Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X

RETRIEVAL-AUGMENTED GENERATION: ARCHITECTURES, ADAPTIVE RETRIEVAL, FEEDBACK-DRIVEN OPTIMIZATION, AND OPEN RESEARCH CHALLENGES

Muhammad Ali Hassan[1], Muhammad Azam*[2], Mehwish Amin[3], Afsheen[4], Ammad Hussain*[5]

[1]0009-0006-9322-2068
[2]0009-0006-8045-0598
[3]0009-0002-4090-439X
[4]0009-0002-2703-1985
[5]0009-0004-0245-2462

[1]alihassanhashmi786@gmail.com, [2] muhammadazam.lashari@gmail.com
[3]mehwishamin2000@gmail.com, [4]afsheenanees41@gmail.com, [5]ammadhussain709@gmail.com

DOI:

Keywords:
Retrieval-Augmented Generation, Large Language Models, Dense Retrieval, Adaptive Retrieval, Feedback Learning, Knowledge-Grounded Generation

Article History
Received on 10 Jan, 2026
Accepted on 02 Feb, 2026
Published on 05 Feb, 2026

Copyright @Author

Corresponding Author*:
Muhammad Azam
Ammad Hussain

Abstract

RAG has become a core paradigm for grounding LLMs into external knowledge sources, preventing hallucinations, and allowing scalable reasoning over dynamic corpora. By integrating parametric language modeling with non-parametric retrieval mechanisms, RAG systems close the gap between fluent natural language generation and factuality. Unlike fully parametric models, RAG provides access to recent and domain specific information at the time of inference. Nevertheless, recent empirical results suggest that a majority of already-deployed RAG pipelines are still brittle because they only rely on static similarity-based retrieval techniques and simple naïve pipeline strategies for context construction and weak or even relevance-only reranking with no feedback-driven adaptivity. The following survey will be based on a very brief overview of current RAG research, including: basic architectural designs; retriever and re-ranker strategies; context construction methodologies; adaptive and reinforcement-learning-based RAGs; feedback aware models; graph based extensions; memory based extensions; long context behaviour; hallucination analysis; and new evaluation benchmarks. We base our work on over forty representative studies to introduce a single taxonomy, provide a metadata-based comparative analysis, and provide comprehensive information on open research challenges. Our argument is that the RAG systems of the future should no longer be in the paradigm of static retrieval, but rather integrate context adaptation mechanisms to feedback to improve the resilience, efficiency and effectiveness of deployment.

Introduction

Large language models (LLMs) such as GPT-3 have proven to be quite successful in a wide range of natural language processing tasks, such as question-answering, summarization, reasoning, and dialogue-generation. However, LLMs also have inherent limitations as they are dependent on parametric memory gained in the course of training, which results in inheriting a number of limitations. These limitations are reflected in the form of hallucinated products, obsolescence or false factual knowledge, reduced transparency, and limited domain adaptability. In turn, these limitations have crippling effects on the implementation of LLMs in the mission-critical environments, including healthcare, finance, or the legal sector.

RAG was presented as a principled response to these issues, by augmenting LLMs with the ability to consult external knowledge resources at inference time. The well-known RAG model first introduced by (Lewis et al., 2021) showed that retrieval-augmented models are a big step above closed-book LLMs when evaluating them on factual QA tasks by conditioning generation on retrieved evidence. Dense Passage Retrieval (DPR) (Karpukhin et al., n.d.), rapidly emerged as the de-facto retrieval backbone by exploiting dual encoders combined with approximate nearest-neighbor search thanks to FAISS (Johnson et al., 2017) and HNSW (Malkov & Yashunin, n.d.).

More recent systems like RETRO (Borgeaud et al., 2022) have taken to include retrieval as part of the Transformer architecture, underscoring the value of an external memory. Nevertheless, despite these improvements, recent works and benchmarks (Gupta & Ranjan, n.d.) suggest that naïve RAG pipelines tend to perform badly in realistic settings, such as multi-hop reasoning tasks on long-context inputs or even domain specific tasks. These findings warrant further investigation on adaptive retrieval, feedback and intelligent context construction.

