Adaptive Retriever Selection Latency Overhead in Long-Context Benchmarks

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: What is the latency overhead of adaptive retriever selection strategies compared to static retrieval methods on long-context benchmarks like HotpotQA. 17 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: When to Retrieve During Reasoning: Adaptive Retrieval for Large Reasoning Models. Research question: What is the latency overhead of adaptive retriever selection strategies compared to static retrieval methods on long-context benchmarks like HotpotQA?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 2.8/10.

3 Results

13 papers retrieved. 17 claims extracted; 0 independently verified. Quality review score: 2.8/10.

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
ReaLM-Retrieve achieves 71.2% F1 on the MuSiQue dataset.	×	0.15
ReaLM-Retrieve uses an average of 1.8 retrieval calls on the MuSiQue dataset.	×	0.14
IRCoT achieves 65.4% F1 on the MuSiQue dataset.	×	0.04
IRCoT uses an average of 3.4 retrieval calls on the MuSiQue dataset.	×	0.12
The improvement of ReaLM-Retrieve over IRCoT on MuSiQue is statistically significant at $p < 0.01$.	×	0.10
ReaLM-Retrieve achieves a +5.8 F1 point improvement over IRCoT on the HotpotQA dataset.	×	0.09
ReaLM-Retrieve achieves a +3.9 F1 point improvement over IRCoT on the 2WikiMultiHopQA dataset.	×	0.09
ReaLM-Retrieve has a token cost per query of 9,489 on the evaluated benchmarks.	×	0.05
ReaLM-Retrieve achieves a 16% cost savings compared to Single RAG.	×	0.08
ReaLM-Retrieve has an average latency of 14.1 seconds per query.	×	0.05
ReaLM-Retrieve achieves an F1/Call ratio of 39.6.	×	0.07
Removing the verbal uncertainty component (w/o Verb.) from ReaLM-Retrieve results in an F1 score of 68.4% on MuSiQue.	×	0.05
Replacing the RSUS policy with a random policy reduces the F1 score on MuSiQue by 8.5 points compared to the full ReaLM-	×	0.07
IRCoT retrieves after each sentence during chain-of-thought reasoning.	×	0.07
FLARE triggers retrieval when token probability falls below a threshold.	×	0.02
Self-RAG requires full model fine-tuning.	×	0.09
Multi-vector retrieval engines such as PLAID and WARP achieve 3–41$\times$ speedups.	×	0.03

References

- http://arxiv.org/abs/2404.14464v1
- http://arxiv.org/abs/2604.26649v1
- http://arxiv.org/abs/2604.18234v1