International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 5, Sep-Oct-2026, ISSN (Online): 2395-566X

Mitigating Hallucinations in Large Language Models Using Retrieval-Augmented Generation and Knowledge Graphs: A Review

Bankar Akshada Santosh[1] Prof. S. K. Said[2] Prof. M. C. Kolse[3]
Department Of Artificial Intelligence And Data Science Jaihind College Of Engineering Pune, Maharashtra, India;

Abstract—Large Language Models (LLMs) can produce fluent but factually incorrect statements, known as hallucinations, which limits their reliability in healthcare, law, education and other critical domains. This paper reviews methods that reduce hallucination by grounding generation in external knowledge. Retrieval-Augmented Generation (RAG) retrieves relevant passages at query time and conditions the model on them, while Knowledge Graphs (KGs) provide structured, verifiable facts and explicit relations that support multi-hop reasoning. The review explains dense passage retrieval, retrieval-augmented pre-training, graph-based RAG that summarizes communities of related entities, and reasoning that walks over knowledge-graph paths. It also covers evaluation, including truthfulness benchmarks and metrics for faithfulness, context relevance and answer relevance. A hybrid architecture is proposed that combines vector retrieval with knowledge-graph lookup, assembles grounded context, generates answers with citations and verifies claims before responding. Research gaps, objectives, a comparison of grounding strategies and challenges such as retrieval errors, outdated knowledge, graph construction cost and latency are discussed. The review concludes that combining unstructured retrieval with structured knowledge offers a practical route to more accurate and transparent LLM applications.

Index Terms—Large language models, hallucination, retrieval-augmented generation, knowledge graphs, GraphRAG, factuality, dense retrieval, grounding

I. INTRODUCTION

LLMs answer questions, summarize documents and write code, but they sometimes state false facts with complete confidence. Ji *et al.* [1] define hallucination as generated content that is nonsensical or unfaithful to the source or to world knowledge. In a hospital chatbot or a legal assistant, such errors can cause real harm.

A key cause is that an LLM stores knowledge only implicitly in its parameters. That knowledge may be incomplete, outdated or blended incorrectly. Two complementary remedies supply knowledge explicitly at inference time: retrieval of relevant text passages, and lookup in a knowledge graph of verified facts. Pan *et al.* [2] describe a roadmap for unifying LLMs and knowledge graphs, arguing that each compensates for the other's weaknesses.

This paper reviews RAG and KG-based grounding, proposes a hybrid architecture and discusses evaluation and open problems.

II. BACKGROUND AND RELATED WORK

2.1 Types Of Hallucination

Fig. 1 distinguishes intrinsic hallucinations, which contradict the provided source, from extrinsic ones, which add unsupported information, and lists the main mitigation families.

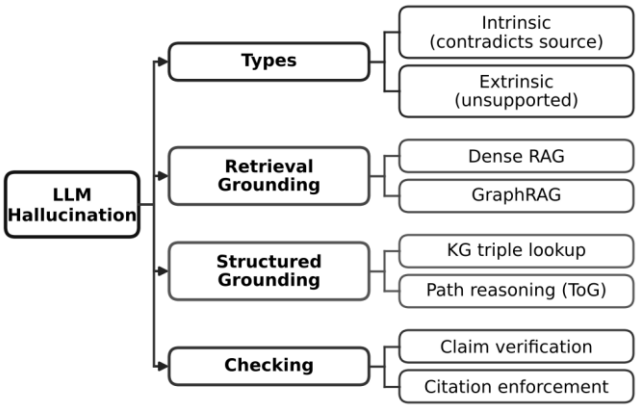


Fig. 1. Hallucination types and mitigation families.

2.2 Knowledge Graphs

A knowledge graph stores facts as (subject, relation, object) triples, such as (Pune, located_in, Maharashtra). Hogan *et al.* [3] describe how knowledge graphs are built, queried and validated. Because each fact is explicit, answers derived from a graph can be traced and checked.

2.3 Review Of Related Studies

Lewis *et al.* [4] introduced RAG, combining a dense retriever with a sequence-to-sequence generator, and showed improved factual accuracy on knowledge-intensive tasks. Karpukhin *et al.* [5] proposed dense passage retrieval with dual encoders, which outperformed keyword search for open-

© 2026 IJSRET

domain question answering. Guu *et al.* [6] incorporated retrieval into language-model pre-training with REALM, so the model learns to use retrieved knowledge.

