Scaling Context Windows in Static and Dynamic RAG Systems: Latency, Throughput, and Accuracy Trade-offs in Sub-10B Models

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How does the scaling of context window size in static vs. dynamic RAG systems affect inference latency and throughput for sub-10B models on domain-specific benchmarks like SIMCOPILOTJ, while. 7 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Investigating Retrieval-Augmented Generation in Quranic Studies: A Study of 13 Open-Source Large Language Models. Research question: How does the scaling of context window size in static vs. dynamic RAG systems affect inference latency and throughput for sub-10B models on domain-specific benchmarks like SIMCOPILOTJ, while maintaining a fixed pass@1 accuracy threshold?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.8/10.

3 Results

12 papers retrieved. 7 claims extracted; 0 independently verified. Quality review score: 3.8/10.

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
The system employs a Retrieval-Augmented Generation (RAG) architecture, combining retrieval-based and generative methodo	×	0.11
The system executes semantic search and retrieval, response generation, and citations and contextualization tasks.	×	0.03
Context Relevance is evaluated using the precision@k metric, where k represents the number of top retrieved results cons	×	0.07
The dataset was chosen according to specific criteria: Authenticity, Descriptive Richness, Clarity and Accessibility, an	×	0.04
The dataset underwent a thorough review to confirm its compliance with recognized Islamic scholarship and the absence of	×	0.01
The dataset must deliver comprehensive, contextually rich descriptions that can be effectively employed for semantic sea	×	0.04
The content needed to be created in a structured and clear manner, facilitating both manual review and computational pro	×	0.01

References

- http://arxiv.org/abs/2601.08844v1
- http://arxiv.org/abs/2503.16581v1
- http://arxiv.org/abs/2506.06962v3