

The Role of Retrieval-Augmented Generation in Improving Factual Accuracy for Medical Large Language Models

By Eason Ni

AUTHOR BIOGRAPHY

Eason Ni is a junior in a high school located in Wallingford, Connecticut, with research interests in the intersection of AI and the biomedical field. Aside from research, he enjoys playing the violin and competing in cross country.

ABSTRACT

Large Language Models (LLMs) that rely solely on parametric memory learned through training have demonstrated strong performance in biomedical question-answering, but their tendency to hallucinate facts and the difficulty of adjusting and adding to the learned knowledge limit their usefulness in clinical settings. Retrieval-Augmented Generation (RAG) has emerged as a promising solution by adding a non-parametric source of memory and grounding LLM outputs to reputable external sources. This paper aims to survey the evolution of RAG methodologies and highlight the most recent developments and shifts in paradigms like Agentic RAG. We highlight current state-of-the-art biomedical RAG systems, the latest evaluation benchmarks, and empirical findings regarding best fine-tuning practices. We also examine technical including lost-in-the-middle effects and scaling behaviour. Ethical concerns surrounding privacy and bias are brought to attention alongside research gaps. Finally, we discuss future directions of RAG in the biomedical field, including integration into Clinical Decision Support systems (CDS).

Keywords: *Retrieval-Augmented Generation, Biomedical Question-Answering, Clinical Decision Support, Large Language Models, Language Modeling, Natural Language Processing, Hallucination, Information Retrieval, Biomedical Informatics, Medical AI, AI ethics, Generative AI, Artificial Intelligence*

INTRODUCTION

Pre-trained Large Language Models (LLMs) have demonstrated incredible ability to learn in-depth knowledge from training data without accessing any external memory. In preliminary tests, GPT-4 achieved an accuracy of around 80% on the US Medical Licensing Exam (Step 1, 2, and 3) multiple-choice components (Bubeck et al., 2023). Moreover, developments like Chain-of-Thought (CoT) prompting have attempted to elicit reasoning in LLMs and have improved their ability to handle complex tasks by breaking them down into multiple steps (Wei et al., 2022). CoT prompting has been shown to improve LLMs at biomedical tasks by leveraging reasoning capabilities (Xiong et al., 2024). However, these methods are prone to hallucinations, where LLMs may incorrectly recall or make up facts, make it hard to update or extend the model's information (Singh et al., 2025), and do not provide transparency on generated responses. Accuracy and recency are very important in the biomedical field, where there are constant breakthroughs and new findings. PubMed, an online database for biomedical literature, for example, increases by approximately 5,000 articles every day (Xiong et al., 2024).

Retrieval-Augmented Generation (RAG) solves these issues by creating a pipeline with a retriever that gets relevant information from a corpus and combining the retrieved information with the question before having the model compose an answer (Singh et al., 2025). This pipeline grounds LLM answers in concrete and verifiable sources. Additionally, this allows updates for domain-specific information in the external corpus and the tracking of accessed knowledge to increase transparency. Since its proposal, this methodology has evolved from Naïve RAG to Advanced RAG, then to Modular RAG frameworks (Gao et al., 2023).

While RAG has been implemented in various scientific and clinical Question-Answering (QA) systems, the best practices for tuning these systems and the numerous flexible components of the RAG pipeline, like the corpus, retriever, and answer-generating LLM, remain unclear. RAG systems are also difficult to benchmark, given the complexity and tunability of the parts (Xiong et al., 2024).

ADVANCEMENTS IN RAG PIPELINES

Gao et al. (2023) describe three major paradigms of RAG as Naïve, Advanced, and Modular (Figure 1). More recent work has also given rise to a fourth paradigm, Agentic RAG.

