International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

Enhancing Generative AI Answer Accuracy Using Retrieval Augmented Generation: An Experimental Study

Author Name: **Bimba A S,** II Sem MCA, Global Institute of Management Sciences, bimba.sreedhara@gmail.com
Author Name: **Nisha G Gaud ,** II Sem MCA, Global Institute of Management Sciences, nishagaud66@gmail.com
Guide Prof. **Sumathi G K,** Associate Professor, Dept of Computer Applications, Global Institute of Management Sciences, Bangalore, sumathi.gk@gimsedu.in

Abstract - Generative Artificial Intelligence (GenAI) systems are increasingly used for question answering and knowledge-based applications, but their responses may contain inaccurate or unsupported information. This study experimentally evaluates whether Retrieval-Augmented Generation (RAG) improves the quality of GenAI-generated answers compared with direct generation without external document support. Twenty fixed questions based on the NIST Generative Artificial Intelligence Profile and UNESCO's Guidance for Generative AI in Education and Research were answered under two conditions using the same ChatGPT system: direct generation and retrieval-supported generation. In the RAG condition, the top five document chunks retrieved using semantic similarity were supplied as context for answer generation. A manually applied scoring rubric assessed accuracy, relevance, faithfulness to source content, and hallucination-free quality, with each criterion scored on a scale of 0–2. The baseline achieved an overall score of 89.38%, while the RAG condition achieved 98.75%, an improvement of 9.38%. Faithfulness increased from 57.5% to 100%, whereas accuracy decreased from 100% to 95%. Relevance and hallucination-free quality remained at 100% under the study's evaluation rubric in both conditions. These findings indicate that retrieval-supported generation can improve source faithfulness, although incomplete or insufficient retrieved context may affect accuracy. The results highlight the importance of evaluating multiple quality dimensions when developing reliable, document-grounded GenAI systems.

Keywords: Generative Artificial Intelligence, Retrieval-Augmented Generation, Answer Accuracy, Faithfulness, Hallucination

1. INTRODUCTION

Generative Artificial Intelligence (GenAI) has become increasingly useful for answering questions, summarizing information, and supporting knowledge-based tasks. Large language models can generate fluent

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

responses to natural-language questions, making information more accessible across fields such as education and research. However, the fluency of generated text does not necessarily indicate factual correctness. Responses may contain inaccurate statements or information that cannot be supported by reliable sources, creating challenges for applications that require trustworthy answers.

Retrieval-Augmented Generation (RAG) offers an approach to addressing this limitation by providing a language model with relevant information retrieved from external documents before answer generation. Rather than relying solely on knowledge encoded during training, the model can use the supplied context to formulate responses grounded in identifiable source material. Nevertheless, retrieval does not guarantee correctness. Relevant information may be missed, retrieved passages may be incomplete, and the model may misinterpret or fail to use the available evidence effectively.

These challenges make it important to evaluate RAG beyond whether its answers appear relevant or well written. Answer accuracy, relevance, faithfulness to source content, and the presence of unsupported information represent distinct dimensions of response quality. An answer may address a question appropriately while failing to reflect the supplied evidence, or it may remain faithful to limited context without fully answering the question.

This study experimentally compares direct ChatGPT answer generation with a retrieval-supported generation procedure using 20 fixed questions based on two authoritative documents: the NIST Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile and UNESCO's Guidance for Generative AI in Education and Research. Both conditions use the same question wording and answer-generation system. In the RAG condition, the top five document chunks retrieved through semantic similarity are supplied to the model along with an instruction to answer using only the retrieved context. Responses are evaluated using a common scoring rubric covering accuracy, relevance, faithfulness, and hallucination-free quality.

The study examines whether retrieval-supported generation improves overall answer quality and whether improvements are consistent across the individual evaluation criteria. By comparing the two conditions, it also identifies potential limitations of relying on retrieved context when generating source-grounded answers.

