International Journal of Engineering, Science and Mathematics
Vol. 15 Issue 8,Aug 2026,
ISSN: 2320-0294 Impact Factor: 8.215
Journal Homepage: http://www.ijesm.co.in, Email: ijesmj@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

Reducing Hallucinations in Large Language Models Using Retrieval-Augmented Generation

Tanmay Sharma s/o Dr. D.K. Sharma,

Class 12th(PCM), Narayana E-techno school, Vashali nagar, Jaipur (Rajasthan) Vaishali Nagar

Abstract

Large Language Models (LLMs) are commonly used for answering questions, writing texts, summarizing documents, and supporting various digital services. Nevertheless, such models may sometimes give out facts which appear to be true yet are incorrect and unsupported by evidence. A hallucination is this issue. Hallucinations are dangerous as people might believe inaccurate information, particularly within education, health care, law enforcement, finance and research. RAG is one of those methods that help with it. An RAG system searches an external knowledge base for information about a user's query. Then it passes this data on to a language model prior to generating an answer. This paper explains hallucination in simple terms and studies how RAG can reduce it. Also it covers some of the key phases for a RAG system such as document creation, search and retrieval, generation and assessment. This paper reviews recent research on RAG, Self-RAG, Corrective RAG, Adaptive-RAG, RAGAS, RAGTruth, and related approaches. RAG improves facts' reliability; it provides current information; it makes it easier for you to check your answer. However, RAG does not completely eliminate hallucination. Bad retrieval, weak sources, conflicting documents, and incorrect use of retrieved evidence can still lead to false answers. RAG works best with reliable sources, good retrieval, clear prompting, source verification and answer evaluation.

Keywords: Large Language Models, hallucination, Retrieval-Augmented Generation, RAG, factual accuracy, information retrieval, artificial intelligence

1. Introduction

Large Language Models are trained with a lot of text and can produce human-like responses. They can explain topics, translate languages, summarize information, python code writing and support question answering system. Due to such capabilities, LLMs are

used for educationally; research-wise; business-wise; customer support and software development.

While LLMs are useful they have a big weakness. Sometimes they give an answer that is flu fluent and confident but factually wrong. The model may invent a name, date, reference, event, or explanation. It may also contain correct and incorrect information in the same response. It's called a hallucination. Ji et al. (2023) describe hallucination as generated information that is not properly supported by the source or by factual knowledge. A hallucination is serious as people may not realize if their answer is wrong.

One of them are how a, for example, an LLM generates text. Predicts probable answers based on, rather than checking an authoritative source on every question asked. Its knowledge may also be incomplete or outdated. Thus, it may give a good-looking response with insufficient credible evidence.

RAG is a method for connecting language generation with external data. Lewis et al. (2020) showed that a model could combine knowledge stored in its parameters with information retrieved from an external collection. It enables it, for example, to base a decision on facts about this particular issue instead of relying solely upon knowledge that is already known by them. Later studies have led to better versions of RAG which improve retrieval, verify retrieved documents, and assess if generated answers are supported by evidence.

This article imperfect paper explains how RAG helps to minimize hallucinations from a large language model. It explains how hallucinations happen, how a basic RAG works, what recent research has found, and why RAG can still fail. This article uses easy-to-understand language to ensure that readers can understand it without any knowledge about artificial intelligence.

2. Objectives of the Study

This study aims at exploring a way to reduce hallucinations with help from Retrieval-Augmented Generation (RAG) for Large Language Models. To explain some of the causes for hallucinations; describe how a RAG works; discuss current techniques for improving it; identify some of its limitations that might lead to wrong answers; and suggest ways that can be taken to improve reliability of these systems.

3. Research Questions

The paper is based on five research questions. Why do Large Language Models generate hallucinated information. Also, what is Retrieval-Augmented Generation. Secondly, how can RAG improve the factual accuracy of generated answers. Fourth, what problems can

still lead to hallucination in a RAG system. Fifth, what techniques could be used to enhance the reliability of RAG for future applications.