Gao *et al.* [7] surveyed RAG and distinguished naive, advanced and modular designs that add query rewriting, re-ranking and iterative retrieval. Edge *et al.* [8] proposed GraphRAG, which extracts an entity graph from documents, groups it into communities and summarizes them, improving answers to broad questions that require connecting many sources.

On the knowledge-graph side, Sun *et al.* [9] introduced Think-on-Graph, in which the LLM iteratively explores paths in a knowledge graph to gather evidence for multi-hop reasoning. For evaluation, Lin *et al.* [10] built TruthfulQA to measure whether models repeat common misconceptions, and Es *et al.* [11] proposed RAGAS metrics for faithfulness, answer relevance and context relevance. Table I summarizes these works.

TABLE I Summary of Literature Survey

Ref.	Method	Key Findings	Limitations
[1]	Survey of hallucination in NLG	Defined intrinsic and extrinsic types	Pre-dates many LLM methods
[4]	Retrieval-augmented generation	Better factuality on knowledge tasks	Retriever errors propagate
[5]	Dense passage retrieval	Beats keyword search for QA	Needs training pairs
[6]	Retrieval-augmented pre-training	Model learns to use retrieval	Costly pre-training
[8]	GraphRAG community summaries	Better global sense-making	Expensive graph indexing
[9]	Think-on-Graph path exploration	Multi-hop reasoning over KG	Depends on KG coverage
[2]	Roadmap for LLM–KG unification	Framework for combining both	Conceptual roadmap
[11]	RAGAS evaluation metrics	Measures faithfulness and relevance	Uses LLM as judge

III. RESEARCH GAP AND OBJECTIVES

3.1 Identified Research Gaps

- **G1:** RAG can still hallucinate when retrieved passages are irrelevant, contradictory or ignored by the model.
- **G2:** Building and maintaining domain knowledge graphs is costly and slow.
- **G3:** Few systems combine unstructured retrieval and structured graphs in a principled way.
- **G4:** Evaluation metrics for grounded generation are not standardized.
- **G5:** Added retrieval steps increase latency and cost.

3.2 Objectives Of The Review

- To analyse the causes and types of LLM hallucinations.
- To review RAG techniques and knowledge-graph-based grounding.
- To propose a hybrid RAG + KG architecture with verification.
- To compare grounding strategies and evaluation metrics.
- To discuss challenges and future directions.

3.3 Design Of The Proposed System

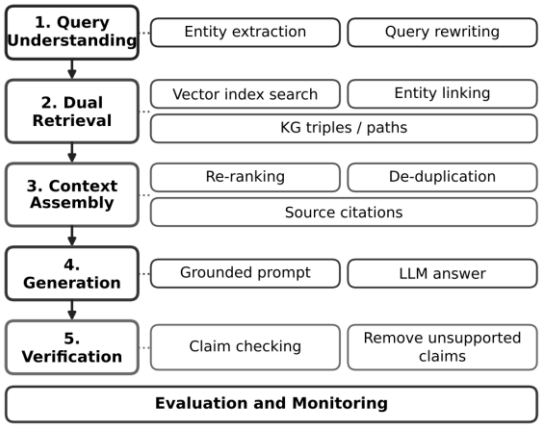


Fig. 2 presents the proposed hybrid grounding architecture.

Fig. 2. Proposed hybrid RAG and knowledge-graph architecture.

3.4 Query Understanding Layer

The user question is analysed to extract entities and intent, and may be rewritten into clearer sub-queries.

3.5 Dual Retrieval Layer

A dense retriever searches a vector index of documents, while entity linking maps detected entities to the knowledge graph and retrieves related triples and paths.

© 2026 IJSRET

3.6 Context Assembly Layer

Passages and triples are re-ranked, de-duplicated and assembled into a prompt that cites each source.

3.7 Generation Layer

The LLM is instructed to answer only from the supplied context and to cite sources, or to say that the answer is not available.

3.8 Verification Layer

Claims in the draft answer are checked against retrieved evidence and graph facts; unsupported claims are removed or flagged.