2. RELATED WORK AND RESEARCH GAP

2.1 Generative AI Reliability and Hallucination

Large language models (LLMs) can answer natural-language questions without task-specific retraining, demonstrating their flexibility across knowledge-intensive applications (Brown et al., 2020). However, fluent responses are not necessarily factually correct. The TruthfulQA benchmark showed that language models can reproduce common misconceptions, highlighting the limitations of relying on generated text alone to assess reliability (Lin et al., 2022). Ji et al. (2023) further examined hallucination in natural language generation, emphasizing the distinction between factual correctness and consistency with source information. These concerns are particularly relevant when generated answers are expected to be supported by authoritative documents.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

2.2 Retrieval-Augmented Generation and Faithfulness

Retrieval-Augmented Generation (RAG) combines information retrieval with language generation to provide external knowledge during response generation. REALM incorporated retrieval into language-model pre-training (Guu et al., 2020), while Lewis et al. (2020) developed a retrieval-augmented framework for knowledge-intensive NLP tasks. Subsequent research has examined the limitations of retrieval-supported generation. Self-RAG uses adaptive retrieval and self-reflection to improve generation quality (Asai et al., 2024), while RAGTruth provides annotated examples of unsupported or contradictory claims in RAG responses (Niu et al., 2024). Liu et al. (2024) also demonstrated that the position of relevant information within long contexts can affect model performance. These studies indicate that retrieval can provide useful evidence, but its effectiveness depends on the relevance and completeness of the retrieved context and the model's use of that context.

2.3 Evaluation of Generated Answers

Evaluating generated answers requires more than checking whether they address a question. Accuracy concerns correctness, relevance concerns alignment with the question, and faithfulness concerns support from the available evidence. FActScore evaluates factual precision by examining individual claims against a knowledge source (Min et al., 2023). RAGAs assesses multiple dimensions of retrieval-supported generation, including context relevance, answer faithfulness, and answer relevance (Es et al., 2024). ARES similarly evaluates retrieval-augmented systems using separate quality dimensions (Saad-Falcon et al., 2024). These approaches motivate evaluating answer quality across multiple criteria rather than relying on a single overall score. The present study uses a manual scoring rubric rather than implementing these automated evaluation frameworks.

2.4 Research Gap

Existing research has investigated retrieval-augmented generation, factuality, hallucination, and answer evaluation. However, the practical effect of retrieval-supported generation can vary with the questions, source documents, retrieval configuration, and evaluation method. This study conducts a small controlled comparison using 20 fixed questions and two authoritative sources: the NIST Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile and UNESCO's Guidance for Generative AI in Education and Research. Answers generated directly by ChatGPT are compared with answers produced using retrieved document context and an explicit context-only instruction. Accuracy, relevance, faithfulness, and hallucination-free quality are assessed separately. This comparison examines whether improvements in source faithfulness are accompanied by improvements in other answer-quality measures.

3. METHODOLOGY AND EXPERIMENTAL SETUP

3.1 Research Design

This study uses an experimental comparison to evaluate answers generated through direct ChatGPT interaction and Retrieval-Augmented Generation (RAG). A fixed set of 20 questions was used in both conditions, resulting in 40 answers for evaluation. The same question wording was maintained across the two conditions to support a consistent comparison.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

3.2 Source Documents and Question Set

The experiment used two knowledge sources: the Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile published by the National Institute of Standards and Technology (NIST), and UNESCO's Guidance for Generative AI in Education and Research. The question set contained 20 questions covering generative AI concepts, risks, reliability, responsible use, and educational applications. Questions Q01–Q10 were categorized under NIST, while Q11–Q20 were categorized under UNESCO.

3.3 Document Processing and Retrieval

The PDF documents were processed using the pypdf library to extract their text. The extracted text was divided into overlapping chunks using LangChain's RecursiveCharacterTextSplitter, with a chunk size of 1,000 characters and an overlap of 200 characters. The all-MiniLM-L6-v2 sentence-transformer model was used to generate normalized embeddings for the chunks. These embeddings were stored in a FAISS IndexFlatIP index. For each question, an embedding was generated using the same model, and the five highest-ranked chunks were retrieved through similarity search. The retrieved passages, source information, and similarity scores were recorded for use in the RAG procedure.

