

A Systematic Review of Hallucination Mitigation in Large Language Models Using Retrieval-Augmented Generation

Gangotri Ravinder Balla, Prea Binu Varghese, Rutuja Suresh Farkande, Ms. Yogeshwari Yawalkar- DYPIMCAM, Akurdi, Pune.

ABSTRACT

Large Language Models (LLMs) have become widely used for question answering, education, research, summarization and information support. Despite their usefulness, LLMs may generate false, unsupported or inconsistent information, commonly described as hallucination. Retrieval-Augmented Generation (RAG) addresses this reliability problem by retrieving relevant information from external knowledge sources and supplying it to the language model as contextual evidence before generating a response. This paper studies the role of RAG in reducing factual hallucinations through a review of existing research and a 50-question pilot comparison of responses generated with and without retrieval support. The submitted report records 49 of 50 responses as correct without RAG and all 50 as correct with RAG; it separately presents a clearly labelled illustrative distribution of 46/3/1 versus 48/2/0 for demonstrating response categories. The distinction is retained to avoid presenting illustrative values as observed experimental results. The study also examines limitations including poor retrieval, unreliable sources, incomplete context and incorrect interpretation of retrieved evidence. The paper concludes that RAG is a practical hallucination-reduction approach, but it should be combined with high-quality retrieval, evidence verification and human oversight for critical applications.

Keywords: *Large Language Models; Hallucination; Retrieval-Augmented Generation; Factuality; Information Retrieval; Generative AI*

1. INTRODUCTION

Large Language Models are artificial intelligence systems trained on large collections of text to understand and generate human-like language. They are now used in question answering, content generation, education, research assistance, software development and customer support. However, fluent language does not guarantee factual correctness; a response can appear convincing while containing unsupported, outdated or incorrect information.

LLM hallucination occurs when a model produces information that is false, unsupported or inconsistent with available evidence. The problem is especially important in information-sensitive applications because users may accept fluent responses without independently checking every statement. Hallucinations can be influenced by limitations in training data, incomplete knowledge, ambiguous questions, outdated information and the model's tendency to produce plausible-sounding text without explicit verification.

Retrieval-Augmented Generation provides a mechanism for improving factual grounding. Instead of depending only on knowledge encoded in model parameters, a RAG system retrieves relevant documents or passages from an external knowledge source and provides them to the LLM as context. The model can then use the retrieved evidence while generating its answer. This is particularly useful when information is domain-specific, current or stored in organizational documents.

The purpose of this research is to study the effectiveness of RAG in reducing factual hallucination and improving the reliability of LLM-generated responses. The study combines a literature review with a 50-question pilot comparison and examines accuracy, evidence support, verifiability and the limitations of retrieval-based generation.

2. LITERATURE REVIEW

Research on hallucination reduction has developed from improving the internal knowledge of language models toward connecting models with external evidence. REALM established the value of retrieval for language-model pre-training, while the original RAG framework combined parametric model knowledge with non-parametric external knowledge. Subsequent research introduced citation-supported generation, evaluation frameworks, adaptive retrieval and self-reflection. These developments show that reliable generation requires more than fluent text generation; the information used by the model must also be relevant, trustworthy and verifiable.

Research on RAG evaluation has emphasized that a final answer should not be judged in isolation. Retrieval relevance, factual faithfulness, completeness and evidence quality all influence reliability. RAGTruth was developed to study hallucinations in retrieval-augmented systems and demonstrates that even when retrieval is present, models can produce unsupported or contradictory claims. Self-RAG explores selective retrieval and self-reflection,

showing that retrieving information indiscriminately is not always beneficial.

Recent research continues to investigate corrective retrieval, multi-step retrieval and evidence-grounded verification. FRAMES provides a challenging evaluation setting for factuality, retrieval and reasoning and reports gains from retrieval on multi-hop questions. These findings support the central argument of this research: RAG can reduce hallucination risk when retrieval quality and evidence use are carefully controlled, but retrieval alone cannot guarantee factual correctness.

