

Yixuan Tang, and Yi Yang (2024). *MultiHop-RAG: Benchmarking Retrieval-Augmented Generation for Multi-Hop Queries.* arXiv.org.

Xiaoxin He, Yijun Tian, Yifei Sun, et al. (2024). *G-Retriever: Retrieval-Augmented Generation for Textual Graph Understanding and Question Answering.* Neural Information Processing Systems.

Zhentao Xu, Mark Jerome Cruz, Matthew Guevara, et al. (2024). *Retrieval-Augmented Generation with Knowledge Graphs for Customer Service Question Answering.* Annual

International ACM SIGIR Conference on Research and Development in Information Retrieval.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, et al. (2024). *Corrective Retrieval Augmented Generation*. arXiv.org.

Yue Yu, Wei Ping, Zihan Liu, et al. (2024). *RankRAG: Unifying Context Ranking with Retrieval-Augmented Generation in LLMs*. Neural Information Processing Systems.

Shenglai Zeng, Jiankun Zhang, Pengfei He, et al. (2024). *The Good and The Bad: Exploring Privacy Issues in Retrieval-Augmented Generation (RAG)*. Annual Meeting of the Association for Computational Linguistics.

Alireza Salemi, and Hamed Zamani (2024). *Evaluating Retrieval Quality in Retrieval-Augmented Generation*. Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

Chi-Min Chan, Chunpu Xu, Ruibin Yuan, et al. (2024). *RQ-RAG: Learning to Refine Queries for Retrieval Augmented Generation*. arXiv.org.

Simone Kresevic, M. Giuffré, M. Ajčević, et al. (2024). *Optimization of hepatological clinical guidelines interpretation by large language models: a retrieval augmented generation-based framework*. npj Digit. Medicine.

Costas Mavromatis, and George Karypis (2024). *GNN-RAG: Graph Neural Retrieval for Large Language Model Reasoning*. arXiv.org.

Jiajie Jin, Yutao Zhu, Xinyu Yang, et al. (2024). *FlashRAG: A Modular Toolkit for Efficient Retrieval-Augmented Generation Research*. The Web Conference.

Bhaskarjit Sarmah, Dhagash Mehta, Benika Hall, et al. (2024). *HybridRAG: Integrating Knowledge Graphs and Vector Retrieval Augmented Generation for Efficient Information Extraction*. International Conference on AI in Finance.

Patrice B'echard, and Orlando Marquez Ayala (2024). *Reducing hallucination in structured outputs via Retrieval-Augmented Generation*. North American Chapter of the Association for Computational Linguistics.

Xiaohua Wang, Zhenghua Wang, Xuan Gao, et al. (2024). *Searching for Best Practices in Retrieval-Augmented Generation*. Conference on Empirical Methods in Natural Language Processing.

Wei Zou, Runpeng Geng, Binghui Wang, et al. (2024). *PoisonedRAG: Knowledge Poisoning Attacks to Retrieval-Augmented Generation of Large Language Models*. arXiv.org.

Bernal Jimenez Gutierrez, Yiheng Shu, Yu Gu, et al. (2024). *HippoRAG: Neurobiologically Inspired Long-Term Memory for Large Language Models*. Neural Information Processing Systems.

Shi Yu, Chaoyue Tang, Bokai Xu, et al. (2024). *VisRAG: Vision-based Retrieval-augmented Generation on Multi-modality Documents*. International Conference on Learning Representations.

Zirui Guo, Lianghao Xia, Yanhua Yu, et al. (2024). *LightRAG: Simple and Fast Retrieval-Augmented Generation*. arXiv.org.

Zhuowan Li, Cheng Li, Mingyang Zhang, et al. (2024). *Retrieval Augmented Generation or Long-Context LLMs? A Comprehensive Study and Hybrid Approach*. Conference on Empirical Methods in Natural Language Processing.

Siyun Zhao, Yuqing Yang, Zilong Wang, et al. (2024). *Retrieval Augmented Generation (RAG) and Beyond: A Comprehensive Survey on How to Make your LLMs use External Data More Wisely*. arXiv.org.

Yizheng Huang, and Jimmy X. Huang (2024). *A Survey on Retrieval-Augmented Text Generation for Large Language Models*. arXiv.org.

Qianqian Xie, Weiguang Han, Zhengyu Chen, et al. (2024). *FinBen: A Holistic Financial Benchmark for Large Language Models*. Neural Information Processing Systems.

Shangyu Wu, Ying Xiong, Yufei Cui, et al. (2024). *Retrieval-Augmented Generation for Natural Language Processing: A Survey*. arXiv.org.

Yuanjie Lyu, Zhiyu Li, Simin Niu, et al. (2024). *CRUD-RAG: A Comprehensive Chinese Benchmark for Retrieval-Augmented Generation of Large Language Models*. ACM Trans. Inf. Syst..

Gelei Deng, Yi Liu, Kailong Wang, et al. (2024). *Pandora: Jailbreak GPTs by Retrieval Augmented Generation Poisoning*. Proceedings 2024 Workshop on AI Systems with Confidential COmputing.