RAG Architecture:

A classic RAG system generally contains three main components: a retriever, context constructor and generator. The retriever is tasked with returning promising document chunks from a large external corpus (commonly through dense or hybrid similarity search). The context builder is used to join the retrieved fragments into purposeful prompts, which are then used by the generator to generate the final product. Early-generation systems

Early-generation RAG systems are based upon dense similarity matching, and assume that semantic similarity between query and document fragment implies usefulness of a fragment. Nonetheless, this is not a universal assumption, in fact, the recent research indicates that even semantically similar fragments can still be misleading, redundant and in general fail to answer the complex questions. (Cuconasu et al., 2024). It follows that the canonical RAG design, while elegant in its concept, directly reveals substantial weaknesses with respect to chunk relevance, redundancy and prioritization.

Metadata Table:

ID	Paper	Category	Key Contribution	Limitation
[1]	(Lewis et al., 2021)	Foundational RAG	Original RAG architecture	Static retrieval
[2]	(Karpukhin et al., n.d.)	Retrieval	DPR dense retriever	Similarity-only
[3]	(Johnson et al., 2017)	ANN Search	FAISS indexing	No ranking adaptivity
[4]	(Malkov & Yashunin, n.d.)	ANN Search	HNSW	Approximate only
[5]	(Borgeaud et al., 2022)	Memory RAG	Retrieval in transformer	Fixed memory
[6]	(Meduri et al.,	Efficiency	Scalable RAG	Retrieval still static

	2024)			
[7]	(Jeong et al., 2024)	Adaptive RAG	Query complexity routing	Classifier errors
[8]	(Wang et al., 2024)	Memory RAG	Iterative notes	High latency
[9]	(Gupta & Ranjan, n.d.)	Survey	RAG gaps	No new method
[10]	(Leto et al., 2024)	Optimal Rag	Toward optimal search and retrieval in rag	Descriptive
[11]	(Cuconasu et al., 2024)	Retrieval Noise	Noise improves RAG	Heuristic
[12]	(Jin et al., 2024)	Long-context RAG	Hard negatives	QA-only
[13]	(Leng et al., 2024)	Long-context Study	Context degradation	Simple RAG
[14]	(Park et al., 2025)	T2DM	Evaluating Phenotyping using optimized RAG	Small dataset
[15]	(Zhang et al., 2025)	OpenGenAlign	Long-Context generation: Reward Modeling in Open-Ended Long-Context Generation.	Only Auto reward model
[16]	(Tang & Yang, 2024)	Multi-hop RAG	Multi-hop benchmark	No method
[17]	(W. Liu et al., n.d.)	Rag-Instruct	Boosting LLMs with diverse RAG instruction	Eval only
[18]	(Finardi et al., 2024)	Retrieval Study	BM25 + rerank	Single dataset
[19]	(Glass et al., 2022)	Reranking	Retrieve–rerank–generate	Expensive
[20]	(Dong et al., 2024)	Graph rerank	AMR-based graph	Preprocessing cost
[21]	(Sanmartin, 2024)	KG-based RAG	Low hallucination	Slow
[22]	(Edge et al., 2025)	Global RAG	Sensemaking RAG base approach	LLM-as-judge
[23]	(Gutiérrez et al., 2025)	Memory RAG	Continual learning	Heavy
[24]	(Oche et al., 2025)	RAG progress	Gaps , progress and future directions regarding RAG	No feedback scoring, metadata enrichment discussed
[25]	(Yang et al., n.d.)	benchmark	CRAG benchmark	RL expensive
[26]	(Shi et al., 2024)	Modular RAG	Trigger + cache	Complex
[27]	(Islam et al., 2024)	Open-source RAG	MoE + reflection	Heavy training