3.4 Answer Generation

In the baseline condition, the 20 questions were submitted directly to ChatGPT without retrieved document context. In the RAG condition, each question was accompanied by the five retrieved chunks and an instruction to answer using only the supplied context. The instruction also directed the model not to use outside knowledge, browse the web, or invent information, and to acknowledge when the retrieved context was insufficient. Answers were limited to two to four sentences.

The generated answers were recorded for evaluation. Since the RAG procedure combines document retrieval with an explicit context-only instruction, the comparison evaluates the combined approach rather than the isolated effect of retrieval alone.

3.5 Prototype Interface

A lightweight web-based prototype was developed to demonstrate the implemented experimental workflow. The prototype allows users to select an experimental question and view the corresponding baseline and RAG-generated responses along with their evaluation scores. It was used to organize and present the outputs of the experiment and provide a practical interface for comparing the two generation approaches. Figure 1 shows the question-selection interface of the developed prototype, while Figure 2 presents the Baseline and RAG responses along with their evaluation scores for a selected experimental question.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

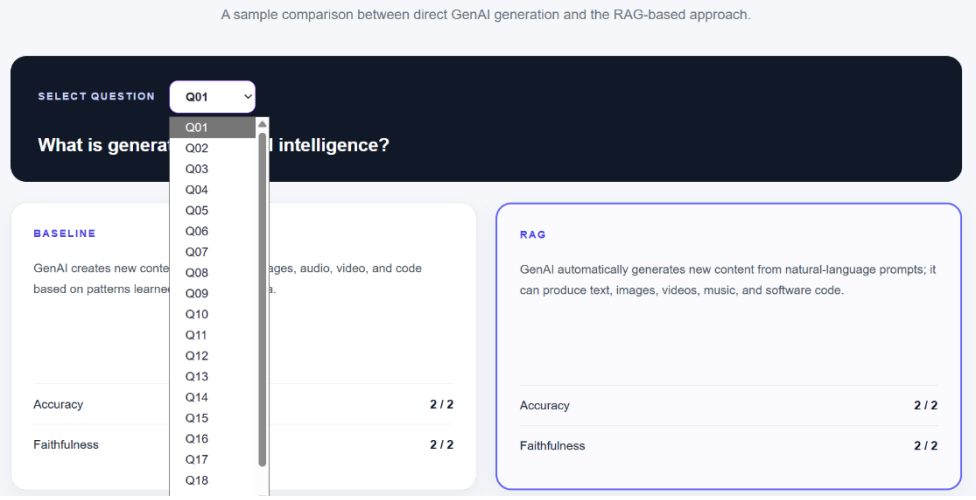


Figure 1: Question – selection interface of the developed web-based prototype

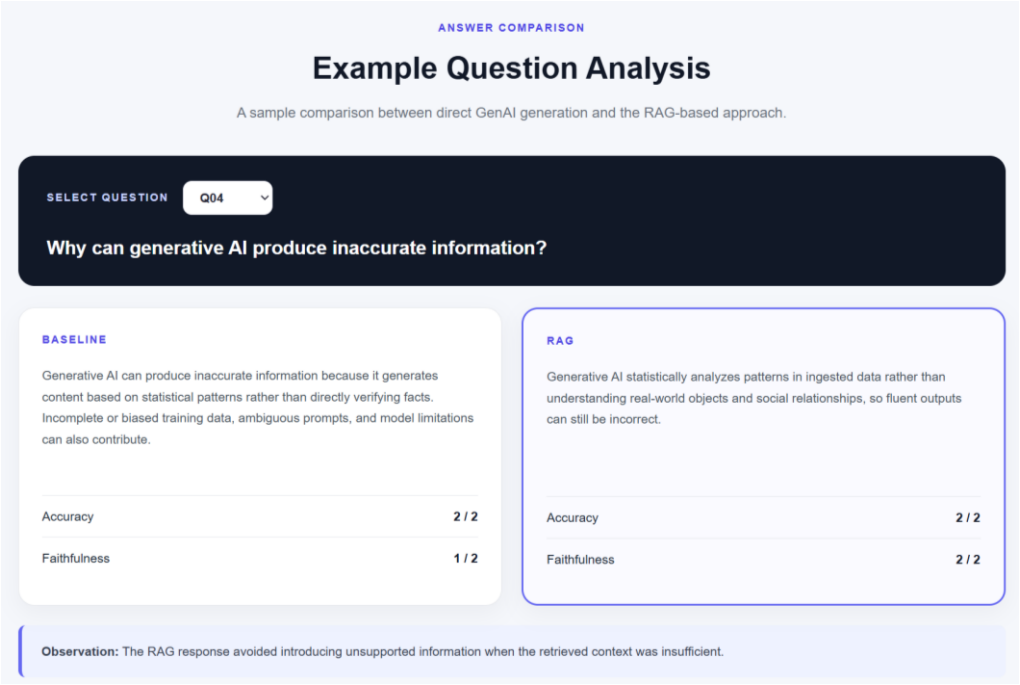


Figure 2: Prototype interface showing Baseline and RAG responses with evaluation scores.

4. EVALUATION CRITERIA

The generated answers were evaluated using four criteria: accuracy, relevance, faithfulness to source content, and hallucination-free quality. Each criterion was assigned a score on a three-point scale from 0 to 2, representing no, partial, or full satisfaction of the criterion, respectively.

