Context Window Size Reduction Effects on Llama-3-8B Throughput in Retrieval-Augmented SQuAD 2.0 Generation

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: What is the impact of context window size reduction (4096 to 1024 tokens) on the throughput of Llama-3-8B when performing retrieval augmented generation on SQuAD 2.0. Retrieval-Augmented Generation (RAG) has emerged as a powerful paradigm to enhance large language models (LLMs) by conditioning generation on external evidence retrieved at inference time. While RAG addresses critical limitations of parametric knowledge storage-such as factual. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Retrieval-Augmented Generation: A Comprehensive Survey of Architectures, Enhancements, and Robustness Frontiers. Research question: What is the impact of context window size reduction (4096 to 1024 tokens) on the throughput of Llama-3-8B when performing retrieval augmented generation on SQuAD 2.0?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.5/10.

3 Results

15 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 6.5/10.

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.

References

- https://arxiv.org/abs/2504.14891
- https://arxiv.org/abs/2501.06713
- https://arxiv.org/abs/2506.00054