4. Research Methodology

The research is based upon reviews. It does not train a new language model nor collect data from humans. Instead, it studies published research on hallucination and Retrieval-Augmented Generation. Selected important papers were taken from reputable sources such as the Association for Computational Linguistics, ACM Computing Surveys, Transactions of the Association for Computational Linguistics, conference proceedings, and widely used research preprints.

Four simple areas of comparison between the reviewed studies are how the system retrieves information, how the model uses retrieved information, how factual accuracy is evaluated, and what types of errors remain. Special attention was given to the original RAG model, Self-RAG, Corrective RAG, Adaptive-RAG, RAGAS, RAGTruth, RAFT, and research on long-context use. It's also appropriate as this approach will help with understanding current approaches and comparing them to existing ones instead of claiming an experimental result.

5. Understanding Hallucination in Large Language Models

1. Simple meaning: Hallucination is when an LLM gives information that sounds correct but is actually wrong, made up, or not supported by a reliable source. The answer may be written clearly and confidently, so a user may believe it. That's because an answer that is fluent does not necessarily mean that it's true. This is why a fluent answer is not always a factual answer.

2. Common examples: It may also create a reference or source that does not exist. At times a model gives both right and wrong answers together. It is more difficult to detect this mistake by people. Sometimes the model mixes correct and incorrect information in the same answer. The model may still attempt to answer rather than saying it is unsure.

3. Why it happens: Hallucination can happen when the model does not have enough information, its learned knowledge is old, the question is unclear, or the prompt contains confusing information. Therefore, an answer from an LLM must be backed up by credible sources that people can verify. Ji et al. (2023) explain that hallucination can come from problems in data, model training, and text generation.

4. Why it is a problem: Why it is a problem: A small factual mistake may be harmless in casual conversation, but wrong information can be serious in healthcare, law, finance,

education, and research. For this reason, LLM answers should be connected with reliable evidence that users can check.

6. What Is Retrieval-Augmented Generation

1. Simple idea: Augmented Generation (RAG) adds a search step before an LLM gives its final answer. Rather than just knowledge gained through training, it then gets useful data from outside sources.

2. Prepare the information: Prepare the information The RAG system stores trusted material such as research papers, books, policies, manuals, or approved websites in a knowledge base.Big files are broken up and split up so that it's easier for a computer to find out what's inside them.

3. Search for the answer: Prepare the information: The RAG system keeps trusted material such as research papers, books, policies, manuals, or approved websites in a knowledge base.

4. Give evidence to the LLM: Give evidence to the LLM The selected information is sent to the language model together with the user's question.Then it answers with help of questions and retrieved information. Search for the answer: When a user asks a question, the system searches the knowledge base and selects the sections that are most related to that question.

5. Easy example and benefit: If a university changes one of its rules, the new rule can be added to the RAG knowledge base. The LLM can then use the new document when students ask about the rule. The whole language model does not need to be trained again, which makes updating information easier.

7. Literature Review

Lewis et al. (2020) provided the foundation for modern RAG by combining a sequence-to-sequence language model with a dense information retriever. The research demonstrated how retrieval-based approaches are capable of providing better factuality and specificity compared with those that rely solely on their own parameters. The study established that external knowledge is a foundation for creation.

Gao et al. (2023) reviewed basic, advanced, and modular RAG. The survey states that RAG can solve for stale information and hallucination, but final quality depends on both retrieval and generation.

Asai et al. (2024) proposed Self-RAG, which lets the model decide when retrieval is useful and reflect on retrieved information and its own response. Factuality and citation reliability were higher for this task compared to others studied by them.

Yan et al. (2024) proposed Corrective Retrieval-Augmented Generation (CRAG). It checks retrieved document quality and can search for better evidence when the first retrieval is weak. It enhances accuracy if a first search yields bad results.

Jeong et al. (2024) introduced Adaptive-RAG, which chooses a retrieval strategy based on question complexity. Simple questions may need little retrieval, while difficult questions can use several steps. It may also enhance precision and productivity.

