

Retrieval-Augmented Generation for Improving Factual Reliability in AI-Based Question Answering Systems

Noah Anderson

Associate Professor, Institute of Data Science (Fictional), USA

Abstract

Artificial intelligence-based question answering systems can generate fluent responses across broad knowledge domains, but linguistic fluency does not necessarily guarantee factual reliability. Generative language models may produce unsupported statements, rely on outdated parametric knowledge, or respond confidently when relevant evidence is unavailable. Retrieval-Augmented Generation provides a methodological approach for addressing this limitation by retrieving external documents and conditioning answer generation on selected evidence. The effectiveness of this approach, however, depends on retrieval quality, evidence relevance, and the ability of the generator to use retrieved material faithfully. This hypothetical methodological study evaluates Retrieval-Augmented Generation for improving factual reliability in knowledge-intensive question answering. The proposed framework compares three conceptual configurations: a parametric-only generative question-answering model, a sparse-retrieval RAG system, and a dense-retrieval RAG system. A fictional corpus of 20,000 knowledge passages supports a hypothetical set of 1,000 question-answer pairs divided into training, validation, and testing partitions.

The illustrative analysis evaluates exact-match accuracy, token-level F1 score, evidence-supported answer rate, and unsupported-claim rate. Under the assigned hypothetical values, the dense-retrieval RAG configuration achieves the strongest overall question-answering and factual-support performance. A second analysis examines performance across factoid, multi-evidence, temporally sensitive, and ambiguous questions. The assigned results show that factual reliability varies according to question structure and retrieval difficulty rather than depending solely on the underlying generator. The study concludes that retrieval augmentation provides a useful framework for grounding generated answers in external evidence, but retrieval does not automatically eliminate hallucination or factual error. Incorrect, incomplete, outdated, or weakly matched passages can themselves become sources of unreliable generation. Future empirical research should therefore evaluate retrieval recall, evidence attribution, answer factuality, source quality, and uncertainty jointly rather than relying exclusively on conventional answer-accuracy metrics.

Keywords: Retrieval-Augmented Generation, Question Answering, Factual Reliability, Knowledge Retrieval, Dense Passage Retrieval, Large Language Models, Hallucination Reduction, Evidence Grounding, Natural Language Processing, Artificial Intelligence.

1. Introduction

Artificial intelligence-based question answering systems have progressed from extractive architectures toward generative language models capable of producing complete natural-language responses. These systems can synthesize information, reformulate evidence, and answer questions without copying exact spans from a source document. However, a generated answer may appear coherent even when particular statements are unsupported by reliable evidence. This distinction between linguistic plausibility and factual correctness has become an important methodological concern in knowledge-intensive natural language processing.

Language models can store substantial factual information within learned parameters. Petroni et al. examined language models as repositories of relational knowledge, while Roberts et al. demonstrated that large generative models can answer some open-domain questions using information encoded within their parameters. Parametric knowledge is useful because it allows rapid generation without explicit retrieval. Nevertheless, updating specific facts within model parameters can be difficult, and the model may not reveal the documentary source supporting a particular answer.

Open-domain question answering historically addressed this problem through information retrieval followed by reading comprehension. Chen et al. developed a system combining document retrieval and machine reading, while Karpukhin et al. demonstrated that dense neural representations could retrieve relevant passages effectively for open-domain question answering. Retrieval provides external context that can be inspected separately from the model's internal representation. The remaining challenge is integrating retrieved information with flexible natural-language generation while maintaining correspondence between evidence and the generated answer.

Lewis et al. introduced Retrieval-Augmented Generation as a framework combining parametric sequence-to-sequence generation with non-parametric external memory. Guu et al. independently developed REALM, in which retrieval was incorporated into language-model pretraining. Izacard and Grave later demonstrated that generative models could aggregate evidence from multiple retrieved passages effectively. These developments established retrieval augmentation as an important research direction for knowledge-intensive language tasks, including question answering, fact verification, and evidence-grounded generation.

The present hypothetical study examines whether retrieval augmentation can improve factual reliability in AI-based question answering under a transparent comparative framework. Parametric-only generation, sparse-retrieval RAG, and dense-retrieval RAG are compared through answer accuracy and evidence-support measures. The study also examines question-type differences and retrieval-related failure modes.