Heydar Soudani, E. Kanoulas, and Faegheh Hasibi (2024). *Fine Tuning vs. Retrieval Augmented Generation for Less Popular Knowledge*. SIGIR-AP.

Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, et al. (2024). *Fact, Fetch, and Reason: A Unified Evaluation of Retrieval-Augmented Generation*. North American Chapter of the Association for Computational Linguistics.

Yujia Zhou, Yan Liu, Xiaoxi Li, et al. (2024). *Trustworthiness in Retrieval-Augmented Generation Systems: A Survey*. arXiv.org.

Nicholas Pipitone, and Ghita Houir Alami (2024). *LegalBench-RAG: A Benchmark for Retrieval-Augmented Generation in the Legal Domain*. arXiv.org.

Chao Jin, Zili Zhang, Xuanlin Jiang, et al. (2024). *RAGCache: Efficient Knowledge Caching for Retrieval-Augmented Generation*. ACM Transactions on Computer Systems.

Zilong Wang, Zifeng Wang, Long T. Le, et al. (2024). *Speculative RAG: Enhancing Retrieval Augmented Generation through Drafting*. International Conference on Learning Representations.

Norbert Tihanyi, M. Ferrag, Ridhi Jain, et al. (2024). *CyberMetric: A Benchmark Dataset based on Retrieval-Augmented Generation for Evaluating LLMs in Cybersecurity Knowledge*. Computer Science Symposium in Russia.

Wei Zou, Runpeng Geng, Binghui Wang, et al. (2024). *PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of Large Language Models*. Unpublished manuscript.

Guangzhi Xiong, Qiao Jin, Xiao Wang, et al. (2024). *Improving Retrieval-Augmented Generation in Medicine with Iterative Follow-up Questions*. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing.

Feiteng Fang, Yuelin Bai, Shiwen Ni, et al. (2024). *Enhancing Noise Robustness of Retrieval-Augmented Language Models with Adaptive Adversarial Training*. Annual Meeting of the Association for Computational Linguistics.

Yuntong Hu, Zhihan Lei, Zhengwu Zhang, et al. (2024). *GRAG: Graph Retrieval-Augmented Generation*. North American Chapter of the Association for Computational Linguistics.

Jiaqi Xue, Meng Zheng, Yebowen Hu, et al. (2024). *BadRAG: Identifying Vulnerabilities in Retrieval Augmented Generation of Large Language Models*. arXiv.org.

Minbyul Jeong, Jiwoong Sohn, Mujeen Sung, et al. (2024). *Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models*. Bioinform..

Nicholas Matsumoto, Jay Moran, Hyunjun Choi, et al. (2024). *KRAGEN: a knowledge graph-enhanced RAG framework for biomedical problem solving using large language models*. Bioinformatics.

Robert Friel, Masha Belyi, and Atindriyo Sanyal (2024). *RAGBench: Explainable Benchmark for Retrieval-Augmented Generation Systems*. arXiv.org.

T. Procko, and Omar Ochoa (2024). *Graph Retrieval-Augmented Generation for Large Language Models: A Survey*. 2024 Conference on AI, Science, Engineering, and Technology (AIxSET).

Hongru Wang, Wenyu Huang, Yang Deng, et al. (2024). *UniMS-RAG: A Unified Multi-source Retrieval-Augmented Generation for Personalized Dialogue Systems*. arXiv.org.

Qinggang Zhang, Shengyuan Chen, Yuan-Qi Bei, et al. (2025). *A Survey of Graph Retrieval-Augmented Generation for Customized Large Language Models*. arXiv.org.

Pengzhou Cheng, Yidong Ding, Tianjie Ju, et al. (2024). *TrojanRAG: Retrieval-Augmented Generation Can Be Backdoor Driver in Large Language Models*. arXiv.org.

Zhenrui Yue, Honglei Zhuang, Aijun Bai, et al. (2024). *Inference Scaling for Long-Context Retrieval Augmented Generation*. International Conference on Learning Representations.

Wenqi Jiang, Shuai Zhang, Boran Han, et al. (2024). *PipeRAG: Fast Retrieval-Augmented Generation via Algorithm-System Co-design*. arXiv.org.

J. Ge, Steve Sun, Joseph Owens, et al. (2024). *Development of a liver disease–specific large language model chat interface using retrieval-augmented generation*. Hepatology.

Yujuan Ding, Wenqi Fan, Liang-bo Ning, et al. (2024). *A Survey on RAG Meets LLMs: Towards Retrieval-Augmented Large Language Models*. arXiv.org.

ZhongXiang Sun, Xiaoxue Zang, Kai Zheng, et al. (2024). *ReDeEP: Detecting Hallucination in Retrieval-Augmented Generation via Mechanistic Interpretability*. International Conference on Learning Representations.