[28]	(J. Liu et al., 2024)	Reasoning RAG	Noise filtering	Extra encoder
[29]	(Nian et al., 2025)	W-RAG	Weakly supervised as dense retrieval used in RAG	LLM cost
[30]	(Barker et al., 2025)	Multi-objective RAG	Cost–latency–safety	Small data
[31]	(Şakar & Emekci, 2025)	Hallucination	Domain analysis	No fix
[32]	(Kulkarni et al., 2024)	Token efficiency	RL saving	Small scope
[33]	(Zamani & Bendersky, 2024)	Stochastic RAG	Gumbel-top-k chunks	Expensive
[34]	(Li et al., 2025)	End-to-end RAG	DPO alignment	Rollout cost
[35]	(Leng et al., 2024)	Evaluation	Context limits	No adaptivity
[36]	(Yuanji Lyu, n.d.)	CRUD-RAG	Comprehensive Benchmark for RAG	Evaluation only
[37]	(Koo et al., 2024)	Retrieval tuning	Query opt	Small gains
[38]	(Ray et al., 2025)	METIS	Fast Quality Aware Rag	Latency
[39]	(Gao et al., 2025)	RL RAG	End-to-end	Heavy
[40]	(B & Purwar, 2024)	Evaluation	Long-context prefs	Not RAG

How RAG Works: Retrieval and Context Construction

For a given query, the retriever identifies the top- document chunks $\{c_1, ..., c_k\}$ based on embedding similarity (Karpukhin et al., n.d.). The retrieval is also optimally efficient and scalable hence making it easy to access the knowledge base quickly. However, it is not very sensitive to downstream reasoning demands and is more likely to deal with retrieval on its own as opposed to connecting it to the following generative steps, as part of a unified joint generative pipeline.

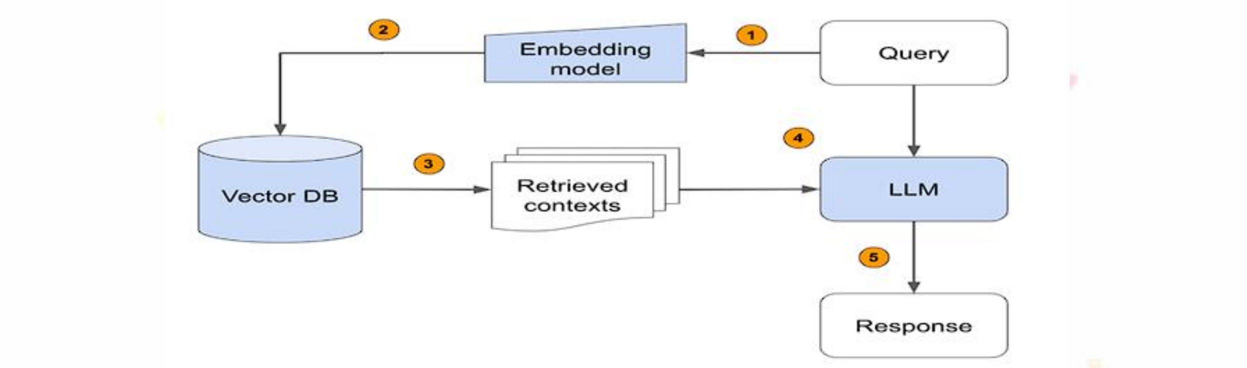

Figure 1:Rag model overview process

Document fragments are then glued with user query and a request command to the language model on retrieval, thus constituting the input. Three other mainstream frameworks, including LangChain and LlamaIndex, also use the same simple methodology. However, it uniformly processes all the retrieved segments regardless of their relative importance, repetition or historical usefulness(Oche et al., 2025). The empirical results point to performance decrease with increasing the number of segments that are retrieved, and it is mainly associated with adding noise, having hard negatives, and the resulting dilution of attention in the context window of the language model (Jin et al., 2024) . We can show that the resilience of a Retrieval -Augmented Generation system should not only depend on the quality of the generated retrievals, but the wise constructions and choices of context.

Taxonomy of Retrieval-Augmented Generation Systems

The Retrieval- Augmented Generation (RAG) systems have gone through spurt in progress, in accordance with the attempts of the researchers to overcome the limitations of the fixed retrieval processes and basic methods of context construction. With the help of studying the coordination of retrieval, ranking, and generation elements, the current RAG strategies may be systematically categorized into six major segments. The resulting taxonomy demonstrates an observable trend, with the move to shift the frameworks that are merely static and based on similarities into dynamical, adaptive, and feedback-recognizing frameworks.