3.9 Methodology

Fig. 3 shows the processing steps. Dense retrieval scores a passage p for query q by the similarity of their embeddings

$$s(q, p) = E_q(q) \cdot E_p(p) \quad (1)$$

and faithfulness is measured as the fraction of answer claims supported by the retrieved context:

$$Faithfulness = |supported\ claims| / |total\ claims| \quad (2)$$

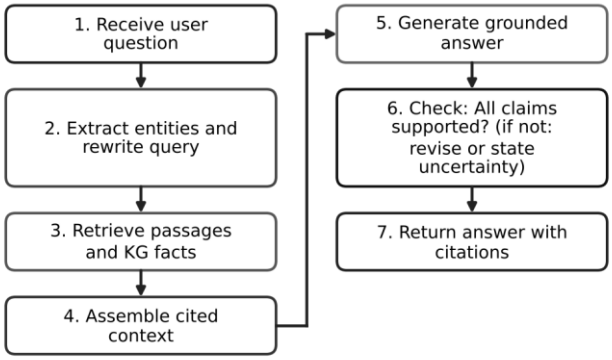


Fig. 3. Hybrid grounding workflow.

IV. DISCUSSION AND COMPARISON

Table II compares grounding strategies. Vector RAG is easy to deploy, GraphRAG handles global questions, KG reasoning gives precise multi-hop facts, and the hybrid approach combines coverage with verifiability.

TABLE II Comparison of Grounding Strategies

Strategy	Strength	Weakness	Latency
No grounding	Simple	Frequent hallucination	Low
Vector RAG	Broad text coverage	Irrelevant retrieval	Medium
GraphRAG	Global summaries	Indexing cost	Medium–High
KG reasoning	Precise, traceable	Limited coverage	Medium
Hybrid + verification	Coverage + checks	Complexity	High

4.1 Challenges And Future Directions

Table III summarizes challenges. Future work includes automatic knowledge-graph construction with LLMs, adaptive retrieval that triggers only when needed, better faithfulness metrics and domain-specific deployments in Indian languages.

TABLE III Challenges and Mitigation

Challenge	Mitigation
Irrelevant retrieval	Re-ranking; query rewriting
Outdated knowledge	Scheduled index and KG updates
KG construction cost	LLM-assisted triple extraction
Ignored context	Citation enforcement; verification
Latency	Caching; adaptive retrieval

V. CONCLUSIONS

Hallucination is one of the main barriers to trustworthy LLM applications. The reviewed literature shows that grounding generation in retrieved text reduces unsupported statements, and that knowledge graphs add structure, traceability and multi-hop reasoning. A hybrid system that retrieves from both sources, cites evidence and verifies claims offers a practical path to more reliable and transparent AI.

5.1 Acknowledgment

The authors express their sincere gratitude to the Department of Artificial Intelligence and Data Science, Jaihind College of Engineering, Pune, for providing the facilities, guidance and encouragement needed to carry out this work. The authors also thank the faculty members and colleagues whose

© 2026 IJSRET

valuable suggestions helped improve the quality of this review.

REFERENCES

1. Z. Ji et al., "Survey of hallucination in natural language generation," ACM Comput. Surv., vol. 55, no. 12, Art. no. 248, 2023.
2. S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu, "Unifying large language models and knowledge graphs: A roadmap," IEEE Trans. Knowl. Data Eng., vol. 36, no. 7, pp. 3580–3599, 2024.
3. A. Hogan et al., "Knowledge graphs," ACM Comput. Surv., vol. 54, no. 4, Art. no. 71, 2021.
4. P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Proc. NeurIPS, vol. 33, 2020, pp. 9459–9474.
5. V. Karpukhin et al., "Dense passage retrieval for open-domain question answering," in Proc. EMNLP, 2020, pp. 6769–6781.
6. K. Guu, K. Lee, Z. Tung, P. Pasupat, and M.-W. Chang, "REALM: Retrieval-augmented language model pre-training," in Proc. ICML, 2020, pp. 3929–3938.
7. Y. Gao et al., "Retrieval-augmented generation for large language models: A survey," arXiv:2312.10997, 2023.
8. D. Edge et al., "From local to global: A graph RAG approach to query-focused summarization," arXiv:2404.16130, 2024.
9. J. Sun et al., "Think-on-Graph: Deep and responsible reasoning of large language model on knowledge graph," in Proc. ICLR, 2024.
10. S. Lin, J. Hilton, and O. Evans, "TruthfulQA: Measuring how models mimic human falsehoods," in Proc. 60th Annu. Meeting ACL, 2022, pp. 3214–3252.
11. S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAS: Automated evaluation of retrieval augmented generation," in Proc. EACL System Demonstrations, 2024, pp. 150–158.

© 2026 IJSRET