Dimensionality Reduction Effects on Multi-Hop RAG Performance in NaturalQuestions

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: How does reducing vector embedding dimensionality from 1024 to 256 impact Exact Match and F1 scores on the NaturalQuestions benchmark for multi-hop RAG systems compared to single-hop queries. 6 claims were extracted from source literature; 6 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 9.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Superposition Prompting: Improving and Accelerating Retrieval-Augmented Generation. Research question: How does reducing vector embedding dimensionality from 1024 to 256 impact Exact Match and F1 scores on the NaturalQuestions benchmark for multi-hop RAG systems compared to single-hop queries?.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 9.0/10.

3 Results

4 papers retrieved. 6 claims extracted; 6 independently verified. Quality review score: 9.0/10.

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
Large language models (LLMs) exhibit significant drawbacks when processing long contexts, including quadratic scaling of	✓	0.31
Superposition prompting is a novel RAG prompting methodology that can be directly applied to pre-trained transformer-bas	✓	0.32
Superposition prompting allows the LLM to process input documents in parallel prompt paths, discarding paths once they a	✓	0.32
Superposition prompting enhances time efficiency across a variety of question-answering benchmarks using multiple pre-tr	✓	0.29
Superposition prompting significantly improves accuracy when the retrieved context is large relative to the context the	✓	0.29
Superposition prompting facilitates a 93x reduction in compute time while improving accuracy by 43% on the NaturalQuesti	✓	0.38

References

- https://doi.org/10.48550/arxiv.2404.06910
- https://doi.org/10.48550/arxiv.2406.18676
- https://doi.org/10.18653/v1/2025.emnlp-main.317