Static Retrieval-Based RAG

The first and most popular type of RAG systems is static retrieval based retrieval augmented generation. In this paradigm, retrieval is done through mechanisms that are based on invariant similarity, the most common being dense vector search and the retrieved documents are not refined any more, but are directly given to the generator. The basic (RAG) model formalised by (Lewis et al., 2021) follows this kind of appraoch wherein a dense retriever is used to select top- ranking passages which are concatenated with the query and fed into a sequence-to-sequence generator. Though the factual question-answering tasks may be performed better using static RAG models than closed-book language models, the latter works on the assumption that the fact that the semantically related text is present in the retrieved passage gives it a certain degree of utility of the generated text. Later empirical studies, which are elaborated in the following sections in this paper, have shown that this assumption does not commonly hold especially on difficult, multi-hop, and long-context queries. Additionally, fixed RAG systems cannot handle change of domains, changing corpus, or changes in historical usage, and hence it is vulnerable to failure in practice.

Hybrid and Reranking-Based RAG

Retrieval-Augmented Generation (RAG) systems of the hybrid and reranking types aim to improve the retrieval performance by incorporating several retrieval cues and inserting intermediate ranking steps. The hybrid retrieval models usually combine the sparse and key-word based approaches like BM25 with the dense and embedding based approaches and thus enhance recall and effectiveness in wide query formulations. This paradigm is represented in the Re2G framework (Glass et al., 2022) hat explicitly decouples retrieval and reranking with generation, which allows a neural reranker chance to tune candidate documents up to the generative step.

Graph-based reranking methods, such as G-RAG (Dong et al., 2024) , further improve this framework by taking into account inter-document relations instead of just scoring documents independently. These methods improve the performance of downstream by eliminating redundancy and ranking more informative evidence highly. However, with improved quality of the overall ranking despite that most reranking based RAG systems are still mainly concerned with relevance; they do not actively optimize to achieve success in generation and thus they only partly support retrieval decisions that are in line with the requirements of language models.

As a result, although there are improvements in rankings of fidelity, reranking-based RAG models still have a relevance-based focus, which provides limited direct maximisation of generation-success metrics, thus limiting the complete correspondence of retrieval results and operational needs of downstream generative frameworks.

Adaptive Retrieval-Augmented Generation

Adaptive RAG systems extend beyond predetermined retrieval strategies and make on-the-fly decisions with regard to when, how, and how much to retrieve responsive passage units to a particular input query. A lightweight classifier is utilized to estimate query complexity in Adaptive-RAG (Jeong et al., 2024) and conditions the query retrieval behavior, scaling down unnecessary retrievals for simple queries while retaining accuracy for hard ones.

More extensions are modular RAG and portable trigger-based RAG frameworks, which are able to do retrieval caching, retrieval skipping, and if-then-else route (Shi et al., 2024) . These methods show us that retrieval is not just an unconditional process. The approaches demonstrate that retrieval is not a process that is not conditional. Nonetheless, the adaptive RAG systems are prone to such errors as misclassification and the overpass might introduce additional elements of architecture that will complicate system design and tuning..

Feedback-Aware RAG

The Retrieval Augmentation Generation (RAG) systems based on feedback differ with the traditional paradigms that only attempt to optimize similarity metrics or rely on a fixed re-ranker. Rather they acquire the salient signals gradually through attuning to real world user feedback, thus achieving an acquisition of preference signals and in some cases, simulated supervisory signals. This adaptive process eventually guides both the retrieval and ranking processes, which would allow the system to rank information in accordance with the user intentions.

W-RAG (Nian et al., 2025) takes this a step further by minimizing the use of manual annotations, replacing them with cues obtained through large language models and, therefore, reducing hallucinations and enhancing the factualization of the results. However, these feedback-sensitive systems present extra costs, including slower inference through computation cost, more intricate feedback acquisition, and more latency in the system, that a practitioner must juggle.

Graph- and Memory-Based RAG

RAG systems based on graphs and memory go beyond the traditional flat vector retrieval paradigm and introduce hierarchy and memory. To give an example, the entity-based graph, GraphRAG (Edge et al., 2025) is built, and information is then aggregated in a layer degree. This helps to provide a successful interpretation of large textual corpora, as well as clarify the connections between the unrelated documents. Graph-based RAG models also have the advantage of being able to reason on scattered traces more efficiently by explicitly mapping inter-entity and document linkages.

