Retrieval-Augmented Generation Performance Under Adversarial Perturbations in Domain-Specific Benchmarks

Assignee Research

June 1, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: What is the comparative performance of retrieval-augmented generation (Vendi-RAG vs. DPR) when evaluated on domain-specific benchmarks (e.g., QuranQA) under adversarial perturbations (e.g., synonym. Retrieval-augmented generation (RAG) enhances large language models (LLMs) with external knowledge to answer questions more accurately. However, research on evaluating RAG systems-particularly the retriever component-remains limited, as most existing work focuses on. 12 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Evaluating Multi-Hop Reasoning in RAG Systems: A Comparison of LLM-Based Retriever Evaluation Strategies. Research question: What is the comparative performance of retrieval-augmented generation (Vendi-RAG vs. DPR) when evaluated on domain-specific benchmarks (e.g., QuranQA) under adversarial perturbations (e.g., synonym substitutions, typos), measured by exact match accuracy and BLEU scores?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.7/10.

3 Results

10 papers retrieved. 12 claims extracted; 3 independently verified. Quality review score: 4.7/10.

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
CARE consistently outperforms existing methods for evaluating multi-hop reasoning in RAG systems.	✓	0.38
The performance gains of CARE are most pronounced in models with larger parameter counts and longer context windows.	✓	0.23
Single-hop queries show minimal sensitivity to context-aware evaluation.	✓	0.30
The indirect evaluation approach derived from the eRAG method [24] involves comparing generated answers with ground trut	×	0.02
The direct evaluation method based on the ARES framework [23] involves determining if a context is crucial to answering	×	0.04
The CARE method involves determining if a context is crucial to answering a question with a ground truth answer, given a	×	0.06
In the indirect method, if the generated answer is empty, the context is labeled as non-relevant.	×	0.03
CARE achieved an accuracy of 0.827$\pm$0.02, F1-Score of 0.814$\pm$0.02, recall of 0.757$\pm$0.04, and precision of 0.880$\pm$0.03 on th	×	0.03
CARE achieved an accuracy of 0.755$\pm$0.02, F1-Score of 0.678$\pm$0.03, recall of 0.517$\pm$0.04, and precision of 0.987$\pm$0.01 on th	×	0.03
The indirect approach led to a significant improvement in F1-Score for the small LLaMa model.	×	0.03
The direct approach resulted in a decline in F1-Score for the reasoning model o4-mini.	×	0.03
CARE consistently outperformed other approaches across all models except for the LLaMa 3.1-8b model.	×	0.05

References

- http://arxiv.org/abs/2604.18234v1

- http://arxiv.org/abs/2510.25518v1
- http://arxiv.org/abs/2502.11228v2