2.1 Comparative Review of Selected Research Studies

Author & Year	Research Topic	Method	Main Finding	Limitation
Guu et al. (2020)	REALM	Retrieval during pre-training	Value of external retrieval for knowledge-intensive tasks.	Additional retrieval/training complexity.
Lewis et al. (2020)	RAG	Retriever + generator	Combines parametric and external knowledge.	Dependent on retrieved evidence.
Gao et al. (2023)	ALCE	Evidence + citations	Improves attribution and verifiability.	Citation completeness remains challenging.
Es et al. (2023)	RAGAS	Evaluation metrics	Evaluates retrieval relevance, faithfulness and quality.	Automated metrics have limits.
Gao et al. (2023)	RAG Survey	Naive, advanced and modular RAG	Organizes major components and challenges.	Does not solve retrieval errors.
Asai et al. (2024)	Self-RAG	Adaptive retrieval + reflection	Selective retrieval and critique can improve factuality.	More complex training/inference.
Niu et al. (2024)	RAGTruth	Hallucination corpus	Supports systematic study of RAG hallucinations.	Benchmark coverage is not universal.
Vu et al. (2024)	FreshLLMs	Search augmentation	Helps with changing factual information.	Depends on search/source quality.
Yan et al. (2024)	Corrective RAG	Retrieval evaluation + correction	Can address weak retrieval.	Adds system complexity.
Yu et al. (2024)	RAG Evaluation Survey	Unified evaluation	Compares relevance, accuracy and faithfulness.	Dynamic-source evaluation remains difficult.
Krishna et al. (2025)	FRAMES	Factuality, retrieval, reasoning	Shows gains on challenging multi-hop evaluation.	Demanding specialized benchmark.
Bécharde & Ayala (2024)	Structured-output RAG	RAG-based generation	Reports reduced hallucination	Domain-specific implementation

Author & Year	Research Topic	Method	Main Finding	Limitation
			in structured generation.	n.

2.2 Comparative Discussion

The reviewed studies demonstrate a progression from making external knowledge available to language models toward evaluating whether retrieved evidence is relevant, whether the final answer is faithful to that evidence, and whether important claims can be verified. This progression is significant because inserting retrieved passages into a prompt does not automatically make a response factual.

Adaptive and corrective retrieval are important later developments. Self-RAG explores retrieval and self-reflection, while corrective approaches evaluate weak retrieval and attempt to obtain better evidence. RAGTruth and FRAMES also demonstrate that difficult questions are more suitable for measuring the contribution of RAG than simple factual questions. Future experiments should therefore include multi-hop, ambiguous, time-sensitive and evidence-conflict questions.

3. RESEARCH OBJECTIVES

The primary objective is to study the effectiveness of Retrieval-Augmented Generation in reducing factual hallucinations produced by Large Language Models. The study also aims to understand major causes of hallucination, compare responses generated with and without retrieval support, examine the contribution of external evidence to factual reliability, identify limitations of RAG, and recommend practices for developing more trustworthy RAG-based applications.

4. RESEARCH METHODOLOGY

The research follows a comparative methodology supported by a small pilot response comparison. The literature review examines research on LLM hallucination, external knowledge retrieval, RAG, citation-supported generation, evaluation and hallucination detection. The experimental component compares two conditions: an LLM answer using internal knowledge without an explicit retrieval step, and an answer generated using retrieved external information as supporting context.

A set of 50 factual questions covering science, history, geography, technology, computer science, artificial intelligence, cybersecurity and public health was considered. Each response was compared with a reliable reference answer. Evaluation considered correctness, partial correctness, incorrectness, evidence support,

factual consistency, completeness and verifiability. The pilot is limited in size and contains relatively straightforward questions, so its findings should not be treated as statistically generalizable.

For research integrity, the submitted report distinguishes recorded pilot values from an illustrative response distribution. The illustrative values were introduced to demonstrate how response categories could be visualized. They should not be described as observed question-level model outputs unless corresponding response transcripts are available.

