Retrieval-Augmented Generation and Chain-of-Thought Prompting Latency in Open-Weight LLMs for Time-Series Data

Assignee Research

June 4, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: What is the impact of retrieval-augmented generation versus chain-of-thought prompting on the inference latency and token throughput of open-weight LLMs when processing high-frequency time-series data. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Riddle Me This! Stealthy Membership Inference for Retrieval-Augmented Generation. Research question: What is the impact of retrieval-augmented generation versus chain-of-thought prompting on the inference latency and token throughput of open-weight LLMs when processing high-frequency time-series data?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.3/10.

3 Results

16 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 5.3/10.

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

- http://arxiv.org/abs/2305.01505v2
- http://arxiv.org/abs/2411.18583v1
- http://arxiv.org/abs/2502.00306v2