Retrieval-Augmented Generation Performance Across Vector Space Dimensionalities on NaturalQuestions

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does the retrieval-augmented generation (RAG) performance on the NaturalQuestions benchmark vary when using different vector space dimensionalities (e.g., 256 vs 1024) for the RGAR system. 18 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. 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 retrieval-augmented generation (RAG) performance on the NaturalQuestions benchmark vary when using different vector space dimensionalities (e.g., 256 vs 1024) for the RGAR system compared to standard RAG, measured by exact match (EM) and F1 scores?.

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 4.2/10.

3 Results

14 papers retrieved. 18 claims extracted; 1 independently verified. Quality review score: 4.2/10.

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
An RAG system handling multi-hop queries is assessed based on retrieval evaluation and generation evaluation.	×	0.14
Retrieval evaluation compares the retrieved set with ground truth evidence, excluding null queries.	×	0.08
Retrieval evaluation metrics used include Mean Average Precision at K (MAP@K), Mean Reciprocal Rank at K (MRR@K), and Hi	×	0.07
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 within the top-K retriev	×	0.02
Hit@K measures the fraction of evidence that appears in the top-K retrieved set.	×	0.02
Response evaluation assesses the LLM's reasoning capability by comparing the LLM response with the ground truth answer.	×	0.07
The MultiHop-RAG dataset was constructed using news articles collected via the mediastack API.	×	0.10
The news data source comprises English-language websites covering entertainment, business, sports, technology, health, a	×	0.02
News articles selected for the dataset were published between September 26, 2023, and December 26, 2023.	×	0.02
The selected publication timeframe extends beyond the knowledge cutoff of ChatGPT and LLaMA as of the time of writing.	×	0.03
Only news articles with a token length greater than or equal to 1,024 were retained for the dataset.	×	0.04
Each news article in the dataset is paired with metadata including title, publish date, author, category, URL, and news	×	0.05
Factual or opinion sentences were extracted from each article using a trained language model to serve as evidence.	×	0.03
Articles were retained only if they contained evidence with overlapping keywords with other news articles to facilitate	×	0.07
RAG improves LLM responses and mitigates the occurrence of hallucinations.	×	0.04
LlamaIndex and LangChain are LLM-based frameworks that specialize in supporting RAG pipelines.	×	0.06
Multi-hop queries require retrieving and reasoning over evidence from multiple documents.	✓	0.26

References

- http://arxiv.org/abs/2401.15391v1
- http://arxiv.org/abs/2106.16020v1
- http://arxiv.org/abs/2506.06962v3