5. DATA ANALYSIS AND RESULTS

5.1 Quantitative Analysis of the 50-Question Pilot

The submitted report records 49 of 50 responses as correct without RAG and all 50 as correct with RAG. It separately presents an explicitly labelled illustrative distribution of 46 correct, 3 partially correct and 1 incorrect response without RAG, compared with 48 correct, 2 partially correct and 0 incorrect responses with RAG. Because these are two different descriptions in the source report, this paper does not merge them. The following table and figures reproduce the illustrative distribution only.

Evaluation Category	Without RAG (n=50)	With RAG (n=50)
Correct	46	48
Partially Correct	3	2
Incorrect	1	0
Accuracy	92%	96%

The illustrative distribution corresponds to 92% accuracy without retrieval and 96% accuracy with retrieval. It indicates a reduction in incorrect responses from 1 to 0 and a reduction in partially correct responses from 3 to 2. These values demonstrate the analysis structure but must not be treated as independently observed findings.

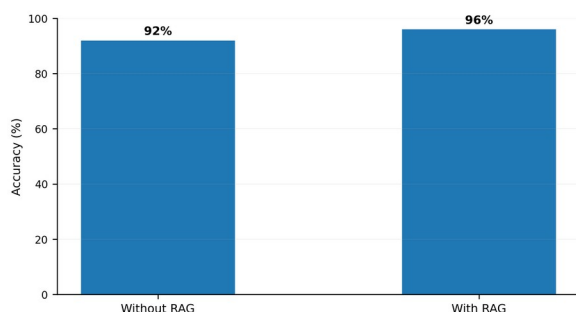


Figure 1. Accuracy comparison for the 50-question pilot.

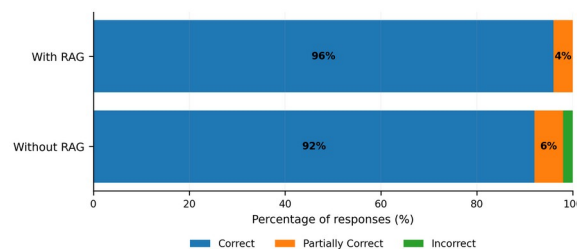


Figure 2. Distribution of response outcomes.

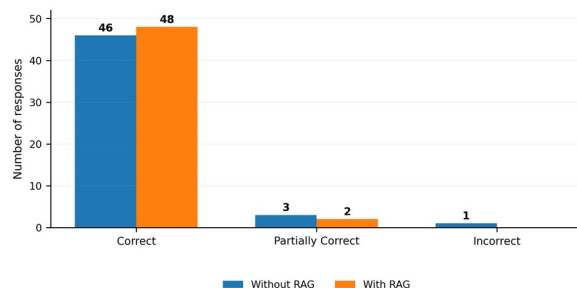


Figure 3. Response classification with and without RAG.

5.2 Question-wise Evaluation

The source report evaluates 50 questions across eight factual domains: Science, History, Geography, Computer Science, Artificial Intelligence, Cybersecurity, Technology and Public Health. Correct means the essential fact is accurate; Partially Correct means the central idea is correct but an important element is incomplete or slightly inaccurate; Incorrect means that the central answer is wrong. The source presents a question-wise illustrative mapping rather than a complete model-response transcript.

Question	Evaluation Item	Without RAG	With RAG
Q11	Who founded the Maurya Empire?	Partially Correct	Correct
Q32	What is an AI hallucination?	Incorrect	Correct
Q34	Main purpose of retrieval in RAG?	Partially Correct	Partially Correct
Q42	What is cloud computing?	Partially Correct	Partially Correct
Other 46 questions	Remaining questions	Correct	Correct

In the illustrative mapping, Q11, Q34 and Q42 are Partially Correct without RAG, while Q32 is the single Incorrect case. With RAG, Q11 and Q32 move to Correct, while Q34 and Q42 remain Partially Correct. All remaining questions are classified as Correct. This produces the illustrative totals of 46/3/1 without RAG and 48/2/0 with RAG.

5.3 Interpretation

The results indicate that retrieval can provide an additional evidence layer and may improve factual grounding. However, the small sample and straightforward factual questions limit the strength of the

conclusion. The study therefore treats RAG as a hallucination-reduction mechanism rather than a guarantee of factual correctness. A stronger empirical evaluation would require recorded model outputs, larger and more difficult datasets, and statistical testing.