Zhang et al. (2024) proposed RAFT for domain-specific RAG. This teaches it how to reference useful information and ignore irrelevant material that is often encountered by people when searching for something.

Niu et al. (2024) created RAGTruth, a large corpus for studying hallucinations in RAG responses. The study included nearly 18,000 model responses with detailed human annotations. It's valuable as this shows how hallucinations may continue to happen despite having access to retrieved information.

RAGAS introduced retrieval relevance and whether answers are supported by retrieved information. Saad-Falcon et al. (2024) proposed ARES to assess context relevance, answer faithfulness, and answer relevance. Both show that the complete RAG pipeline should be evaluated.

8. How RAG Reduces Hallucinations

1. It gives the model evidence: RAG provides useful passages to an LLM prior to generating an answer. It implies that this model is less likely to guess based upon prior knowledge it has learned before. If a retrieved text directly answers your query then it will probably remain factual. When the retrieved passage directly answers the question, the final response is more likely to stay close to the facts.

2. It can use newer information: The documents can also be updated periodically and so this model will be able to respond to queries with current data. It can use newer information: A RAG system can search a current knowledge base even when the LLM was trained earlier. RAFT shows that a model can be trained to use helpful domain documents while ignoring distracting information.

3. It supports special subjects: Then the user will be able to access it from source code to see if it's correct. It increases openness but does not ensure accuracy for all questions asked.

4. It makes answers easier to check: It would be more efficient to save on costs and also provide better access to current data for your organization's needs. The user can then open the source and check whether the answer is supported. This improves transparency, even though it cannot guarantee that every answer will be correct.

5. It makes updating easier: It makes updating easier: When information changes, an organisation can update the documents in the knowledge base instead of training the complete LLM again. This can save time and cost while helping the system use newer and more relevant information.

9. Limitations and Remaining Problems

RAG reduces hallucinations, but it cannot eliminate them entirely. One of them are bad retrieval problems. If the retriever chooses a document that is unrelated or only partly related to the question, the LLM may use that information incorrectly. Thus, search quality is of great importance.

Secondly, it's about sourcing. RAG can provide inaccurate data when it, for example, it retrieves information from a bad database that is outdated and incorrect. Evidence retrieval provides a factuality for models; however, this does not guarantee that they are accurate. Therefore, a trusted application should, should then restrict to sources that it uses.

Conflicting information is, the second issue. Two documents may disagree because one is older or comes from a different authority. Information like date, author, version, and type of source will assist it to choose correctly.

Fourthly, there is too much information. Liu et al. (2024) found that language models do not always use information equally across long inputs. Performance may suffera fall if vital data is put at an intermediate position within an extended text. It means to say retrieving more documents does not always give an answer.

FDRIVER is the fifth problem as it can ignore valid evidence. Even if there is a correct passage available, the LLM might include unsupported information from its internal knowledge. RAGTruth shows that hallucinations are still present in retrieval-supported responses. Thus, retrieval should also include directions for retrieving information and verification techniques to ensure that answers are close enough to evidence.

10. Improving the Reliability of RAG

Some practical ways to enhance an RAG system are provided below. Secondly, it must have reliable and up-to-date information available to it. Duplicate, obsolete, and poor-quality documents must be deleted if they are available.

Also, document should be divided by meaningful segments. Small pieces may lose context and big ones might contain irrelevant details.

Secondly, it must also be tested for retrieving correct results from top of search engine results page. The system may re-rank results and perform a corrective retrieval if it's a poor initial result.

Fourth, the prompt should tell the model how to answer based on given evidence. Also, it should enable the model to state that there is not enough information available for this task. It's better not to force a response.

Fifth, answer should cite where possible. Cite a specific source instead of just giving you some general references to use as answers.

Sixth, RAG should be evaluated regularly. RAGAS and ARES show that testing should measure retrieval relevance, answer relevance, and faithfulness to evidence.