2. Objectives of the Study

The primary objective of this study is to examine the methodological potential of Retrieval-Augmented Generation for improving factual reliability in artificial intelligence-based question answering. Particular attention is given to retrieval quality, answer accuracy, evidence support, unsupported generation, and variation across different forms of knowledge-intensive questions.

2.1 Examination of Retrieval-Augmented Question Answering

The first objective is to examine how external retrieval can be incorporated into generative question answering. The proposed framework converts a user question into a retrieval query, identifies potentially

relevant passages from an indexed knowledge corpus, and provides selected passages to the answer generator. The generator is therefore able to condition its response on externally accessible information rather than relying exclusively on parametric memory. This architecture creates a methodological basis for connecting answer generation with identifiable supporting evidence.

The objective also examines alternative retrieval mechanisms. Sparse retrieval identifies documents using lexical correspondence, whereas dense retrieval represents questions and passages as learned semantic vectors. These approaches can behave differently when a question and its relevant evidence use different terminology. The study does not assume that dense retrieval always performs better. Instead, it establishes a comparative design in which retrieval quality and downstream factual reliability can be examined jointly under identical questions and knowledge sources.

2.2 Evaluation of Factual Reliability

The second objective is to evaluate generated answers through both conventional question-answering metrics and evidence-oriented reliability indicators. Exact match measures whether a generated answer corresponds exactly to a reference answer after normalization, while token-level F1 allows partial lexical agreement. These measures are useful but do not independently establish whether every statement within a generated response is supported by retrieved evidence. A separate factual-support indicator is therefore incorporated into the hypothetical framework.

The study additionally examines unsupported-claim frequency. An answer may contain a correct primary response while introducing an additional unsupported explanation. Evaluating only exact match can overlook this behavior. The hypothetical factual-support procedure therefore assesses whether substantive claims within the answer can be linked to the supplied evidence passages. This separation between answer correctness and evidential support provides a more informative basis for examining factual reliability in retrieval-augmented systems.

2.3 Identification of Retrieval and Generation Challenges

The third objective is to identify methodological problems that can limit the reliability of RAG-based question answering. Relevant issues include retrieval failure, semantically similar but incorrect passages, conflicting documents, outdated knowledge, insufficient context, ambiguous questions, and generator disregard of retrieved evidence. These failure modes demonstrate that retrieval augmentation adds new components whose errors can propagate through the complete question-answering pipeline.

The study further aims to clarify the importance of source provenance and uncertainty. A system should not present a generated statement as firmly established merely because a passage was retrieved. Retrieval scores measure relevance rather than absolute truthfulness, and external corpora may themselves contain incorrect information. Reliable RAG evaluation should therefore examine document quality, retrieval relevance, evidence-answer correspondence, and appropriate abstention when the available evidence is insufficient.

3. Review of Literature

3.1 Open-Domain Question Answering and Retrieval

Chen et al. developed DrQA as an open-domain question-answering system combining document retrieval with a neural reader. Their work demonstrated that question answering could be performed over large document collections rather than using a single provided passage. Kwiatkowski et al. later

introduced Natural Questions, a large-scale dataset based on real information-seeking queries, while Joshi et al. developed TriviaQA using question-answer pairs with supporting evidence documents. These benchmarks strengthened evaluation of retrieval-dependent question answering.

Karpukhin et al. introduced Dense Passage Retrieval for open-domain question answering using learned dual-encoder representations. Their work demonstrated that semantic vector retrieval could outperform a strong sparse-retrieval baseline on several passage-retrieval benchmarks. Dense retrieval is particularly relevant when lexical overlap between a question and its supporting passage is limited. Nevertheless, sparse methods remain useful because they are computationally efficient, interpretable, and effective for queries containing distinctive terminology.

3.2 Retrieval-Augmented Language Generation

Guu et al. developed REALM, which augments language-model pretraining with a latent retriever accessing external textual knowledge. This approach demonstrated that information retrieval could be incorporated directly into neural language modeling rather than treated as a completely separate preprocessing stage. Lewis et al. subsequently introduced Retrieval-Augmented Generation, combining dense document retrieval with sequence-to-sequence generation and explicit non-parametric memory for knowledge-intensive natural-language processing tasks.

