Scale-Dependent Correlation Between Retrieval Precision and Hallucination in RAG Pipelines

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: How does the scale of the generator model (7B vs. 70B) influence the correlation between retrieval precision and downstream hallucination rates in sensitive domain RAG pipelines. 11 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering. Research question: How does the scale of the generator model (7B vs. 70B) influence the correlation between retrieval precision and downstream hallucination rates in sensitive domain RAG pipelines?.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.5/10.

3 Results

11 papers retrieved. 11 claims extracted; 0 independently verified. Quality review score: 3.5/10.

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
Most ODQA datasets like Natural Questions, TriviaQA, WebQuestions, and CuratedTrec are answered with Wikipedia-based kno	×	0.06
Neural retrievers like DPR are already trained with Wikipedia-based datasets.	×	0.07
Three domain-specific datasets were selected for the experiment: COVID-19 QA, News QA, and Conversation QA.	×	0.11
The COVID-19 QA domain knowledge base was created with 250,000 100-word passages extracted from 5,000 full-text scientif	×	0.07
RAG-end2end outperforms RAG-original even in other Wikipedia-based datasets.	×	0.12
RAG-end2end updates the context encoder and embeddings during the training process.	×	0.08
The retriever component is crucial in domain-specific question answering.	×	0.13
Future research directions include exploring RAG-end2end on tasks like Fact Checking, Summarisation, and conversational	×	0.04
Exploring generative capabilities with qualitative metrics could improve factual consistency and reduce hallucinations i	×	0.03
Updating the retriever and document embeddings during the training phase could improve factual consistency and reduce ha	×	0.06
The statement reconstruction signal acts as a good auxiliary signal for improving the overall performance of RAG models.	×	0.09

References

- http://arxiv.org/abs/2404.07220v2
- http://arxiv.org/abs/2504.01346v4
- http://arxiv.org/abs/2210.02627v1