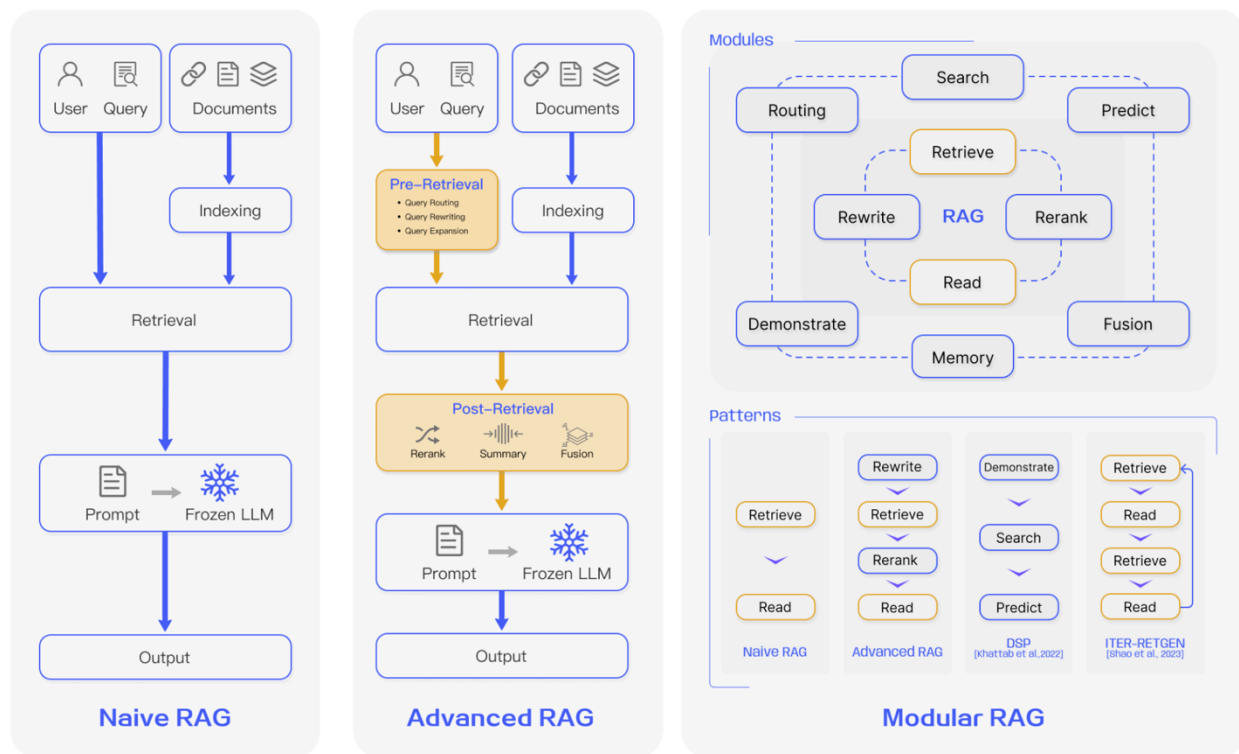


Figure 1: Comparison between three paradigms of RAG. Adapted from Gao et al. (2023).

The simplest RAG pipeline, Naïve RAG, consists of indexing, retrieving, and generation. The indexing step in the Naïve RAG pipeline includes gathering and extracting diverse documents into uniform plain text, then storing the encoded vectors from an embedding model of those chunks in a vector database. Once the user provides a query, the pipeline uses the same embedder to convert the query into a vector. This vector is then used to compute similarity, often with cosine similarity, between the query vector and the vectors of chunks in the vector database. The system then retrieves the top k chunks that demonstrate the most similarity to the query and passes the chunks as context into a combined prompt with the user query. However, Naïve RAG has many flaws. The retrieval step can gather irrelevant chunks or leave out important information for response generation, as the most semantically similar chunk may not always be the chunk that contains the necessary answer. Additionally, for multi-step questions, retrieving information one time may not be adequate.