Izacard and Grave developed a passage-fusion approach in which a generative model processes multiple retrieved passages before producing an answer. Their results highlighted the value of aggregating evidence from several retrieved documents. Petroni et al. introduced KILT as a benchmark covering question answering, fact checking, slot filling, entity linking, and dialogue while retaining provenance information. These studies collectively emphasize that retrieval quality and generation quality should be evaluated together rather than as independent components.

3.3 Factual Reliability and Knowledge Grounding

Thorne et al. introduced FEVER as a large-scale dataset for fact extraction and verification, emphasizing the connection between factual judgments and explicit supporting evidence. Kryściński et al. later developed FactCC for evaluating factual consistency in abstractive generation. Although FactCC focused on summarization, its underlying motivation is relevant to RAG: linguistic overlap or fluency alone does not establish that generated content is consistent with the source material.

Lin et al. introduced TruthfulQA to evaluate whether language models generate truthful answers rather than reproducing common misconceptions. Their findings demonstrated that scale alone does not guarantee factual reliability. Retrieval-based systems provide one potential mechanism for improving grounding because answers can be conditioned on external evidence. However, retrieved evidence must itself be relevant and used faithfully. Consequently, factuality evaluation should consider both the retriever and the generator.

4. Research Methodology

4.1 Research Design and RAG Architecture

The present study adopts a hypothetical comparative methodological design involving three question-answering configurations. The first is a parametric-only generative model receiving only the question. The second combines a sequence-to-sequence generator with sparse lexical retrieval. The third combines the same conceptual generator with dense passage retrieval. Keeping the generator conceptually constant

across the two RAG configurations allows differences to be associated more clearly with the retrieval mechanism rather than with unrelated changes in generation architecture.

Each RAG system conceptually retrieves the top five passages from an indexed knowledge collection before generating an answer. Retrieved passages are ranked according to relevance and provided alongside the question. The generator is instructed conceptually to answer only when sufficient supporting information is present and otherwise indicate insufficient evidence. No actual model, index, retrieval engine, or prompting procedure was implemented for the numerical results reported in this manuscript.

4.2 Hypothetical Dataset and Knowledge Corpus

The fictional question-answering dataset contains 1,000 knowledge-intensive questions. Seven hundred questions are allocated to training, one hundred fifty to validation, and one hundred fifty to independent testing. The supporting knowledge corpus contains 20,000 hypothetical passages representing encyclopedic, technical, scientific, historical, and institutional information. Every passage and question-answer pair is fictional and does not correspond to an existing benchmark, website, organization, or proprietary knowledge base.

The test set is conceptually divided into forty-five factoid questions, forty multi-evidence questions, thirty-five temporally sensitive questions, and thirty ambiguous or underspecified questions. Reference answers and evidence passages are assumed to have been created consistently for methodological illustration. A genuine experiment would require transparent annotation guidelines, evidence verification, version-controlled source documents, and inter-annotator assessment for factual-support judgments.

4.3 Evaluation Measures and Reliability Assessment

Question-answering performance is evaluated through normalized exact match and token-level F1 score. Exact match requires the predicted core answer to correspond completely with the reference answer after basic normalization. Token F1 measures overlap between predicted and reference answer tokens and therefore gives partial credit when wording differs. These metrics are supplemented by an evidence-supported answer rate representing the proportion of generated answers whose substantive factual claims are supported by the provided evidence.

Unsupported-claim rate represents the proportion of evaluated answers containing at least one substantive factual statement not justified by the retrieved passages. This measure is not assumed to be a standardized universal factuality metric; it is a hypothetical human-evaluation indicator constructed for the present methodology. An empirical study would require multiple evaluators, explicit support criteria, agreement measurement, and separate analysis of retrieval errors and generation errors.

5. Analysis

5.1 Comparative Question-Answering and Factual Reliability

The first hypothetical analysis compares parametric-only generation, sparse-retrieval RAG, and dense-retrieval RAG across the 150-question test set. The parametric model is assigned an exact-match score of 61.3% and token F1 of 70.8%. Sparse retrieval improves these illustrative values to 67.3% and 76.2%, while dense retrieval produces 72.0% exact match and 80.4% token F1. These values are deliberately constructed and do not represent experimental model outputs.