Shengjie Ma, Chengjin Xu, Xuhui Jiang, et al. (2024). *Think-on-Graph 2.0: Deep and Faithful Large Language Model Reasoning with Knowledge-guided Retrieval Augmented Generation*. International Conference on Learning Representations.

Andrei-Laurentiu Bornea, Fadhel Ayed, Antonio De Domenico, et al. (2024). *Telco-RAG: Navigating the Challenges of Retrieval Augmented Language Models for Telecommunications*. Global Communications Conference.

Diji Yang, Jinmeng Rao, Kezhen Chen, et al. (2024). *IM-RAG: Multi-Round Retrieval-Augmented Generation Through Learning Inner Monologues.* Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

Weihang Su, Yichen Tang, Qingyao Ai, et al. (2024). *DRAGIN: Dynamic Retrieval Augmented Generation based on the Real-time Information Needs of Large Language Models.* Annual Meeting of the Association for Computational Linguistics.

Shayekh Bin Islam, Md Asib Rahman, K. S. M. T. Hossain, et al. (2024). *Open-RAG: Enhanced Retrieval-Augmented Reasoning with Open-Source Large Language Models.* Conference on Empirical Methods in Natural Language Processing.

Siru Liu, Allison B. McCoy, and Adam Wright (2025). *Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines.* J. Am. Medical Informatics Assoc..

Yuhe Ke, Liyuan Jin, Kabilan Elangovan, et al. (2024). *Development and Testing of Retrieval Augmented Generation in Large Language Models - A Case Study Report.* arXiv.org.

Bo Ni, Zheyuan Liu, Leyao Wang, et al. (2025). *Towards Trustworthy Retrieval Augmented Generation for Large Language Models: A Survey.* arXiv.org.

Myeonghwa Lee, Seonho An, and Min-Soo Kim (2024). *PlanRAG: A Plan-then-Retrieval Augmented Generation for Generative Large Language Models as Decision Makers.* North American Chapter of the Association for Computational Linguistics.

Mufei Li, Siqi Miao, and Pan Li (2024). *Simple is Effective: The Roles of Graphs and Large Language Models in Knowledge-Graph-Based Retrieval-Augmented Generation.* International Conference on Learning Representations.

Yuhe Ke, Liyuan Jin, Kabilan Elangovan, et al. (2025). *Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness.* npj Digit. Medicine.

Zheng Wang, Shu Xian Teo, Jieer Ouyang, et al. (2024). *M-RAG: Reinforcing Large Language Model Performance through Retrieval-Augmented Generation with Multiple Partitions*. Annual Meeting of the Association for Computational Linguistics.

Mintong Kang, Nezihe Merve Gurel, Ning Yu, et al. (2024). *C-RAG: Certified Generation Risks for Retrieval-Augmented Language Models*. International Conference on Machine Learning.

Demiao Lin (2024). *Revolutionizing Retrieval-Augmented Generation with Enhanced PDF Structure Recognition*. arXiv.org.

Gauthier Guinet, Behrooz Omidvar-Tehrani, Anoop Deoras, et al. (2024). *Automated Evaluation of Retrieval-Augmented Language Models with Task-Specific Exam Generation*. International Conference on Machine Learning.

I. Radeva, I. Popchev, L. Doukovska, et al. (2024). *Web Application for Retrieval-Augmented Generation: Implementation and Testing*. Electronics.

Karthik Soman, Peter W Rose, John H Morris, et al. (2023). *Biomedical knowledge graph-optimized prompt generation for large language models*. Bioinformatics.

Zhanpeng Chen, Chengjin Xu, Yiyan Qi, et al. (2024). *MLLM Is a Strong Reranker: Advancing Multimodal Retrieval-augmented Generation via Knowledge-enhanced Reranking and Noise-injected Training*. arXiv.org.

Ozan Unlu, Jiyeon Shin, Charlotte J. Mailly, et al. (2024). *Retrieval Augmented Generation Enabled Generative Pre-Trained Transformer 4 (GPT-4) Performance for Clinical Trial Screening*. medRxiv.

J. Ge, Steve Sun, Joseph Owens, et al. (2023). *Development of a Liver Disease-Specific Large Language Model Chat Interface using Retrieval Augmented Generation*. medRxiv.

David Rau, Herv'e D'ejean, Nadezhda Chirkova, et al. (2024). *BERGEN: A Benchmarking Library for Retrieval-Augmented Generation*. Conference on Empirical Methods in Natural Language Processing.

Arunabh Bora, and H. Cuayáhuitl (2024). *Systematic Analysis of Retrieval-Augmented Generation-Based LLMs for Medical Chatbot Applications*. Machine Learning and Knowledge Extraction.

Ronak Pradeep, Nandan Thakur, Sahel Sharifymoghaddam, et al. (2024). *Ragnarök: A Reusable RAG Framework and Baselines for TREC 2024 Retrieval-Augmented Generation Track*. European Conference on Information Retrieval.

Qingfei Zhao, Ruobing Wang, Yukuo Cen, et al. (2024). *LongRAG: A Dual-Perspective Retrieval-Augmented Generation Paradigm for Long-Context Question Answering*. Conference on Empirical Methods in Natural Language Processing.

