

Retrieval-Augmented Generation Latency-Accuracy Trade-offs in Llama3-70B and Gemini 1.5 Pro on CodeXGLUE Security

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 9 peer-reviewed papers addressing the following research question: What is the trade-off between inference latency and accuracy when using retrieval-augmented generation for Llama3-70B versus Gemini 1.5 Pro on the CodeXGLUE security subset under few-shot learning. The advent of large language models (LLMs) has significantly advanced the field of code translation, enabling automated translation between programming languages. However, these models often struggle with complex translation tasks due to inadequate contextual understanding. 0 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: Enhancing Code Translation in Language Models with Few-Shot Learning via Retrieval-Augmented Generation. Research question: What is the trade-off between inference latency and accuracy when using retrieval-augmented generation for Llama3-70B versus Gemini 1.5 Pro on the CodeXGLUE security subset under few-shot learning settings?.

2 Methodology

Systematic literature search across multiple databases yielded 9 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

9 papers retrieved. 0 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.

References

- <http://arxiv.org/abs/2402.12317v2>
- <http://arxiv.org/abs/2206.00092v1>
- <http://arxiv.org/abs/2407.19619v1>