Memory-augmented models like HippoRAG-2 (Gutiérrez et al., 2025) push this further. They use association-based recollection and lifelong learning through a stable non-parametric memory hence mastering multi-hop reasoning problems. These advantages however are at the expense of higher preprocessing requests, higher memory footprints and slower inference times. The overhead that is produced can form an important bottleneck in time-sensitive applications.

Reinforcement-Learning-Based RAG

Retrieval-augmented generation(RAG) systems referring to reinforcement learning model retrieval and generation as a decision-making process have been suggested. These methodologies do not independently adjust the parameters of each of the components; instead, they follow a single policy, which simultaneously controls retrieval decisions, reranking and answer generation. The latter is exemplified by the work of Smart-RAG (Gao et al., 2025) , that learns a joint policy based on reinforcement learning signals induced from successful answers and retrieval cost.

RAG systems, based on RL, have performed better and are principled; however, they are computationally expensive to train, and they need careful reward design and exploration. The above challenges limit their direct application into industry, requiring simplified formulations, but the frameworks provide significant insights on end-to-end RAG optimisation.

Long-Context RAG and Hallucination Analysis

Paradoxically, there is no guarantee that an increase in the size of context window will be associated with an increase in RAGs performance. Massive inspections show that accuracy of most systems increases dramatically as number of reasonable words used and becomes decreasing as more chunks are added (Leng et al., 2024) . False negative (-) also worsen hallucination since they distort the generator (Jin et al., 2024).

Certain domain-specific studies, such as in the financial domains reveal that mis-aligned retrievals can make RAG fall behind closed-book systems (Park et al., 2025). These results highlight the value of context prioritization and structured reasoning.

Open Research Challenges and Future Directions

Nevertheless, certain significant gaps remain in question. First, the degree of chunk utility continues to be approximated primarily in similarity/relevance, as opposed to what contribution to accurate generation it makes. Second, because we are computing ally expensively integrating the feedback and never in practice use real human feedback at scale. Third, There are various incompleteness's about the current context construction procedure: Thirdly, the context construction is performed quite informally through gluing naively the sentences together. Lastly, deployment is in practice impractical due to scalability and speed issues. Finally, evaluation benchmarks are also not adequate in capturing human-conceptualized concept of factuality and reliability. The next generation RAG systems will be faced with the need to address these problem.

Conclusion

A popular method to ground large language models in knowledge and reduce the occurrence of hallucination and knowledge staleness has been Retrieval-Augmented Generation (RAG). Initially founded on basic dense retrieval, RAG systems have been extended to hybrid retrieval, reranking, adaptive control, feedback-based optimization, reinforcement learning, and graph and memory reasoning. This development is an indicator of a change in perception towards retrieval as a dynamic and learnable component of generation and not a detached preprocessing phase. This review concludes that the present limitations of RAG are based not as much on the lack of retrieval but on the selection and utilization of retrieved information. Naive similarity-based retrieval and plain context concatenation tend to add noise, particularly in the long-context and multi-hop tasks, i.e. larger contexts do not imply higher accuracy of the facts. More recent adaptive and feedback-sensitive techniques enhance robustness by gaining knowledge of what information is valuable to generate, but at the cost of higher system complexity and cost. In general, it is possible to say that the future of RAG lies in the adaptive context optimization dynamically determining what, when, and how to access and present knowledge with the assistance of lightweight feedback mechanisms and more realistic evaluation benchmarks.

References

1. B, G., & Purwar, A. (2024). *Evaluating the Efficacy of Open-Source LLMs in Enterprise-Specific RAG Systems: A Comparative Study of Performance and Scalability.* http://arxiv.org/abs/2406.11424
2. Barker, M., Bell, A., Thomas, E., Carr, J., Andrews, T., & Bhatt, U. (2025). *Faster, Cheaper, Better: Multi-Objective Hyperparameter Optimization for LLM and RAG Systems.* http://arxiv.org/abs/2502.18635
3. Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., Driessche, G. van den, Lespiau, J.-B., Damoc, B., Clark, A., Casas, D. de Las, Guy, A., Menick, J., Ring, R., Hennigan, T., Huang, S., Maggiore, L., Jones, C., Cassirer, A., … Sifre, L. (2022). *Improving language models by retrieving from trillions of tokens.* http://arxiv.org/abs/2112.04426
4. Cuconasu, F., Trappolini, G., Siciliano, F., Filice, S., Campagnano, C., Maarek, Y., Tonellotto, N., & Silvestri, F. (2024). The Power of Noise: Redefining Retrieval for RAG Systems. *SIGIR 2024 - Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 719–729. https://doi.org/10.1145/3626772.3657834
5. Dong, J., Fatemi, B., Perozzi, B., Yang, L. F., & Tsitsulin, A. (2024). *Don't Forget to Connect! Improving RAG with Graph-based Reranking.* http://arxiv.org/abs/2405.18414
6. Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., Metropolitansky, D., Ness, R. O., & Larson, J.