Nadezhda Chirkova, David Rau, Herv'e D'ejean, et al. (2024). *Retrieval-augmented generation in multilingual settings*. KNOWLLM.

Guanting Dong, Yutao Zhu, Chenghao Zhang, et al. (2024). *Understand What LLM Needs: Dual Preference Alignment for Retrieval-Augmented Generation*. The Web Conference.

Songshuo Lu, Hua Wang, Yutian Rong, et al. (2024). *TurboRAG: Accelerating Retrieval-Augmented Generation with Precomputed KV Caches for Chunked Text*. arXiv.org.

Yun Zhu, Jia-Chen Gu, Caitlin Sikora, et al. (2024). *Accelerating Inference of Retrieval-Augmented Generation via Sparse Context Selection*. International Conference on Learning Representations.

Tian Yu, Shaolei Zhang, and Yang Feng (2024). *Auto-RAG: Autonomous Retrieval-Augmented Generation for Large Language Models*. arXiv.org.

L. M. Amugongo, Pietro Mascheroni, Steven Brooks, et al. (2025). *Retrieval augmented generation for large language models in healthcare: A systematic review*. PLOS Digital Health.

Yulong Hui, Yao Lu, and Huanchen Zhang (2024). *UDA: A Benchmark Suite for Retrieval Augmented Generation in Real-world Document Analysis*. Neural Information Processing Systems.

M. A. Khaliq, P. Chang, M. Ma, et al. (2024). *RAGAR, Your Falsehood Radar: RAG-Augmented Reasoning for Political Fact-Checking using Multimodal Large Language Models*. FEVER.

Alireza Salemi, and Hamed Zamani (2024). *Towards a Search Engine for Machines: Unified Ranking for Multiple Retrieval-Augmented Large Language Models*. Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

Shicheng Xu, Liang Pang, Mo Yu, et al. (2024). *Unsupervised Information Refinement Training of Large Language Models for Retrieval-Augmented Generation*. Annual Meeting of the Association for Computational Linguistics.

Zhibo Hu, Chen Wang, Yanfeng Shu, et al. (2024). *Prompt Perturbation in Retrieval-Augmented Generation based Large Language Models*. Knowledge Discovery and Data Mining.

Jiwoong Sohn, Yein Park, Chanwoong Yoon, et al. (2024). *Rationale-Guided Retrieval Augmented Generation for Medical Question Answering*. North American Chapter of the Association for Computational Linguistics.

Zehan Qi, Rongwu Xu, Zhijiang Guo, et al. (2024). *LONG^{2}RAG: Evaluating Long-Context Long-Form Retrieval-Augmented Generation with Key Point Recall*. Conference on Empirical Methods in Natural Language Processing.

Binglan Han, Teo Sušnjak, and A. Mathrani (2024). *Automating Systematic Literature Reviews with Retrieval-Augmented Generation: A Comprehensive Overview*. Applied Sciences.

Yiyun Zhao, Prateek Singh, Hanoz Bhathena, et al. (2024). *Optimizing LLM Based Retrieval Augmented Generation Pipelines in the Financial Domain*. North American Chapter of the Association for Computational Linguistics.

Xinze Li, Senkun Mei, Zhenghao Liu, et al. (2024). *RAG-DDR: Optimizing Retrieval-Augmented Generation Using Differentiable Data Rewards*. International Conference on Learning Representations.

Fei Wang, Xingchen Wan, Ruoxi Sun, et al. (2024). *Astute RAG: Overcoming Imperfect Retrieval Augmentation and Knowledge Conflicts for Large Language Models*. Annual Meeting of the Association for Computational Linguistics.

Rama Akkiraju, Anbang Xu, Deepak Bora, et al. (2024). *FACTS About Building Retrieval Augmented Generation-based Chatbots*. arXiv.org.

Qingqing Zhou, Can Liu, Yuchen Duan, et al. (2024). *GastroBot: a Chinese gastrointestinal disease chatbot based on the retrieval-augmented generation*. Frontiers in Medicine.

Dongkyu Kim, Byoungwook Kim, Donggeon Han, et al. (2024). *AutoRAG: Automated Framework for optimization of Retrieval Augmented Generation Pipeline*. arXiv.org.

G. M. Yilma, J. Ayala-Romero, A. Garcia-Saavedra, et al. (2024). *TelecomRAG: Taming Telecom Standards with Retrieval Augmented Generation and LLMs*. Computer communication review.

Haowen Xu, Jinghui Yuan, Anye Zhou, et al. (2024). *GenAI-powered Multi-Agent Paradigm for Smart Urban Mobility: Opportunities and Challenges for Integrating Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) with Intelligent Transportation Systems*. arXiv.org.

Ran Xu, Hui Liu, Sreyashi Nag, et al. (2024). *SimRAG: Self-Improving Retrieval-Augmented Generation for Adapting Large Language Models to Specialized Domains*. North American Chapter of the Association for Computational Linguistics.