Advanced RAG adds pre-retrieval and post-retrieval processes, as well as improvements in the indexing step. Methods such as adding metadata to the chunks of the corpus provide more structure and filtered searching capabilities (Gao et al., 2023). Sliding window approaches for indexing corpora create overlap between

chunks to preserve semantic completeness. Dynamic chunking decides where to set boundaries of chunks based on semantic similarity changes, also aiming for the same objective (Liu et al., 2025). Moreover, knowledge graph indexing can be used to capture complicated relationships with nodes and connections between concepts and entities (Gao et al., 2023). The pre-retrieval process aims to transform the user's query into something more suitable for the retrieval process. It accomplishes this task through strategies like query rewriting, transformation, and expansion, among other techniques (Gao et al., 2023). Post-retrieval involves integrating the retrieved context with the query optimally. Methods commonly used are re-ranking the chunks of context and summarizing less important parts of the context to keep the total context size for the final generation step small (Gao et al., 2023).

By separating the RAG pipeline components into separate flexible modules, the Modular RAG architecture makes the process more adaptable (Gao et al., 2023). Different modules can be swapped, rearranged, or fine-tuned. New modules like the search module allow more adaptive data retrieval by enabling the use of a diverse set of search engines, databases, and knowledge graphs. A fusion module expands the user query to multiple similar queries and retrieves for each query in parallel to get more information. Modular RAG has also brought new patterns, adding diversity to the existing retrieve-read pattern that Naïve and Advanced RAG use. Examples of other patterns are Retrieve-Read-Retrieve-Read. These workflows allow for iterative retrieval based on missing information (Gao et al., 2023).

The most recent frontier of RAG research, Agentic RAG, brings the agentic paradigm to RAG. Even though some agentic behavior has already been experimented with in RAG systems, Agentic RAG represents the new paradigm shift. In Agentic RAG systems, machine learning algorithms and LLMs have control over deciding the workflow. Design patterns include reflection, planning, prompt chaining, tool use, and multi-agent collaboration. Example frameworks include Corrective RAG, which utilizes self-correction mechanisms to score the overall quality of the retrieved documents and uses the score to trigger varying actions (Yan et al., 2024). Adaptive-RAG uses a small language model first to classify the complexity of the user's query and then tailor the processing strategy (Jeong et al., 2024). Despite its many strengths, Agentic RAG faces challenges with computational overhead and scalability, given the complexity of the interactions between agents and the complexity of the user's question (Singh et al., 2025).

CASE STUDIES OF RAG APPLIED TO MEDICAL APPLICATIONS

The RAG pipeline has been quickly implemented in biomedicine for its ability to provide more accuracy and transparency. To excel at medical domain-specific questions and tasks, many RAG pipelines have been developed with different retrievers and corpora and include different strategies (Xiong et al., 2024).

MedRAG is a novel RAG framework designed to enhance the diagnosis accuracy of LLMs introduced by Zhao et al. (2025). MedRAG uses Knowledge Graph (KG)-elicited reasoning with four tiers to capture the comprehensive differences critical for diagnostics. This pipeline can also retrieve data from Electronic Health Records (EHRs). By combining retrieved information from these data sources, MedRAG can provide precise and personalized diagnoses and proactively ask follow-up questions to clarify ambiguous patient information. In quantitative evaluations, MedRAG outperformed state-of-the-art RAG models on the Chronic Pain Diagnostic Dataset (CPDD) and the DDXPlus dataset, exceeding next-best scores by 11.32% and 1.23% respectively, in the task of identifying the exact disease name (Zhao et al., 2025). Additionally, the proactive questioning mechanism was shown to increase this accuracy from 52.83% to 66.04% (Zhao et al., 2025).

