The Correlation Between Retrieval Precision And Hallucination Rates In Rag-End2End Vary Across Different Domains (E.G.,

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does the correlation between retrieval precision and hallucination rates in RAG-end2end vary across different domains (e.g., healthcare vs. news) when comparing 7B and 70B generator models. 16 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: MultiHop-RAG: Benchmarking Retrieval-Augmented Generation for Multi-Hop Queries. Research question: How does the correlation between retrieval precision and hallucination rates in RAG-end2end vary across different domains (e.g., healthcare vs. news) when comparing 7B and 70B generator models?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.5/10.

3 Results

14 papers retrieved. 16 claims extracted; 1 independently verified. Quality review score: 4.5/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
The quality of the retrieval set Rq determines the final generation quality in an RAG system handling multi-hop queries.	×	0.12
Retrieval evaluation metrics used include Mean Average Precision at K (MAP@K), Mean Reciprocal Rank at K (MRR@K), and Hi	×	0.03
MAP@K measures the average top-K retrieval precision across all queries.	×	0.03
MRR@K calculates the average of the reciprocal ranks of the first relevant chunk for each query, considering the top-K r	×	0.02
Hit@K metric measures the fraction of evidence that appears in the top-K retrieved set.	×	0.03
The MultiHop-RAG dataset is constructed using news articles from various English-language websites covering categories l	×	0.06
The news articles in the MultiHop-RAG dataset were published from September 26, 2023, to December 26, 2023.	×	0.05
The selected timeframe for the news articles extends beyond the knowledge cutoff of some widely-used LLMs, including Cha	×	0.04
Only articles with a token length greater than or equal to 1,024 are kept in the MultiHop-RAG dataset.	×	0.07
Each news article in the MultiHop-RAG dataset is paired with metadata, including the title, publish date, author, catego	×	0.08
Factual or opinion sentences are extracted from each article using a trained language model for use as evidence in answe	×	0.14
Only news articles containing evidence with overlapping keywords with other news articles are retained in the MultiHop-R	×	0.08
RAG improves LLM's response and mitigates the occurrence of hallucinations, thereby enhancing the models' credibility.	×	0.07
LLM-based frameworks like LlamaIndex and LangChain specialize in supporting RAG pipelines.	×	0.04
Multi-hop queries in RAG applications require retrieving and reasoning over evidence from multiple documents.	✓	0.21
Traditional similarity matching methods like cosine similarity may not be sufficient for multi-hop queries involving inf	×	0.09

References

- http://arxiv.org/abs/2401.15391v1
- http://arxiv.org/abs/2502.11228v2
- http://arxiv.org/abs/2603.01710v1