The larger hypothetical distinction occurs in factual-support measures. The parametric system receives a 76% evidence-equivalent factual-support rating based on reference verification, whereas sparse RAG reaches 86% and dense RAG 91%. Unsupported-claim rates decline correspondingly. The pattern illustrates the methodological hypothesis that access to relevant evidence may improve factual grounding, while also showing that retrieval does not eliminate unsupported output completely. Incorrect retrieval or unfaithful generation can still produce errors.

Table 1: Hypothetical Comparison of Question-Answering Configurations

Model Configuration	Exact Match (%)	Token F1 (%)	Evidence-Supported Answers (%)	Unsupported-Claim Rate (%)
Parametric-Only Generator	61.3	70.8	76.0	18.0
Sparse-Retrieval RAG	67.3	76.2	86.0	10.0
Dense-Retrieval RAG	72.0	80.4	91.0	6.0

5.2 Performance Across Question Categories

The second hypothetical analysis examines the dense-retrieval RAG configuration across four question categories. Factoid questions produce the strongest illustrative exact-match and factual-support values because a single passage is assumed to contain the required answer clearly. Multi-evidence questions record slightly weaker results because relevant information must be combined across passages. Temporally sensitive questions are more difficult when source freshness becomes important, while ambiguous questions produce the lowest assigned performance.

The category-level comparison illustrates that factual reliability cannot be summarized fully through one overall score. Retrieval quality depends on question specificity, corpus coverage, evidence distribution, and document freshness.

Table 2: Hypothetical Dense-RAG Performance by Question Type

Question Type	Number of Test Questions	Exact Match (%)	Token F1 (%)	Evidence-Supported Answers (%)
Factoid	45	80.0	86.2	95.0
Multi-Evidence	40	72.5	81.0	92.0
Temporally Sensitive	35	68.6	78.1	88.0
Ambiguous / Underspecified	30	60.0	71.4	83.0

6. Discussion

6.1 Retrieval Augmentation and Factual Grounding

The hypothetical results illustrate the central rationale of retrieval-augmented question answering: externally retrieved evidence can provide information that does not need to be stored entirely within model parameters. Lewis et al. showed that combining parametric generation with non-parametric memory can improve knowledge-intensive task performance, while REALM similarly demonstrated the value of retrieval-augmented language modeling. The present framework extends this reasoning toward explicit factual-support evaluation.

However, retrieval grounding should not be equated automatically with factual correctness. A generator can ignore retrieved evidence, combine incompatible passages, or add unsupported details. The hypothetical unsupported-claim rate remains above zero even for dense RAG to emphasize this limitation. Reliable system design should therefore evaluate whether generated statements are actually entailed or supported by retrieved sources rather than merely confirming that retrieval occurred somewhere within the processing pipeline.

6.2 Dense Retrieval and Evidence Selection

Dense retrieval is designed to identify semantic similarity between questions and passages even when exact terminology differs. Karpukhin et al. demonstrated its effectiveness for open-domain question answering, while later retrieval-generation systems used dense retrievers as important components. The hypothetical results assign stronger performance to dense RAG than sparse RAG because relevant evidence is assumed to require semantic matching in a portion of the test questions.

Sparse retrieval nevertheless remains an important benchmark. Lexical matching can be highly effective for names, dates, identifiers, and specialized terminology, and sparse systems may offer greater transparency concerning why a passage was selected. Hybrid retrieval combining sparse and dense signals may therefore be valuable in real systems. Comparative evaluation should examine retrieval recall directly rather than inferring retriever quality solely from final answer accuracy.

6.3 Factuality Metrics and Answer Evaluation

Traditional question-answering metrics such as exact match and token F1 evaluate correspondence with reference answers but cannot fully determine factual reliability. A generated answer may express the correct core answer and still append an unsupported explanation. Conversely, a factually correct paraphrase may receive a lower lexical score. FEVER, FactCC, and TruthfulQA demonstrate broader efforts to evaluate evidence, consistency, and truthfulness beyond surface-form similarity.