(2025). *From Local to Global: A Graph RAG Approach to Query-Focused Summarization*. http://arxiv.org/abs/2404.16130

7. Finardi, P., Avila, L., Castaldoni, R., Gengo, P., Larcher, C., Piau, M., Costa, P., & Caridá, V. (2024). *The Chronicles of RAG: The Retriever, the Chunk and the Generator*. http://arxiv.org/abs/2401.07883
8. Gao, J., Li, L., Li, W., Fu, Y., & Dai, B. (2025). *SmartRAG: Jointly Learn RAG-Related Tasks From the Environment Feedback*. http://arxiv.org/abs/2410.18141
9. Glass, M., Rossiello, G., Chowdhury, M. F. M., Naik, A. R., Cai, P., & Gliozzo, A. (2022). *Re2G: Retrieve, Rerank, Generate*. http://arxiv.org/abs/2207.06300
10. Gupta, S., & Ranjan, R. (n.d.). *A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions*.
11. Gutiérrez, B. J., Shu, Y., Qi, W., Zhou, S., & Su, Y. (2025). *From RAG to Memory: Non-Parametric Continual Learning for Large Language Models*. http://arxiv.org/abs/2502.14802
12. Islam, S. Bin, Rahman, M. A., Hossain, K. S. M. T., Hoque, E., Joty, S., & Parvez, M. R. (2024). *Open-RAG: Enhanced Retrieval-Augmented Reasoning with Open-Source Large Language Models*. http://arxiv.org/abs/2410.01782
13. 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*. http://arxiv.org/abs/2403.14403
14. Jin, B., Yoon, J., Han, J., & Arik, S. O. (2024). *Long-Context LLMs Meet RAG: Overcoming Challenges for Long Inputs in RAG*. http://arxiv.org/abs/2410.05983
15. Johnson, J., Douze, M., & Jégou, H. (2017). *Billion-scale similarity search with GPUs*. http://arxiv.org/abs/1702.08734
16. Karpukhin, V., O˘ Guz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., Yih, W.-T., & Ai, F. (n.d.). *Dense Passage Retrieval for Open-Domain Question Answering*. https://github.com/facebookresearch/DPR.
17. Koo, H., Kim, M., & Hwang, S. J. (2024). *Optimizing Query Generation for Enhanced Document Retrieval in RAG*. http://arxiv.org/abs/2407.12325
18. Kulkarni, M., Tangarajan, P., Kim, K., & Trivedi, A. (2024). *Reinforcement Learning for Optimizing RAG for Domain Chatbots*. http://arxiv.org/abs/2401.06800
19. Leng, Q., Portes, J., Havens, S., Zaharia, M., & Carbin, M. (2024). *Long Context RAG Performance of Large Language Models*. http://arxiv.org/abs/2411.03538
20. Leto, A., Aguerrebere, C., Bhati, I., Willke, T., Tepper, M., & Vo, V. A. (2024). *Toward Optimal Search and Retrieval for RAG*. http://arxiv.org/abs/2411.07396
21. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2021). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. http://arxiv.org/abs/2005.11401
22. Li, X., Mei, S., Liu, Z., Yan, Y., Wang, S., Yu, S., Zeng, Z., Chen, H., Yu, G., Liu, Z., Sun, M., & Xiong, C. (2025). *RAG-DDR: Optimizing Retrieval-Augmented Generation Using Differentiable Data Rewards*. http://arxiv.org/abs/2410.13509
23. Liu, J., Lin, J., & Liu, Y. (2024). *How Much Can RAG Help the Reasoning of LLM?* http://arxiv.org/abs/2410.02338
24. Liu, W., Chen, J., Ji, K., Zhou, L., Chen, W., & Wang, B. (n.d.). *RAG-Instruct: Boosting LLMs with Diverse Retrieval-Augmented Instructions*. https://github.com/FreedomIntelligence/RAG-
25. Malkov, Y. A., & Yashunin, D. A. (n.d.). *Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs*.
26. Meduri, K., Nadella, G. S., Gonaygunta, H., Maturi, M. H., Fatima, F., & Member, I. (2024). *Efficient RAG Framework for Large-Scale Knowledge Bases* (Vol. 9, Issue 4). www.ijnrd.org
27. Nian, J., Peng, Z., Wang, Q., & Fang, Y. (2025). W-RAG: Weakly Supervised Dense Retrieval in RAG for Open-domain Question