6. APPLICATIONS OF RAG

RAG can be applied wherever an LLM needs to generate answers grounded in a controlled information source. In education, it can retrieve textbooks, academic material, course notes and institutional documents. In research, it can retrieve papers and supporting passages for literature summaries. In customer support, product manuals, frequently asked questions and internal documentation can be connected to the model. In healthcare and legal applications, RAG can support information retrieval from approved sources, although critical decisions must remain subject to professional verification.

7. BENEFITS OF RAG

The primary benefit of RAG is improved factual grounding because the language model can use information retrieved at inference time instead of relying entirely on stored parametric knowledge. This is useful for current, domain-specific and organization-specific information. Retrieved evidence can also improve transparency and verifiability because users can inspect the source material behind important claims. RAG can connect existing language models to specialized document collections without requiring complete retraining.

8. CHALLENGES OF RAG

RAG does not eliminate hallucination because retrieved evidence itself may be irrelevant, incomplete, outdated or unreliable. A language model may also misinterpret correct evidence or combine multiple passages incorrectly. Other challenges include document chunking, embedding quality, ranking, context-window limitations, citation completeness, latency and computational cost. For high-risk domains, a RAG system should provide source evidence and retain human review rather than presenting generated output as automatically authoritative.

9. RESEARCH GAP

The reviewed literature shows that important gaps remain in reliable hallucination reduction. Existing RAG systems do not guarantee that retrieved information will be relevant or trustworthy, and evaluation results can vary according to the dataset, retriever, model and question difficulty. There is a need for evaluation methods that

jointly measure retrieval quality, factual faithfulness, completeness and claim-level evidence support. Simple factual questions are insufficient for demonstrating the full benefit of RAG, creating a need for more challenging and realistic evaluation datasets. A further gap concerns verification after retrieval: systems should evaluate evidence, identify contradictions, verify important claims and decide whether additional retrieval is necessary.

10. FUTURE SCOPE

Future research can extend the present pilot by using a larger dataset with hundreds or thousands of questions and balancing questions across difficulty levels and application domains. Statistical testing can determine whether observed improvements are significant rather than relying only on percentages. Another direction is adaptive retrieval, in which the system retrieves information only when necessary and performs additional retrieval when initial evidence is weak. Future systems can also combine hybrid retrieval, knowledge graphs, citation generation, claim-level verification and uncertainty estimation. In sensitive domains, systems should provide transparent evidence, communicate uncertainty and retain human oversight.

11. CONCLUSION

Large Language Model hallucination is a significant challenge for reliable generative artificial intelligence because fluent responses may still contain unsupported or incorrect information. Retrieval-Augmented Generation provides a practical approach by connecting an LLM with external information and using retrieved evidence as context during generation. The reviewed literature demonstrates that retrieval can improve factual grounding, while newer research emphasizes adaptive retrieval, corrective retrieval, citation support and explicit verification. The submitted pilot records 49 correct responses without RAG and 50 correct responses with RAG, while its illustrative analysis demonstrates how a distribution of 46/3/1 versus 48/2/0 can be visualized. Because the dataset is small and relatively simple, neither presentation should be interpreted as proof that RAG eliminates hallucination. Overall, RAG is best understood as a practical hallucination-reduction technique whose effectiveness depends on source quality, retrieval relevance and correct use of evidence.

12. REFERENCES

1. Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M. (2020). REALM: Retrieval-Augmented Language Model Pre-Training. Proceedings of ICML.
2. Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems.
3. Gao, T., Yen, H., Yu, J., & Chen, D. (2023). Enabling Large Language Models to Generate Text with Citations. Proceedings of EMNLP.
4. Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2023). RAGAS: Automated Evaluation of Retrieval Augmented Generation. arXiv:2309.15217.
5. Gao, Y., Xiong, Y., Gao, X., et al. (2023). Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv:2312.10997.
6. Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. ICLR 2024.
7. 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 ACL 2024.
8. Vu, T., Iyyer, M., Wang, X., et al. (2024). FreshLLMs: Refreshing Large Language Models with Search Engine Augmentation. Findings of ACL 2024.