BiomedRAG is another novel RAG framework designed to enhance LLMs in biomedical Natural Language Processing (NLP) tasks by reducing hallucinations (Li et al., 2025). BiomedRAG does this by utilizing a learnable, tailored chunk scorer that identifies which chunks from the corpora improve the LLM's perplexity, signaling that the retrieved information is what the LLM needs to answer the question. BiomedRAG was benchmarked across four biomedical NLP tasks: triple extraction; relation extraction; text classification; and link prediction against other State-of-the-Art (SOTA) models, where the BiomedRAG framework produced superior results across the board. BiomedRAG improved the original MedLLaMA 13B and LLaMA2 13B LLMs by 34.67% and 22.43%, respectively, in F1 score for triple extraction on the GIT dataset (Li et al., 2025).

ATTEMPTS AT BENCHMARKING MEDICAL RAG SYSTEMS

The RAG system is comprised of several flexible and interdependent components. Given the diversity of RAG architectures, consistent benchmarking is essential for fair comparisons and reproducibility. To address this need, the Medical Information Retrieval-Augmented Generation Evaluation (MIRAGE) benchmark was introduced by Xiong et al. (2024) to compare medical question-answering RAG systems. It is comprised of 7,663 questions from 5 datasets, including three medical examination QA datasets (MMLU-Med, MedQA-US, MedMCQA) and two biomedical research QA datasets (PubMedQA, BioASQ-Y/N). A key feature of MIRAGE not found in benchmarks before is the ability to have Question-Only Retrieval (QOR), which does not provide answer options during the retrieval phase, making it reflect real-world scenarios more realistically. To facilitate comprehensive comparisons on MIRAGE, Xiong et al. (2024) developed the MEDRAG toolkit. MEDRAG offers systematic implementations of RAG pipelines for medical QA, incorporating a wide range of corpora, retrievers, and LLMs. However, MIRAGE currently only supports the vanilla RAG architecture, lacking support for active or agentic RAG systems.

TECHNICAL CHALLENGES AND LIMITATIONS

Despite the massive improvements RAG has provided for medical domain-specific and in-depth questions, there are still limits and ways to improve.

Table 1: Average accuracy (%) of GPT-3.5 across retrievers for different corpora on MIRAGE. Retrievers include BM25, Contriever, SPECTER, MedCPT, RRF-2, and RRF-4. Shows the impact of corpus choice on overall performance. Data derived from (Xiong et al., 2024).

Benchmark Dataset	Corpus					
	None (CoT)	PubMed	StatPearls	Textbooks	Wikipedia	MedCorp
MMLU-Med	72.91	73.20	72.99	74.83	73.17	74.58
MedQA-US	65.04	64.77	65.59	66.55	64.62	66.12
MedMCQA	55.25	54.66	54.10	54.44	54.21	56.21
PubMedQA	36.00	64.87	27.97	28.83	27.67	64.33
BioASQ-Y/N	74.27	85.17	59.68	59.52	71.63	85.03
Average	60.69	68.53	56.07	56.83	58.26	69.25

As the well-known saying goes, “garbage in, garbage out.” This certainly applies to the RAG paradigm. By design, RAG systems rely on the corpus or corpora to act as the ground truth in response generation (He et al., 2025). As shown in Table 1, the average values of accuracy (%) vary widely across different corpora for various evaluation datasets. Therefore, the output quality can only achieve the quality of the input context, meaning that the curation of corpora is a cardinal part of the fine-tuning. In the biomedical field, this curation is particularly challenging due to the rapid advancements and high volume of literature being produced. Consequently, real-world implementations of RAG systems need robust strategies for curating and updating corpora. There also needs to be a robust solution to contradicting information in corpora. For RAG systems, retrieving noisy or contradictory information has been observed to be detrimental to the final output. As Gao et al. (2023) noted, for LLMs, “Misinformation can be worse than no information at all.”