The present methodology therefore separates answer accuracy from evidence support. This distinction is particularly important for RAG because the availability of source passages enables attribution-oriented evaluation.

7. Challenges

7.1 Retrieval Failure and Irrelevant Evidence

The first major challenge is retrieval failure. If the relevant evidence is absent from the top-ranked passages, the generator may be unable to answer reliably even when the knowledge corpus contains the required information. Dense retrievers can also return semantically similar passages that discuss related

but different entities or events. Such retrieval errors can be especially misleading because the returned text may appear relevant enough to encourage confident but incorrect generation.

Evaluating retrieval recall is therefore necessary independently of final answer quality. Researchers should determine whether the required evidence appears within the retrieved set before attributing an incorrect answer entirely to the generator. Retrieval depth also introduces a trade-off: increasing the number of passages can improve evidence coverage but can introduce distracting or contradictory context. Optimal retrieval therefore depends on both recall and evidence precision.

7.2 Source Quality, Freshness, and Contradictory Information

RAG systems inherit the quality characteristics of their external knowledge sources. A retrieved document may be outdated, incomplete, poorly sourced, or factually incorrect. Temporally sensitive questions are particularly challenging because an answer can become obsolete even when it was correct when the knowledge index was created. Updating the external corpus can be easier than retraining a complete language model, but indexing freshness still requires active maintenance and version control.

Conflicting sources create an additional problem. Multiple retrieved passages may provide different values, dates, interpretations, or conclusions. A generator should not silently select one source without acknowledging uncertainty when the evidence is genuinely inconsistent. Source provenance, publication date, authority, and internal agreement can therefore become important retrieval-ranking or answer-generation signals. Factual reliability requires evaluation of source quality in addition to semantic relevance.

7.3 Evidence Utilization and Hallucinated Generation

A generator can produce unsupported content even when correct evidence is present in its context. Neural language generation is probabilistic, and the model may rely partly on parametric associations rather than strictly following retrieved passages. This creates an evidence-utilization problem distinct from retrieval failure. Systems should therefore examine whether final claims can be attributed directly to the supplied context and whether unsupported elaboration is appropriately discouraged.

Abstention represents an important reliability mechanism when evidence is insufficient. A system optimized exclusively for answer completion may generate a plausible response even when retrieval fails. In contrast, a reliability-oriented question-answering system should be capable of indicating that available evidence does not justify a confident answer. Designing and evaluating such behavior requires calibrated confidence measures, explicit insufficiency criteria, and test questions for which abstention is genuinely appropriate.

8. Conclusion

Retrieval-Augmented Generation provides a methodologically important framework for connecting generative question answering with external evidence. By combining a parametric language model with a searchable knowledge source, RAG can make relevant information available during answer generation and provide a pathway toward more transparent evidence grounding.

The present hypothetical study compared parametric-only generation with sparse- and dense-retrieval configurations using both answer-accuracy and factual-support measures. The illustrative results assigned the strongest performance to dense-retrieval RAG, including higher exact-match accuracy, greater evidence-supported answer rates, and fewer unsupported claims. These values are

entirely fictional and do not establish empirical superiority. Their purpose is to demonstrate how a factual-reliability study can distinguish retrieval quality, answer correctness, and evidence correspondence instead of treating fluent generation as sufficient evidence of accuracy.

The study concludes that RAG can reduce dependence on static parametric knowledge but does not eliminate factual errors automatically. Retrieval failure, poor source quality, conflicting evidence, outdated documents, and generator hallucination remain important risks. Future empirical research should evaluate retrieval recall, source provenance, claim-level support, uncertainty, abstention behavior, and knowledge-base freshness alongside conventional question-answering metrics.

