Scholars Journal of Engineering and Technology

OPEN ACCESS

Abbreviated Key Title: Sch J Eng Tech
ISSN 2347-9523 (Print) | ISSN 2321-435X (Online)
Journal homepage: https://saspublishers.com

Retrieval-Augmented Generation and Hallucination in Large Language Models: A Scholarly Overview

Sagar Gupta[1*]

[1]Enterprise Resources Planning Implementation Leader, Advertising Vehicles, Cincinnati, OH, USA

DOI: https://doi.org/10.36347/sjet.2025.v13i05.003 | **Received:** 12.04.2025 | **Accepted:** 16.05.2025 | **Published:** 19.05.2025

***Corresponding author:** Sagar Gupta
Enterprise Resources Planning Implementation Leader, Advertising Vehicles, Cincinnati, OH, USA

Abstract **Review Article**

Large Language Models (LLMs) have revolutionized natural language processing tasks, yet they often suffer from "hallucination" the confident generation of factually incorrect information. Retrieval-Augmented Generation (RAG) has emerged as a promising technique to mitigate hallucinations by grounding model responses in external documents. This article explores the underlying causes of hallucinations in LLMs, the mechanisms and architectures of RAG systems, their effectiveness in reducing hallucinations, and ongoing challenges. We conclude with a discussion of future directions for integrating retrieval mechanisms more seamlessly into generative architecture.

Keywords: Large Language Models (LLMs), Hallucination, Retrieval-Augmented Generation (RAG), Factual Inaccuracy, External Document Retrieval.

1. INTRODUCTION

Large Language Models (LLMs), such as GPT-4, PaLM, and LLaMA, have demonstrated remarkable capabilities across a wide range of natural language tasks, including question answering, summarization, and dialogue (Brown *et al*., 2020; Chowdhery *et al*., 2022; Touvron *et al*., 2023). However, despite their impressive performance, LLMs often generate text that is fluent but factually inaccurate phenomenon known as hallucination (Ji *et al*., 2023). This limitation poses significant challenges, particularly in high-stakes domains such as healthcare, law, and scientific research.

Retrieval-Augmented Generation (RAG) is a hybrid approach that incorporates external document retrieval into the generation process, aiming to ground responses in factual information (Lewis *et al*., 2020). By combining the strengths of traditional information retrieval with neural language modeling, RAG offers a mechanism to reduce hallucinations and enhance factual consistency.

2. Understanding Hallucination in LLMs

2.1 Definition and Taxonomy

Hallucination refers to the generation of content that is not supported by the input or underlying knowledge. Ji *et al*., (2023) classify hallucinations into two broad types:

- **Intrinsic Hallucinations**: Fabrications that conflict with known facts or entail logical inconsistencies.
- **Extrinsic Hallucinations**: Information that cannot be verified against any known source, even if not inherently illogical.

In the context of LLMs, hallucination arises from the model's reliance on probabilistic associations learned from large corpora rather than explicit knowledge of facts.

2.2 Causes

The causes of hallucination in LLMs include:

- **Data Limitations**: LLMs are trained on large but incomplete and noisy datasets.
- **Training Objectives**: Next-token prediction incentivizes fluency over factuality.
- **Lack of Grounding**: Models lack access to up-to-date or verifiable information during inference.
- **Prompt Ambiguity**: Vague prompts can lead the model to "fill in" gaps with plausible sounding but inaccurate content (Shuster *et al*., 2021).

3. Retrieval-Augmented Generation: An Overview

3.1 Concept and Architecture

RAG systems enhance generative models by integrating an information retrieval component that retrieves relevant documents based on the input query. These documents are then used to condition the language model's generation process. Lewis *et al*., (2020) introduced a prominent RAG architecture, consisting of:

1. **Retriever**: Often based on dense passage retrieval (DPR) or BM25, retrieves top-k documents from a corpus.
2. **Reader/Generator**: A sequence-to-sequence model, typically a Transformer, that generates answers based on the retrieved context.

This two-stage pipeline can be trained end-to-end or with frozen components, depending on resource constraints and application requirements.

3.2 Variants and Improvements

Several RAG variants have emerged:

- **Fusion-in-Decoder (FiD)**: Processes multiple documents independently and fuses them within the decoder (Izacard & Grave, 2021).
- **REPLUG**: Uses a plug-and-play retriever with frozen LLMs (Shi *et al*., 2023).
- **Atlas**: Combines dense retrieval with instruction tuning (Izacard *et al*., 2022).