How LLMs handle context is another limitation. In an ideal scenario, LLMs would perform the same regardless of how much context they are given, so retrievers can feed the LLM as much context as it needs and have it utilize the necessary snippets. However, LLMs suffer decreased performance in accuracy for tasks when given larger contexts. It is a confirmed limitation that LLMs do not process context uniformly, and their performance grows increasingly unreliable as the input length expands, even on simple tasks (Du et al., 2025). Du et al. (2025) demonstrated that in retrieval tasks like “Needle in a Haystack,” where a simple question has its answer surrounded by irrelevant information, models that achieve near-perfect scores on short contexts begin to

show significant performance degradation as token counts grow, while remaining within their claimed context windows, with reported performance drops ranging from 13.5% to 85% across evaluated models.

The “lost-in-the-middle” phenomenon, when LLMs tend to have trouble utilizing information in the middle of their context (Liu et al., 2024), also limits the performance of RAG systems (Xiong et al., 2024). This phenomenon is confirmed by examining model accuracy relative to the position of the crucial ground-truth snippet within the context. Xiong et al. (2024) revealed that when evaluating RAG systems on biomedical datasets like PubMedQA and BioASQ-Y/N, there is a clear U-shaped (i.e. decreasing-then-increasing) pattern in the accuracy corresponding to context locations of crucial information. This pattern proves that performance is lowest when the relevant information is placed in the middle of the prompt, indicating that LLMs tend to concentrate their attention on the beginning and end of the input sequence. Even though some systems attempt to mitigate this by re-ranking context chunks to have the most important chunks at the beginning and end of the prompt, this phenomenon effectively reduces the amount of the working context size.

Additionally, Xiong et al. (2024) identified a log-linear scaling relationship between model performance and the number of retrieved snippets in RAG systems. Generally, as the number of retrieved context chunks (k) increases, the performance of the system improves because the likelihood of including helpful knowledge grows. However, this scaling has an upper limit; if k becomes too large, the signal-to-noise ratio begins to decrease, and the presence of excessive irrelevant information can hinder the LLM’s ability to derive the correct answer because of the reasons mentioned above.

Moreover, there is the issue of latency. As RAG systems become increasingly complex in the goal of attaining higher accuracy, latency will increase. This will also make these systems more computationally costly. As these systems evolve from the simple “retrieve-read” workflow to more complex paradigms like Agentic RAG, the use of multiple autonomous agents will significantly increase the computational overhead (Singh et al., 2025).

ETHICAL CONCERNS

While RAG has proven to be a promising solution to enhance LLMs in biomedicine, medicine, and healthcare by integrating real-time non-parametric memory retrieval to provide contextually relevant and up-to-date responses, its implementation raises several ethical concerns (Bunnell et al., 2025). These concerns are similar to or stem from existing issues with LLMs that remain largely unresolved. RAG tries to address the main problem of transparency and hallucinations, but it does not eliminate all fundamental LLM challenges and even introduces new ones. (Bunnell et al., 2025) summarizes that most ethical concerns about RAG systems fall under these categories:

Privacy and data security. LLM conversations are often stored, and this practice is the same with RAG-enhanced LLM conversations. This raises issues regarding potential privacy breaches and data usage (Tegsten, 2024). Including sensitive patient data as context for LLMs increases cybersecurity risks (Quidwai & Lagana, 2024; Tegsten, 2024).

Bias and cultural sensitivity. Recent studies have highlighted the risk of algorithmic biases and the importance of ensuring that diverse cultural medical practices, such as Traditional Chinese Medicine, are represented fairly (Su & Gu, 2024). Biases may stem from limited training data and existing disparities (AlGhadban et al., 2023; Long et al., 2024).

Transparency and explainability. Even augmented by RAG pipelines, LLM generations are still a black box, giving them some degree of unpredictability (Bunnell et al., 2025). Without clear mechanisms for interpreting the pipeline's behavior, it becomes difficult to evaluate reliability or assign responsibility when errors occur. Studies have consistently emphasized that adequate transparency is essential for safe integration of AI into clinical settings.