Accuracy assessed the factual correctness of each answer, while relevance measured how directly it addressed the question. Faithfulness evaluated whether the answer was supported by the relevant source content. Hallucination-free quality assessed whether the answer contained unsupported or invented information.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

Scores for the baseline and RAG conditions were recorded in an evaluation sheet containing separate fields for each criterion. Python was used to aggregate the scores and generate comparative results. For each criterion, the scores across the 20 questions were summed, producing a maximum of 40 points. The percentage score was calculated by dividing the total by 40 and multiplying by 100.

An overall score was calculated by summing the scores for all four criteria, resulting in a maximum of 160 points per condition. The overall percentage was obtained by dividing the total score by 160 and multiplying by 100. The resulting summaries were exported to a CSV file, and charts were generated to compare baseline and RAG performance across the evaluation criteria.

5. RESULTS

The experiment compared direct ChatGPT generation with retrieval-supported generation across 20 questions. Four evaluation criteria were considered: accuracy, relevance, faithfulness to source content, and hallucination-free quality. Each criterion had a maximum score of 40, while the overall maximum across all four criteria was 160.

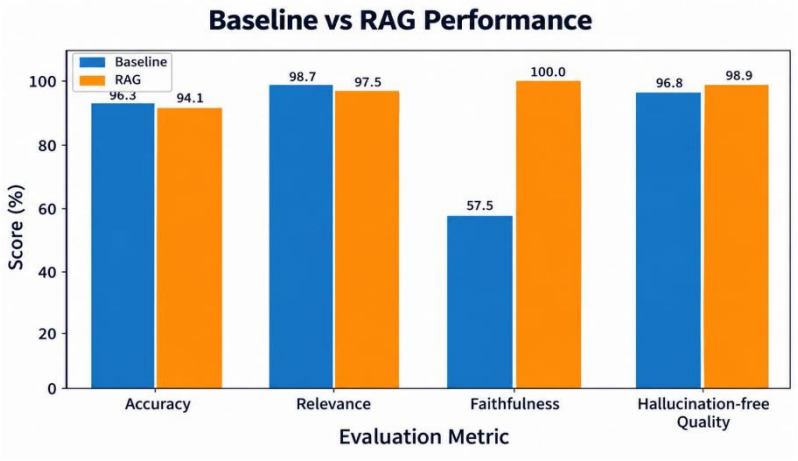


Figure 3: Baseline and RAG performance across evaluation criteria

5.1 Comparison Across Evaluation Criteria

The baseline achieved an overall score of 143/160 (89.38%), whereas the RAG condition achieved 158/160 (98.75%), representing an improvement of 9.38 percentage points.

