How does the inference efficiency (throughput, latency) of SecLM-fine-tuned Llama3, Codestral, and Deepseek R1

Assignee Research

May 29, 2026

Abstract

Large language models (LLMs) such as GPT-4o and Claude Sonnet 4.5 have demonstrated strong capabilities in open-ended reasoning and generative language tasks, leading to their widespread adoption across a broad range of NLP applications. However, for structured text classification problems with fixed label spaces, model selection is often driven by predictive performance alone, overlooking operational constraints encountered in production systems. In this work, we present a systematic comparison of two contrasting paradigms for text classification: zero- and few-shot prompt-based large langu

1 Introduction

This paper examines: Cost-Aware Model Selection for Text Classification: Multi-Objective Trade-offs Between Fine-Tuned Encoders and LLM Prompting in Production. Research question: How does the inference efficiency (throughput, latency) of SecLM-fine-tuned Llama3, Codestral, and Deepseek R1 vary across different programming languages (Python, Java, C/C++) when deployed on edge devices with limited compute resources?.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.1/10.

3 Results

4 papers retrieved. 9 claims extracted; 1 independently verified. Quality review score: 5.1/10.

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
In production environments, engineering teams repeatedly face similar decision points—whether to rely on hosted LLM APIs	×	0.07
A rigorously designed benchmark can be viewed as a persistent knowledge artifact.	×	0.02
Performance metrics are analyzed jointly with inference latency and monetary cost through Pareto frontier projections an	✓	0.17
The released artifacts are intended to function as living reference points that facilitate sustainable system evolution,	×	0.03
We consider three primary operational constraints: latency budgets, throughput requirements, and budget constraints.	×	0.04
In production-grade NLP systems, model selection is rarely a single-objective optimization problem driven solely by pred	×	0.10
A model that offers marginal gains in F1 score may be unsuitable if it introduces unstable latency profiles, opaque infe	×	0.03
The objective of this work is to compare encoder-based architectures and prompt-based LLM approaches, and to construct a	×	0.13
By jointly quantifying predictive quality, inference latency, and economic cost across representative datasets, we provi	×	0.08

References

- http://arxiv.org/abs/2602.06370v1
- http://arxiv.org/abs/2510.22531v1

- `http://arxiv.org/abs/2308.10783v2`