RESEARCH GAPS

There is a dearth of research on the algorithmic biases of retrievers used in RAG pipelines. Uneven distribution of demographic attributes in knowledge sources can lead to intrinsic biases in generated responses (He et al., 2025). Even with attempts to establish evaluation standards, there is a persistent lack of specialized benchmarks and datasets tailored to capture the unique aspects of agentic capabilities, such as multi-agent collaboration and dynamic adaptability. Multilingual support remains another significant gap, as English dominates public biomedical repositories. As a result, the performance of RAG systems in low-resource languages has not been adequately explored due to a scarcity of available knowledge sources (He et al., 2025). A deeper understanding of the Retriever-Generator Interaction is also required, as these components currently remain loosely connected, often allowing a model's inherent knowledge boundaries or noise in the retrieved text to lead to incorrect outputs. There are contradictory findings that have not been investigated thoroughly. For example, Cuconasu et al. (2024) found that adding random documents in the prompt improves LLM accuracy by up to 35%. This directly contradicts the initial assumption that irrelevant context reduces quality, and suggests that the underlying assumptions that fueled RAG development may be faulty.

FUTURE DIRECTIONS AND APPLICATIONS

One of the most promising directions for RAG in medicine is integration into Clinical Decision Support (CDS) systems. Currently deployed real-world CDS systems use LLMs in limited early-stage ways at most and utilize static knowledge bases (Zhao et al., 2025). A RAG-enhanced CDS system could draw on regularly maintained corpora containing medical information and new research findings, while also accessing a patient's electronic health record in parallel, enabling grounded outputs that reflect both up-to-date knowledge and patient-specific details.

Many studies have suggested the feasibility of this approach in narrow clinical domains. For example, MedRAG demonstrated that hierarchical knowledge graph-guided retrieval can improve diagnostic differentiation in complex cases. While these systems remain largely experimental and domain-specific, they provide proof-of-concept evidence that agentic RAG architectures can support clinically meaningful reasoning tasks beyond single-turn question answering.

Multimodal RAG systems that integrate imaging (e.g., X-rays), time series (e.g., ECG), and audio (e.g., lung sound recordings) may further expand the scope of AI-assisted clinical decision-making. Because the focus has been on text-based RAG systems and only a small subset of works has studied image-based retrieval, many other modalities have been inadequately explored, providing an opportunity for future improvements (He et al., 2025).

CONCLUSION

The introduction of the RAG pipeline has significantly enhanced LLM accuracy in domain-specific areas like healthcare and biomedicine. As surveyed in this paper, the evolution from Naïve and Advanced RAG pipelines to Modular and, more recently, Agentic RAG, reflects a shift toward increasingly adaptive and autonomous systems capable of addressing the unique demands of medical question-answering.

Despite these advances, the deployment of RAG in medical settings is constrained by technical and ethical challenges. The accuracy of these systems is highly sensitive to corpus quality and retrieval strategies, and limitations in LLM context management and phenomena such as lost-in-the-middle continue to restrict performance. Moreover, the growing complexity of agentic workflows introduces additional concerns related to computational cost and scalability. To address the lack of understanding between retriever and generator interactions, benchmarking efforts, such as MIRAGE, represent important progress toward the standardized evaluation of individual components of RAG. However, these current benchmarks fall behind the rapidly evolving architectures of modern RAG systems, particularly those with autonomous or multi-agent behaviors.

Ethical concerns further complicate real-world adoption, as while RAG improves traceability of information, it does not eliminate risks related to data privacy and bias.