References

1. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474. <https://arxiv.org/abs/2005.11401>
2. Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M.-W. (2020). REALM: Retrieval-augmented language model pre-training. *Proceedings of the 37th International Conference on Machine Learning*, 119, 3929–3938. <https://proceedings.mlr.press/v119/guu20a.html>
3. Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-T. (2020). Dense passage retrieval for open-domain question answering. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 6769–6781. <https://doi.org/10.18653/v1/2020.emnlp-main.550>
4. Izacard, G., & Grave, E. (2021). Leveraging passage retrieval with generative models for open domain question answering. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, 874–880. <https://doi.org/10.18653/v1/2021.eacl-main.74>
5. Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., van den Driessche, G., Lespiau, J.-B., Damoc, B., Clark, A., de Las Casas, D., Guy, A., Menick, J., Ring, R., Hennigan, T., Huang, S., Maggiore, L., Jones, C., Cassirer, A., ... Simonyan, K. (2022). Improving language models by retrieving from trillions of tokens. *Proceedings of the 39th International Conference on Machine Learning*, 162, 2206–2240. <https://proceedings.mlr.press/v162/borgeaud22a.html>
6. Izacard, G., Lewis, P., Lomeli, M., Hosseini, L., Petroni, F., Schick, T., Dwivedi-Yu, J., Joulin, A., Riedel, S., & Grave, E. (2022). Few-shot learning with retrieval augmented language models. *arXiv*. <https://arxiv.org/abs/2208.03299>
7. Asai, A., Min, S., Zhong, Z., & Chen, D. (2022). Retrieval-based language models and applications. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. <https://aclanthology.org/>
8. Ram, O., Levine, Y., Dalmedigos, I., Muhlga, D., Shashua, A., Leyton-Brown, K., & Shoham, Y. (2023). In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11, 1316–1331. https://doi.org/10.1162/tacl_a_00605
9. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024). Retrieval-augmented generation for large language models: A survey. *arXiv*. <https://doi.org/10.48550/arXiv.2312.10997>

10. Jiang, Z., Xu, F. F., Gao, L., Sun, Z., Liu, Q., Dwivedi-Yu, J., Yang, Y., Callan, J., & Neubig, G. (2023). Active retrieval augmented generation. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. <https://arxiv.org/abs/2305.06983>
11. Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. International Conference on Learning Representations. <https://arxiv.org/abs/2310.11511>
12. 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. Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics. <https://arxiv.org/abs/2403.14403>
13. Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2024). RAGAS: Automated evaluation of retrieval augmented generation. Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, 150–158. <https://aclanthology.org/2024.eacl-demo.16/>
14. Li, S., Park, S., Lee, I., & Bastani, O. (2024). TRAQ: Trustworthy retrieval augmented question answering via conformal prediction. Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics, 3799–3821. <https://doi.org/10.18653/v1/2024.naacl-long.210>
15. Hwang, J., Park, J., Park, H., Kim, D., Park, S., & Ok, J. (2024). Retrieval-augmented generation with estimation of source reliability. arXiv. <https://arxiv.org/abs/2410.22954>
16. Zhou, Y., Liu, Y., Li, X., Jin, J., Qian, H., Liu, Z., Li, C., Dou, Z., Ho, T.-Y., & Yu, P. S. (2024). Trustworthiness in retrieval-augmented generation systems: A survey. arXiv. <https://arxiv.org/abs/2409.10102>
17. Yang, X., Sun, K., Xin, H., Sun, Y., Bhalla, N., Chen, X., Choudhary, S., Gui, R. D., Jiang, Z. W., Jiang, Z., Kong, L., Moran, B., Wang, J., Xu, Y. E., Yan, A., Yang, C., Yuan, E., Zha, H., Tang, N., ... Dong, X. L. (2024). CRAG: Comprehensive RAG benchmark. Advances in Neural Information Processing Systems, 37. https://proceedings.neurips.cc/paper_files/paper/2024/hash/1435d2d0fca85a84d83ddcb754f58c29-Abstract-Datasets_and_Benchmarks_Track.html
18. Zhang, T., Patil, S. G., Jain, N., Sheng, V., Zaharia, M., Gonzalez, J. E., & Stoica, I. (2024). RAFT: Adapting language model to domain specific RAG. arXiv. <https://arxiv.org/abs/2403.10131>
19. Chen, J., Lin, H., Han, X., & Sun, L. (2024). Benchmarking large language models in retrieval-augmented generation. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. <https://aclanthology.org/>
20. Yu, H., et al. (2024). FoRAG: Factuality-optimized retrieval augmented generation for web-enhanced long-form question answering. Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. <https://doi.org/10.1145/3637528.3672065>