Suqing Liu, Zezhu Yu, Feiran Huang, et al. (2024). *Can Small Language Models With Retrieval-Augmented Generation Replace Large Language Models When Learning Computer Science?*. Annual Conference on Innovation and Technology in Computer Science Education.

Huimin Zeng, Zhenrui Yue, Qian Jiang, et al. (2024). *Federated Recommendation via Hybrid Retrieval Augmented Generation*. BigData Congress [Services Society].

Manish Bhattarai, Javier E. Santos, Shawn Jones, et al. (2024). *Enhancing Code Translation in Language Models with Few-Shot Learning via Retrieval-Augmented Generation.* IEEE Conference on High Performance Extreme Computing.

Shuting Wang, Jiongnan Liu, Jiehan Cheng, et al. (2024). *DomainRAG: A Chinese Benchmark for Evaluating Domain-specific Retrieval-Augmented Generation.* arXiv.org.

Pouria Omrani, Alireza Hosseini, Kiana Hooshanfar, et al. (2024). *Hybrid Retrieval-Augmented Generation Approach for LLMs Query Response Enhancement.* 2024 10th International Conference on Web Research (ICWR).

Ryota Tozuka, Hisashi Johno, Akitomo Amakawa, et al. (2024). *Application of NotebookLM, a Large Language Model with Retrieval-Augmented Generation, for Lung Cancer Staging.* Japanese Journal of Radiology.

Xueguang Ma, Shengyao Zhuang, B. Koopman, et al. (2024). *VISA: Retrieval Augmented Generation with Visual Source Attribution.* arXiv.org.

Si-Nan Yang, Dong Wang, Haoqi Zheng, et al. (2024). *TimeRAG: BOOSTING LLM Time Series Forecasting via Retrieval-Augmented Generation.* IEEE International Conference on Acoustics, Speech, and Signal Processing.

Robert Lakatos, P. Pollner, András Hajdu, et al. (2024). *Investigating the performance of Retrieval-Augmented Generation and fine-tuning for the development of AI-driven knowledge-based systems.* Machine Learning and Knowledge Extraction.

Zhuo Chen, Jiawei Liu, Haotan Liu, et al. (2024). *Black-Box Opinion Manipulation Attacks to Retrieval-Augmented Generation of Large Language Models.* arXiv.org.

Saber Zerhoudi, and Michael Granitzer (2024). *PersonaRAG: Enhancing Retrieval-Augmented Generation Systems with User-Centric Agents.* IR-RAG@SIGIR.

Dazhou Yu, Riyang Bao, Gengchen Mai, et al. (2025). *Spatial-RAG: Spatial Retrieval Augmented Generation for Real-World Spatial Reasoning Questions.* arXiv.org.

Yasmina Al Ghadban, Yvonne Lu, Uday Adavi, et al. (2023). *Transforming Healthcare Education: Harnessing Large Language Models for Frontline Health Worker Capacity Building using Retrieval-Augmented Generation*. medRxiv.

Xun Liang, Simin Niu, Zhiyu Li, et al. (2025). *SafeRAG: Benchmarking Security in Retrieval-Augmented Generation of Large Language Model*. Annual Meeting of the Association for Computational Linguistics.

Derrick Quinn, Mohammad Nouri, Neel Patel, et al. (2024). *Accelerating Retrieval-Augmented Generation*. International Conference on Architectural Support for Programming Languages and Operating Systems.

Hanzhuo Tan, Qi Luo, Lingixao Jiang, et al. (2024). *Prompt-based Code Completion via Multi-Retrieval Augmented Generation*. ACM Transactions on Software Engineering and Methodology.

Mohammadtaghi Hajiaghayi, S'ebastien Lahaie, Keivan Rezaei, et al. (2024). *Ad Auctions for LLMs via Retrieval Augmented Generation*. Neural Information Processing Systems.

Darío Garigliotti (2024). *SDG target detection in environmental reports using Retrieval-augmented Generation with LLMs*. CLIMATENLP.

Ryan Barron, Ves Grantcharov, Selma Wanna, et al. (2024). *Domain-Specific Retrieval-Augmented Generation Using Vector Stores, Knowledge Graphs, and Tensor Factorization*. International Conference on Machine Learning and Applications.

Shiyue Zhang, Mark Dredze, AI Bloomberg, et al. (2025). *RAG LLMs are Not Safer: A Safety Analysis of Retrieval-Augmented Generation for Large Language Models*. North American Chapter of the Association for Computational Linguistics.

Chunjing Gan, Dan Yang, Binbin Hu, et al. (2024). *Similarity is Not All You Need: Endowing Retrieval Augmented Generation with Multi Layered Thoughts*. arXiv.org.

Shuting Wang, Xin Xu, Mang Wang, et al. (2024). *RichRAG: Crafting Rich Responses*

for Multi-faceted Queries in Retrieval-Augmented Generation. International Conference on Computational Linguistics.