REFERENCES

- Alghamdi, H. M., & Mostafa, A. (2024). Towards reliable healthcare LLM agents: A case study for pilgrims during hajj. *Information*, 15(7), 371. <https://doi.org/10.3390/info15070371>
- Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Sharick, E. (2023). Paper Review: 'Sparks of Artificial General Intelligence: Early experiments with GPT-4'.
- Bunnell, D. J., Bondy, M. J., Fromtling, L. M., Ludeman, E., & Gourab, K. (2025, April 1). *Bridging AI and healthcare: A scoping review of retrieval-augmented generation—ethics, bias, transparency, improvements, and applications*. Cold Spring Harbor Laboratory. <https://doi.org/10.1101/2025.04.01.25325033>
- Cuconasu, F., Trappolini, G., Siciliano, F., Filice, S., Campagnano, C., Maarek, Y., ... & Silvestri, F. (2024, July). The power of noise: Redefining retrieval for rag systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 719-729).
- Du, Y., Tian, M., Ronanki, S., Rongali, S., Bodapati, S., Galstyan, A., ... & Peng, H. (2025). Context length alone hurts LLM performance despite perfect retrieval. *arXiv preprint arXiv:2510.05381*.
- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1).
- He, J., Zhang, B., Rouhizadeh, H., Chen, Y., Yang, R., Lu, J., ... & Teodoro, D. (2025). Retrieval-Augmented Generation in Biomedicine: A Survey of Technologies, Datasets, and Clinical Applications. *arXiv preprint arXiv:2505.01146*.
- Jeong, S., Baek, J., Cho, S., Hwang, S. J., & Park, J. C. (2024). Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. *arXiv preprint arXiv:2403.14403*.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33, 9459-9474.
- Li, M., Kilicoglu, H., Xu, H., & Zhang, R. (2025). Biomedrag: A retrieval augmented large language model for biomedicine. *Journal of Biomedical Informatics*, 162, 104769.
- Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12, 157-173.

-
- Liu, S., McCoy, A. B., & Wright, A. (2025). Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. *Journal of the American Medical Informatics Association : JAMIA*, 32(4), 605–615. <https://doi.org/10.1093/jamia/ocaf008>
- Long, C., Subburam, D., Lowe, K., dos Santos, A., Zhang, J., Hwang, S., Saduka, N., Horev, Y., Su, T., Côté, D. W., & Wright, E. D. (2024). ChatENT: Augmented large language model for expert knowledge retrieval in otolaryngology–head and neck surgery. *Otolaryngology–Head and Neck Surgery*, 171(4), 1042-1051. <https://doi.org/10.1002/ohn.864>
- Lozano, A., Fleming, S. L., Chiang, C. C., & Shah, N. (2024). Clinfo.ai: An Open-Source Retrieval-Augmented Large Language Model System for Answering Medical Questions using Scientific Literature. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 29, 8–23.
- Quidwai, M. A., & Lagana, A. (2024, March 18). *A RAG chatbot for precision medicine of multiple myeloma*. OpenRxiv. <https://doi.org/10.1101/2024.03.14.24304293>
- Singh, A., Ehtesham, A., Kumar, S., & Khoei, T. T. (2025). Agentic retrieval-augmented generation: A survey on agentic rag. *arXiv preprint arXiv:2501.09136*.
- Su, X., & Gu, Y. (2024, June 11). *Implementing retrieval-augmented generation (RAG) for large language models to build confidence in traditional chinese medicine*. Center for Open Science. <https://doi.org/10.31219/osf.io/ns2v3>
- Tegsten, S. (2024). Tailored Query Resolution for Medical Data Interaction: Integrating LangChain4j, LLMs, and Retrieval Augmented Generation: Utilizing Real Time Embedding Techniques.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35, 24824-24837.
- Xiong, G., Jin, Q., Lu, Z., & Zhang, A. (2024, August). Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics ACL 2024* (pp. 6233-6251).
- Yan, S., Gu, J., Zhu, Y., & Ling, Z. (2024). Corrective Retrieval Augmented Generation. *ArXiv, abs/2401.15884*.
- Zhan, Z., Wang, J., Zhou, S., Deng, J., & Zhang, R. (2025). Mmrag: Multi-mode retrieval-augmented generation with large language models for biomedical in-context learning. *arXiv preprint arXiv:2502.15954*.
- Zhao, X., Liu, S., Yang, S. Y., & Miao, C. (2025, April). Medrag: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. In *Proceedings of the ACM on Web Conference 2025* (pp. 4442-4457).