Semantic Overlap and Hallucination Rates in Retrieval-Augmented Generation Systems

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: To what extent does varying the level of semantic overlap in retrieved documents affect the hallucination rates of large language models in retrieval-augmented generation settings. Retrieval-augmented generation (RAG) enhances large language models (LLMs) for domain-specific question-answering (QA) tasks by leveraging external knowledge sources. However, traditional RAG systems primarily focus on relevance-based retrieval and often struggle with. 8 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Vendi-RAG: Adaptively Trading-Off Diversity And Quality Significantly Improves Retrieval Augmented Generation With LLMs. Research question: To what extent does varying the level of semantic overlap in retrieved documents affect the hallucination rates of large language models in retrieval-augmented generation settings?.

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 5.2/10.

3 Results

14 papers retrieved. 8 claims extracted; 2 independently verified. Quality review score: 5.2/10.

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
Vendi-RAG was evaluated on three multi-hop QA benchmark datasets: MuSiQue, HotpotQA, and 2WikiMultiHopQA.	✓	0.23
The sensitivity analysis of the VSR process was conducted using 100 randomly sampled queries from the dataset.	×	0.02
The sensitivity analysis evaluated the retrieval pipeline across multiple s values ranging from 0.0 to 1.0.	×	0.03
Setting s = 0.0 serves as a baseline representing a pure similarity search scenario.	×	0.05
Kendall's τ and Spearman's ρ were used to quantify deviations from the baseline in the sensitivity analysis.	×	0.05
As s increases from 0.0 to 1.0, both Kendall's τ and Spearman's ρ decrease progressively.	×	0.03
Vendi-RAG uses a retrieval approach based on the Vendi Score (VS) to quantify semantic diversity in a set of documents.	✓	0.20
The Vendi Score (VSk(D)) reflects the effective number of unique documents in D, attaining its maximum value n when all	×	0.06

References

- http://arxiv.org/abs/2409.03708v2
- http://arxiv.org/abs/2503.16581v1
- http://arxiv.org/abs/2502.11228v2