For accuracy, the baseline scored 40/40 (100%), while RAG scored 38/40 (95%). Relevance remained at 40/40 (100%) for both conditions. Faithfulness increased from 23/40 (57.5%) in the baseline to 40/40 (100%) with RAG. Hallucination-free quality also remained at 40/40 (100%) in both conditions under the applied scoring rubric.

5.2 Overall Performance and Faithfulness

The results show that the improvement in overall score was primarily associated with the increase in faithfulness to source content. The RAG condition achieved full marks for faithfulness, indicating that its

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

answers received higher scores for alignment with the supplied source material. However, its accuracy score was slightly lower than that of the baseline.

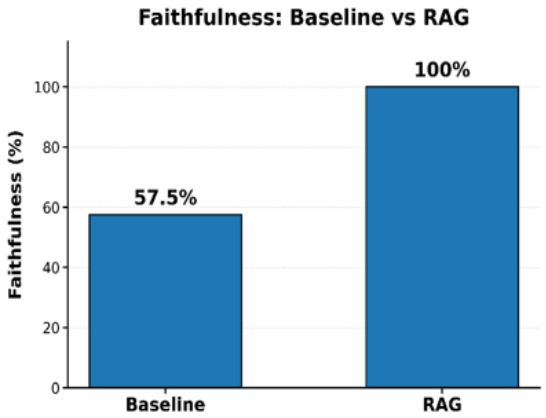


Figure 4: Faithfulness comparison between Baseline and RAG

These findings demonstrate that improvements across answer-quality criteria were not uniform. Although RAG achieved a higher overall score, the reduction in accuracy highlights the importance of evaluating individual criteria alongside aggregate performance.

Evaluation Criterion	Baseline Score	Baseline(%)	RAG Score	RAG (%)
Accuracy	40/40	100.00%	38/40	95.00%
Relevance	40/40	100.00%	40/40	100.00%
Faithfulness	23/40	57.50%	40/40	100.00%
Hallucination-free Quality	40/40	100.00%	40/40	100.00%
Overall Score	**143/160**	**89.38%**	**158/160**	**98.75%**

Table 1: Comparison of Baseline and RAG Performance

6. DISCUSSION

The experimental results indicate that Retrieval-Augmented Generation (RAG) improved the overall quality score of generated answers compared with direct ChatGPT generation. The overall score increased from 89.38% in the baseline condition to 98.75% in the RAG condition, an improvement of 9.38% points. However,

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

the improvement was not consistent across all evaluation criteria. The most notable difference was observed in faithfulness to source content, which increased from 57.50% in the baseline condition to 100% in the RAG condition. This suggests that supplying retrieved document passages and instructing the model to rely only on the provided context helped produce answers more closely aligned with the selected sources. The result highlights the potential value of combining external document retrieval with explicit source-grounding instructions. In contrast, accuracy decreased slightly from 100% in the baseline condition to 95% in the RAG condition. This indicates that source-grounded generation does not necessarily guarantee a fully correct answer. If the retrieved passages do not contain sufficient information, the model may provide an incomplete response or acknowledge that the available context is insufficient. The drop in accuracy under RAG suggests that gaps in the retrieved context carry through directly into the final answer. Relevance and hallucination-free quality both received scores of 100% under the applied evaluation rubric in both conditions. These scores should be interpreted within the scope of the selected questions, source documents, and scoring method; they do not establish that either system is universally free from hallucinations. Overall, the findings suggest that RAG can improve the alignment of generated answers with selected source documents, but its effectiveness should be assessed using multiple evaluation criteria. Since this experiment used 20 questions and two source documents, further testing with larger question sets, additional sources, and independent evaluators would help establish how consistently these findings apply to other tasks.

7. LIMITATIONS AND FUTURE WORK

This study has certain limitations that should be considered when interpreting the findings. First, the experiment was conducted using a fixed set of questions derived from two source documents: the NIST Generative Artificial Intelligence Profile and UNESCO's Guidance for Generative AI in Education and Research. Although these sources provide authoritative and relevant content, the limited question set and document collection may affect the generalizability of the findings to other subject areas, document collections, and question types.