Answering. *ICTIR 2025 - Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval*, 136–146. https://doi.org/10.1145/3731120.3744578

28. Oche, A. J., Folashade, A. G., Ghosal, T., & Biswas, A. (2025). *A Systematic Review of Key Retrieval-Augmented Generation (RAG) Systems: Progress, Gaps, and Future Directions*. http://arxiv.org/abs/2507.18910
29. Park, H., Rees, M., Kruger, N., Fuse, K., Castro, V. M., Gainer, V., Wattanasin, N., Benoit, B., Wagholikar, K. B., & Murphy, S. N. (2025). *Evaluation of T2DM Phenotyping Using Optimized Retrieval-Augmented Generation (RAG) and the Impact of Embedding Model, Context, and Prompt*. https://doi.org/10.1101/2025.04.29.25326696
30. Ray, S., Pan, R., Gu, Z., Du, K., Feng, S., Ananthanarayanan, G., Netravali, R., & Jiang, J. (2025). *METIS: Fast Quality-Aware RAG Systems with Configuration Adaptation*. http://arxiv.org/abs/2412.10543
31. Şakar, T., & Emekci, H. (2025). Maximizing RAG efficiency: A comparative analysis of RAG methods. *Natural Language Processing*, *31*(1), 1–25. https://doi.org/10.1017/nlp.2024.53
32. Sanmartin, D. (2024). *KG-RAG: Bridging the Gap Between Knowledge and Creativity*. http://arxiv.org/abs/2405.12035
33. Shi, Y., Zi, X., Shi, Z., Zhang, H., Wu, Q., & Xu, M. (2024). *Enhancing Retrieval and Managing Retrieval: A Four-Module Synergy for Improved Quality and Efficiency in RAG Systems*. http://arxiv.org/abs/2407.10670
34. Tang, Y., & Yang, Y. (2024). *MultiHop-RAG: Benchmarking Retrieval-Augmented Generation for Multi-Hop Queries*. http://arxiv.org/abs/2401.15391
35. Wang, X., Sen, P., Li, R., & Yilmaz, E. (2024). *Adaptive Retrieval-Augmented Generation for Conversational Systems*. http://arxiv.org/abs/2407.21712
36. Yang, X., Sun, K., Xin, H., Sun, Y., Bhalla, N., Chen, X., Choudhary, S., Daniel Gui, R., Will Jiang, Z., Jiang, Z., Kong, L., Moran, B., Wang, J., Ethan Xu, Y., Yan, A., Yang, C., Yuan, E., Zha, H., Tang, N., … Reality Labs, M. (n.d.). *CRAG-Comprehensive RAG Benchmark*. https://www.aicrowd.com/challenges/meta-comprehensive-rag-benchmark-kdd-cup-2024
37. *yuanji (s40)*. (n.d.).
38. Zamani, H., & Bendersky, M. (2024). Stochastic RAG: End-to-End Retrieval-Augmented Generation through Expected Utility Maximization. *SIGIR 2024 - Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2641–2646. https://doi.org/10.1145/3626772.3657923
39. Zhang, H., Song, J., Zhu, J., Wu, Y., Zhang, T., & Niu, C. (2025). *OpenGenAlign: A Preference Dataset and Benchmark for Trustworthy Reward Modeling in Open-Ended, Long-Context Generation*. http://arxiv.org/abs/2501.13264