These architectures aim to improve the relevance and integration of retrieved information, thereby enhancing factual accuracy.

4. RAG vs. Hallucination: Empirical Insights

4.1 Hallucination Reduction

Empirical studies have demonstrated that RAG significantly reduces hallucination rates across various tasks. For example, Shuster *et al*., (2021) reported up to a 35% reduction in hallucinated responses in open-domain question answering. Similarly, Borgeaud *et al*., (2022) showed that RETRO, a retrieval-based model, outperforms similarly sized LLMs in factual benchmarks.

4.2 Limitations

Despite its promise, RAG is not a panacea. Hallucinations can still occur due to:

- **Retriever Errors**: Irrelevant or low-quality documents may be retrieved.
- **Reader Overgeneration**: The model may ignore retrieved evidence or generate beyond the evidence boundaries.
- **Latency and Cost**: Retrieval introduces overhead, particularly in real-time applications.

5. Evaluation Methods

Evaluating hallucination and RAG effectiveness remains challenging. Common metrics include:

- **Factual Consistency Metrics**: e.g., FactCC, FEVER, QAGS (Kryściński *et al*., 2020; Thorne *et al*., 2018).
- **Human Evaluations**: Essential for nuanced judgment of truthfulness.
- **Information Sufficiency**: Metrics like F1-overlap or retrieval precision assess whether retrieved documents support the answer.

Recent efforts are aimed at developing reference-free factuality metrics (Dziri *et al*., 2022).

6. Future Directions

6.1 Differentiable Retrieval

Differentiable retrievers, such as ColBERTv2 (Santhanam *et al*., 2022), aim to make retrieval fully end-to-end trainable, improving relevance and alignment between retrieval and generation.

6.2 Memory-Augmented Models

Memory systems, including recurrent retrieval mechanisms and long-context attention, may reduce the need for external retrieval altogether, or complement it (Khandelwal *et al*., 2022).

6.3 Fact-Aware Training

Incorporating factual objectives directly into pretraining or fine-tuning, such as through reinforcement learning from human feedback (RLHF), may further reduce hallucinations (Ouyang *et al*., 2022).

7. CONCLUSION

Retrieval-Augmented Generation represents a promising paradigm for enhancing the factual accuracy of LLMs and reducing hallucinations. While challenges remain, particularly in aligning retriever and generator components, the approach has shown robust empirical gains and continues to evolve. As LLMs become increasingly integrated into high-stakes domains, the combination of generative power with retrieval-based grounding offers a path toward more reliable and trustworthy AI systems.

REFERENCES

- Borgeaud, S., Mensch, A., Hoffmann, J., *et al*. (2022). Improving language models by retrieving from trillions of tokens. *Nature*, *610*(7930), 607–616. https://doi.org/10.1038/s41586-022-04590-6
- Brown, T., Mann, B., Ryder, N., *et al*. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 1877–1901.
- Chowdhery, A., Narang, S., Devlin, J., *et al*. (2022). PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.
- Dziri, N., Kamalloo, E., Mathew, B., *et al*. (2022). FaithDial: A faithful benchmark for information-seeking dialogue. *Findings of ACL 2022*.

- Izacard, G., & Grave, E. (2021). Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.
- Izacard, G., Grusky, M., Joulin, A., & Grave, E. (2022). Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2112.04426*.
- Ji, Z., Lee, N., Frieske, R., *et al*. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), 1–38.
- Khandelwal, U., Fan, A., Jurafsky, D., & Lewis, M. (2022). Generalization through memorization: Nearest neighbor language models. *arXiv preprint arXiv:1911.00172*.
- Kryściński, W., McCann, B., Xiong, C., & Socher, R. (2020). Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1910.12840*.
- Lewis, P., Perez, E., Piktus, A., *et al*. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 9459–9474.
- Ouyang, L., Wu, J., Jiang, X., *et al*. (2022). Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.
- Santhanam, K. G., Khattab, O., Zuccon, G., & Craswell, N. (2022). ColBERTv2: Effective and efficient retrieval via lightweight late interaction. *Proceedings of the 45th International ACM SIGIR Conference*, 941–951.
- Shi, Y., Pang, R., Li, D., *et al*. (2023). Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.
- Shuster, K., Ju, D., Roller, S., *et al*. (2021). Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*.
- Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). FEVER: A large-scale dataset for fact extraction and verification. *NAACL-HLT*, 809–819.
- Touvron, H., Lavril, T., Izacard, G., *et al*. (2023). LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.