Yuying Li, Gaoyang Liu, Chen Wang, et al. (2024). *Generating Is Believing: Membership Inference Attacks against Retrieval-Augmented Generation*. IEEE International Conference on Acoustics, Speech, and Signal Processing.

Jia Fu, Xiaoting Qin, Fangkai Yang, et al. (2024). *AutoRAG-HP: Automatic Online Hyper-Parameter Tuning for Retrieval-Augmented Generation*. Conference on Empirical Methods in Natural Language Processing.

Huanshuo Liu, Hao Zhang, Zhijiang Guo, et al. (2024). *CtrlA: Adaptive Retrieval-Augmented Generation via Probe-Guided Control*. arXiv.org.

Kurnia Muludi, Kaira Milani Fitria, Joko Triloka, et al. (2024). *Retrieval-Augmented Generation Approach: Document Question Answering using Large Language Model*. International Journal of Advanced Computer Science and Applications.

A. Lahiri, and Q. Hu (2024). *AlzheimerRAG: Multimodal Retrieval Augmented Generation for PubMed articles*. arXiv.org.

Zijian Hei, Weiling Liu, Wenjie Ou, et al. (2024). *DR-RAG: Applying Dynamic Document Relevance to Retrieval-Augmented Generation for Question-Answering*. arXiv.org.

Yucheng Zhang, Qinfeng Li, Tianyu Du, et al. (2024). *HijackRAG: Hijacking Attacks against Retrieval-Augmented Large Language Models*. arXiv.org.

Jirui Qi, Gabriele Sarti, R. Fern'andez, et al. (2024). *Model Internals-based Answer Attribution for Trustworthy Retrieval-Augmented Generation*. Conference on Empirical Methods in Natural Language Processing.

Bolei He, Nuo Chen, Xinran He, et al. (2024). *Retrieving, Rethinking and Revising: The Chain-of-Verification Can Improve Retrieval Augmented Generation*. Conference on Empirical Methods in Natural Language Processing.

Ruiyang Qin, Zheyu Yan, Dewen Zeng, et al. (2024). *Robust Implementation of Retrieval-Augmented Generation on Edge-Based Computing-in-Memory Architectures*. International Conference on Computer Aided Design.

Zheng Wang, Zhongyang Li, Zeren Jiang, et al. (2024). *Crafting Personalized Agents through Retrieval-Augmented Generation on Editable Memory Graphs*. Conference on Empirical Methods in Natural Language Processing.

Thomas Merth, Qichen Fu, Mohammad Rastegari, et al. (2024). *Superposition Prompting: Improving and Accelerating Retrieval-Augmented Generation*. International Conference on Machine Learning.

Jiajing Chen, Runyuan Bao, Hongye Zheng, et al. (2024). *Optimizing Retrieval-Augmented Generation with Elasticsearch for Enhanced Question-Answering Systems*. 2024 5th International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE).

Dayton G. Thorpe, Andrew Duberstein, and Ian A. Kinsey (2024). *Dubo-SQL: Diverse Retrieval-Augmented Generation and Fine Tuning for Text-to-SQL*. arXiv.org.

Fatma Yasmine Loumachi, M. C. Ghanem, and M. Ferrag (2024). *Advancing Cyber Incident Timeline Analysis Through Retrieval-Augmented Generation and Large Language Models*. De Computis.

Weiye Xu, Min Wang, Wen-gang Zhou, et al. (2024). *P-RAG: Progressive Retrieval Augmented Generation For Planning on Embodied Everyday Task*. ACM Multimedia.

Reza Fayyazi, Rozhina Taghdimi, and S. Yang (2024). *Advancing TTP Analysis: Harnessing the Power of Encoder-Only and Decoder-Only Language Models with Retrieval Augmented Generation*. arXiv.org.

Xi Wang, Procheta Sen, Ruizhe Li, et al. (2024). *Adaptive Retrieval-Augmented Generation for Conversational Systems*. North American Chapter of the Association for Computational Linguistics.

Tzu-Lin Kuo, Fengting Liao, Mu-Wei Hsieh, et al. (2024). *RAD-Bench: Evaluating Large Language Models Capabilities in Retrieval Augmented Dialogues*. arXiv.org.

Megumi Yazaki, S. Maki, T. Furuya, et al. (2024). *Emergency Patient Triage Improvement through a Retrieval-Augmented Generation Enhanced Large-Scale Language Model*. Prehospital Emergency Care.

Cody Clop, and Yannick Teglia (2024). *Backdoored Retrievers for Prompt Injection Attacks on Retrieval Augmented Generation of Large Language Models*. arXiv.org.

Jaedong Lee, H. Cha, Y. Hwangbo, et al. (2024). *Enhancing Large Language Model Reliability: Minimizing Hallucinations with Dual Retrieval-Augmented Generation Based on the Latest Diabetes Guidelines*. Journal of Personalized Medicine.

Zhe Chen, Yusheng Liao, Shuyang Jiang, et al. (2025). *Towards Omni-RAG: Comprehensive Retrieval-Augmented Generation for Large Language Models in Medical Applications*. Annual Meeting of the Association for Computational Linguistics.