11. Findings and Discussion

Secondly, RAG helps to reduce hallucination as it provides an external source of information to an LLM atlas when answering questions. Retrieval alone isn't sufficient either. Evidence should be retrieved by a system, and then models should utilize this information properly. Secondly, more data isn't always better. Too much information may confuse a machine learning algorithm and make it harder for them to access or access vital facts. Fourth, evaluation is important as an answer may be good but not supported. From basic RAG to Self-RAG, CRAG, Adaptive-R-R-RAG, and RAFT there is an obvious trend to current research.

Today's systems are moving towards "retrieve-and-generate"" pipelines. They are becoming more selective and include checking, reflection, correction, and adaptation. It's crucial as this would not be an effective way of solving a problem with hallucinations by adding an internet search engine. All stages of a process should ensure that it is evidence-based and the, and how evidence relates to conclusions drawn from it. RAG is more of an approach to risk reduction rather than, as an ultimate solution for this problem.

A basic RAG system might suffice for low-risk uses. For high-risk areas, stronger source control, human review and citation, and automated verification should also be

included. In high-risk fields, stronger source control, human review, citations, and automated verification should be added.

12. Conclusion

One of a major problem with Large Language Models is hallucination. An ai model is capable of writing coherent, and confident statements despite having incorrect facts.

Retrieval Augmented Generation (generates an effective way of addressing this issue through linking the model to external data prior to generating an answer. This study shows how RAG improves factual improves accuracy of facts; provides up-to-date and specific information; supports answers by providing sources. However, RAG does not, as well not be completely accurate.

Bad retrieval, low-quality sources; poor evidence; contradictory information; long context; and unsupported generation may also cause hallucinations. Therefore, an effective effective RAG tool needs to be based on credible sources; good retrieval; proper chunking; relevance checking; clear instructions; citations; and regular evaluation. Self-RAG, Corrective RAG, Adaptive-RAG, RAFT are examples of how intelligent and selective retrieval systems have become.

Future research needs to do more on verification of facts; it must also be easier for a model to tell if there is no reliable data available. RAG is an important step towards safer and more trustworthy use of Large Language Models, but it should be combined with careful system design and human responsibility. Future work should continue to improve evidence checking and should make it easier for models to say when reliable information is not available. RAG is therefore an important step toward safer and more trustworthy use of Large Language Models, but it should be combined with careful system design and human responsibility.

References

1. Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In International Conference on Learning Representations (ICLR). https://arxiv.org/abs/2310.11511
2. Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2024). RAGAs: Automated evaluation of Retrieval Augmented Generation. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (pp. 150-158). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.eacl-demo.16

3. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2023). Retrieval-Augmented Generation for Large Language Models: A survey. arXiv. https://arxiv.org/abs/2312.10997
4. Jeong, S., Baek, J., Cho, S., Hwang, S. J., & Park, J. (2024). Adaptive-RAG: Learning to adapt Retrieval-Augmented Large Language Models through question complexity. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Vol. 1, pp. 7036-7050). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.389
5. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), Article 248, 1-38. https://doi.org/10.1145/3571730
6. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., Yih, W.-T., Rocktaschel, 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
7. 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
8. 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. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Vol. 1, pp. 10862-10878). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.585
9. Saad-Falcon, J., Khattab, O., Potts, C., & Zaharia, M. (2024). ARES: An automated evaluation framework for Retrieval-Augmented Generation systems. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics. https://aclanthology.org/2024.naacl-long.20/
10. Yan, S.-Q., Gu, J.-C., Zhu, Y., & Ling, Z.-H. (2024). Corrective Retrieval Augmented Generation. arXiv. https://arxiv.org/abs/2401.15884

11. Zhang, T., Patil, S. G., Jain, N., Shen, S., Zaharia, M., Stoica, I., & Gonzalez, J. E. (2024). RAFT: Adapting language model to domain specific RAG. arXiv. https://arxiv.org/abs/2403.10131