Second, the RAG procedure employed a fixed retrieval configuration, using chunks of 1,000 characters with an overlap of 200 characters and retrieving the top five chunks for each question. Different chunk sizes, embedding models, and retrieval configurations may influence system performance. In addition, the experiment combined retrieval with an explicit instruction to generate answers only from the supplied context. Consequently, the individual contribution of retrieval and prompting was not evaluated separately.

Third, answer quality was assessed using a four-criterion scoring rubric. While this provided a structured basis for comparing the baseline and RAG approaches, the results may be influenced by the definition and application of the selected evaluation criteria. The study did not include independent evaluation by multiple reviewers or a larger evaluation dataset, which could provide further evidence regarding the consistency of the observed results.

Future work can extend this study by employing a larger and more diverse set of questions, additional authoritative document collections, and multiple language models. Experiments with different retrieval configurations and separate evaluation of retrieval quality and answer generation could provide further insight into the factors affecting response reliability. Independent evaluation by multiple reviewers could also strengthen the robustness and consistency of the evaluation process. These extensions would provide a broader

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

understanding of the conditions under which retrieval-supported generation can improve the reliability of document-grounded question answering.

8. CONCLUSION

This study experimentally compared direct ChatGPT generation with a Retrieval-Augmented Generation (RAG) approach for answering questions based on selected authoritative documents. Using a fixed set of 20 questions, the experiment evaluated answer accuracy, relevance, faithfulness to source content, and hallucination-free quality.

The RAG approach achieved an overall score of 98.75%, compared with 89.38% for the baseline approach, representing an improvement of 9.38%. The most notable improvement was observed in faithfulness, which increased from 57.50% to 100%. However, accuracy decreased slightly from 100% to 95%, indicating that retrieval-supported generation does not necessarily improve every evaluation dimension.

The findings indicate that combining document retrieval with explicit context-based instructions can improve the alignment of generated answers with the selected source material. However, the findings should be interpreted within the scope of the experimental conditions and evaluation criteria used in this study. Further evaluation using larger and more diverse datasets, additional source collections, and independent reviewers would help assess the broader applicability of the approach.

Overall, this study demonstrates the potential of RAG for improving the reliability of source-grounded question answering and highlights the importance of evaluating GenAI systems across multiple quality criteria rather than relying on a single measure of answer quality.

REFERENCES

[1] 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 (ICLR 2024).
[2] Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi, E., Hall, P., & Roberts, K. (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.600-1
[3] Brown, T. B., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33.
[4] 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://doi.org/10.18653/v1/2024.eacl-demo.16
[5] 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:2312.10997.
[6] Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M.-W. (2020). Retrieval augmented language model pre-training. Proceedings of the 37th International Conference on Machine Learning, 119, 3929–3938.
[7] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), Article 248. https://doi.org/10.1145/3571730
[8] 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.
[9] Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 3214–3252. https://doi.org/10.18653/v1/2022.acl-long.229
[10] 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. https://doi.org/10.1162/tacl_a_00638
[11] Miao, F., & Holmes, W. (2023). Guidance for generative AI in education and research. UNESCO. https://doi.org/10.54675/EWZM9535
[12] Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.-T., Koh, P., Iyyer, M., Zettlemoyer, L., & Hajishirzi, H. (2023). FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 12076–12100. https://doi.org/10.18653/v1/2023.emnlp-main.741

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

[13] Niu, C., Wu, Y., Zhu, J., Xu, S., Shum, K., Zhong, R., Song, J., & Zhang, T. (2024). RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 10862–10878. https://doi.org/10.18653/v1/2024.acl-long.585

[14] Saad-Falcon, J., Khattab, O., Potts, C., & Zaharia, M. (2024). ARES: An automated evaluation framework for retrieval-augmented generation systems. Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 338–354. https://doi.org/10.18653/v1/2024.naacl-long.20

(This work is licensed under a Creative Commons Attribution 4.0 International License.)