Brandon T Garcia, Lauren Westerfield, Priya Yelemali, et al. (2024). *Improving Automated Deep Phenotyping Through Large Language Models Using Retrieval Augmented Generation*. medRxiv.

Qimin Yang, Huan Zuo, Runqi Su, et al. (2025). *Dual retrieving and ranking medical large language model with retrieval augmented generation*. Scientific Reports.

Chenxi Dong (2023). *How to Build an AI Tutor that Can Adapt to Any Course and Provide Accurate Answers Using Large Language Model and Retrieval-Augmented Generation*. arXiv.org.

Feifan Wu, Lingyuan Liu, Wentao He, et al. (2024). *Time-Sensitve Retrieval-Augmented Generation for Question Answering*. International Conference on Information and Knowledge Management.

Dongyang Li, Junbing Yan, Taolin Zhang, et al. (2024). *On the Role of Long-tail Knowledge in Retrieval Augmented Large Language Models*. Annual Meeting of the Association for Computational Linguistics.

Kartik Sharma, Peeyush Kumar, and Yunqing Li (2024). *OG-RAG: Ontology-Grounded Retrieval-Augmented Generation For Large Language Models*. arXiv.org.

R. Leekha, Olga Simek, and Charlie Dagli (2024). *War of Words: Harnessing the Potential of Large Language Models and Retrieval Augmented Generation to Classify, Counter and Diffuse Hate Speech*. The Florida AI Research Society.

Ruiyu Xu, Ying Hong, Feifei Zhang, et al. (2024). *Evaluation of the integration of retrieval-augmented generation in large language model for breast cancer nursing care responses*. Scientific Reports.

Y. Low, Michael L. Jackson, Rebecca J. Hyde, et al. (2025). *Answering real-world clinical questions using large language model, retrieval-augmented generation, and agentic systems*. Digital Health.

Weijie Chen, Ting Bai, Jinbo Su, et al. (2024). *KG-Retriever: Efficient Knowledge Indexing for Retrieval-Augmented Large Language Models*. arXiv.org.

Yutao Zhu, Zhaoheng Huang, Zhicheng Dou, et al. (2024). *One Token Can Help! Learning Scalable and Pluggable Virtual Tokens for Retrieval-Augmented Large Language Models*. AAAI Conference on Artificial Intelligence.

Sourav Verma (2024). *Contextual Compression in Retrieval-Augmented Generation for Large Language Models: A Survey*. arXiv.org.

Chengyuan Yao, and Satoshi Fujita (2024). *Adaptive Control of Retrieval-Augmented Generation for Large Language Models Through Reflective Tags*. Electronics.

Marcus Vinicius Leite, J. Abe, Marcos Leandro Hoffmann Souza, et al. (2025). *Enhancing Environmental Control in Broiler Production: Retrieval-Augmented Generation for Improved Decision-Making with Large Language Models*. AgriEngineering.

Cara Burgan, Josiah Kowalski, and Weidong Liao (2024). *Developing a Retrieval Augmented Generation (RAG) Chatbot App Using Adaptive Large Language Models (LLM) and LangChain Framework*. Proceedings of the West Virginia Academy of Science.

Yun-Wei Chu, Kai Zhang, Christopher Malon, et al. (2025). *Reducing Hallucinations of Medical Multimodal Large Language Models with Visual Retrieval-Augmented Generation*. arXiv.org.

Sefika Efeoglu, and Adrian Paschke (2024). *Relation Extraction with Fine-Tuned Large Language Models in Retrieval Augmented Generation Frameworks*. arXiv.org.

Jeffy Yu (2024). *Retrieval Augmented Generation Integrated Large Language Models in Smart Contract Vulnerability Detection*. arXiv.org.

Kan Feng, Lijun Luo, Yongjun Xia, et al. (2024). *Optimizing Microservice Deployment in Edge Computing with Large Language Models: Integrating Retrieval Augmented Generation and Chain of Thought Techniques*. Symmetry.

Kieran Pichai (2023). *A Retrieval-Augmented Generation Based Large Language Model Benchmarked On a Novel Dataset*. Journal of student-scientists' research.

Reza Fayyazi, Rozhina Taghdimi, and S. Yang (2023). *Advancing TTP Analysis: Harnessing the Power of Large Language Models with Retrieval Augmented Generation*. 2024 Annual Computer Security Applications Conference Workshops (ACSAC Workshops).

Viju Sudhi, Sinchana Ramakanth Bhat, Max Rudat, et al. (2024). *RAG-Ex: A Generic Framework for Explaining Retrieval Augmented Generation*. Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

Shijie Wang, Wenqi Fan, Yue Feng, et al. (2025). *Knowledge Graph Retrieval-Augmented Generation for LLM-based Recommendation*. Annual Meeting of the Association for Computational Linguistics.

Yin Wu, Quanyu Long, Jing Li, et al. (2025). *Visual-RAG: Benchmarking Text-to-Image Retrieval Augmented Generation for Visual Knowledge Intensive Queries*. arXiv.org.

