Adaptive Retriever Selection Strategies for Hallucination Reduction in Long-Context QA

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: Can adaptive retriever selection strategies reduce hallucination rates in long-context QA tasks evaluated on the NarrativeQA dataset. 10 claims were extracted from source literature; 0 were 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: Efficient Context Selection for Long-Context QA: No Tuning, No Iteration, Just Adaptive-k. Research question: Can adaptive retriever selection strategies reduce hallucination rates in long-context QA tasks evaluated on the NarrativeQA dataset?.

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. 10 claims extracted; 0 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
The experimental settings include datasets HotpotQA, Natural Questions (NQ), and TriviaQA, curated by HELMET for long-co	×	0.05
Due to high computational cost, evaluation is conducted on a subset of 100 examples per dataset.	×	0.03
HoloBench is used for aggregation tasks, providing 90 evaluation samples with a fixed total context of 100k tokens and v	×	0.05
Three embedding models are tested: Meta's contriver-msmarco2 (109M params), BAAI's bge-en-large-v1.5 (335M params), and	×	0.02
Five closed and open models are used as readers: GPT-4o-mini, GPT-4o, Gemini-2.5-Flash, Llama4-Scout, and Llama4-Maveric	×	0.05
Adaptive-k retrieval is compared against zero-shot LLMs, LLMs with full context, SELF-ROUTE, and fixed-n retrieval metho	×	0.13
The best-performing fixed-n setting is regarded as the oracle for comparison with adaptive-k.	×	0.04
Context recall is used as a metric to evaluate retrieval performance.	×	0.06
The University of California, Los Angeles (UCLA) has a student population exceeding 45,000.	×	0.00
San Diego State University and California State University, Fullerton each enroll more than 30,000 students annually.	×	0.01

References

- `http://arxiv.org/abs/2506.08479v3`

- http://arxiv.org/abs/2404.14464v1
- http://arxiv.org/abs/2604.18234v1