Rui Yang (2024). *CaseGPT: a case reasoning framework based on language models and retrieval-augmented generation*. arXiv.org.

Zhongzhen Huang, Kui Xue, Yongqi Fan, et al. (2024). *Tool Calling: Enhancing Medication Consultation via Retrieval-Augmented Large Language Models*. arXiv.org.

Peizhuo Lv, Mengjie Sun, Hao Wang, et al. (2025). *RAG-WM: An Efficient Black-Box Watermarking Approach for Retrieval-Augmented Generation of Large Language Models*. arXiv.org.

Yang Jiao, Xiaodong Wang, and Kai Yang (2025). *PR-Attack: Coordinated Prompt-RAG Attacks on Retrieval-Augmented Generation in Large Language Models via Bilevel Optimization*. Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

Haijin Wang, Mianrong Zhang, Zheng Chen, et al. (2024). *Carbon Footprint Accounting Driven by Large Language Models and Retrieval-augmented Generation*. arXiv.org.

Hetul Niteshbhai Patel, Azara Surti, Parth Goel, et al. (2024). *A Comparative Analysis of Large Language Models with Retrieval-Augmented Generation based Question Answering System*. 2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC).

Asen Hikov, and Laura Murphy (2024). *Information retrieval from textual data: Harnessing large language models, retrieval augmented generation and prompt engineering*. Journal of AI, Robotics amp; Workplace Automation.

Soroosh Tayebi, A. Aripoli, P. Iglar, et al. (2024). *RadioRAG: Factual Large Language Models for Enhanced Diagnostics in Radiology Using Dynamic Retrieval Augmented Generation*. arXiv.org.

Nguyen Quang Duc, Le Hai Son, Nguyen Duc Nhan, et al. (2024). *Towards Comprehensive Vietnamese Retrieval-Augmented Generation and Large Language Models*. arXiv.org.

Michael DeBellis, Nivedita Dutta, Jacob Gino, et al. (2024). *Integrating Ontologies and Large Language Models to Implement Retrieval Augmented Generation.* Appl. Ontology.

A. Pelletier, Joseph Ramirez, Irsyad Adam, et al. (2024). *Explainable Biomedical Hypothesis Generation via Retrieval Augmented Generation enabled Large Language Models.* arXiv.org.

Xinyi Lin, Gelei Deng, Yuekang Li, et al. (2024). *GeneRAG: Enhancing Large Language Models with Gene-Related Task by Retrieval-Augmented Generation.* bioRxiv.

Dane A Weinert, and A. Rauschecker (2025). *Enhancing Large Language Models with Retrieval-augmented Generation: A Radiology-specific Approach..* Radiology: Artificial Intelligence.

Siru Liu, A. Wright, Allison B. McCoy, et al. (2025). *Detecting emergencies in patient portal messages using large language models and knowledge graph-based retrieval-augmented generation.* J. Am. Medical Informatics Assoc..

Shuliang Liu, Xinze Li, Zhenghao Liu, et al. (2025). *Judge as A Judge: Improving the Evaluation of Retrieval-Augmented Generation through the Judge-Consistency of Large Language Models.* Annual Meeting of the Association for Computational Linguistics.

Quang Nguyen, Duy-Anh Nguyen, Khang Dang, et al. (2024). *Advancing Question-Answering in Ophthalmology with Retrieval Augmented Generations (RAG): Benchmarking Open-source and Proprietary Large Language Models.* medRxiv.

Erik Rehulka, and Marek Suppa (2024). *RAG Meets Detox: Enhancing Text Detoxification Using Open Large Language Models with Retrieval Augmented Generation.* Conference and Labs of the Evaluation Forum.

Jie Huang, Mo Wang, Yunpeng Cui, et al. (2024). *Layered Query Retrieval: An Adaptive Framework for Retrieval-Augmented Generation in Complex Question Answering for Large Language Models.* Applied Sciences.

Zakaria Hammane, Fatima-Ezzahraa Ben-Bouazza, and A. Fennan (2024). *SelfRewardRAG: Enhancing Medical Reasoning with Retrieval-Augmented Generation and Self-Evaluation in Large Language Models*. International Symposium on Computer Vision.

Chamod Samarajeewa, Daswin De Silva, Evgeny Osipov, et al. (2024). *Causal Reasoning in Large Language Models using Causal Graph Retrieval Augmented Generation*. International Conference on Human System Interaction.

Yu Hou, J. R. Bishop, Hongfang Liu, et al. (2024). *Improving Dietary Supplement Information Retrieval: Development of a Retrieval-Augmented Generation System With Large Language Models*. Journal of Medical Internet Research.

Mohammad Affan Habib, Shehryar Amin, Muhammad Oqba, et al. (2024). *TaxTajweez: A Large Language Model-based Chatbot for Income Tax Information In Pakistan Using Retrieval Augmented Generation (RAG)*. The Florida AI Research Society.