Prompt Engineering Strategies and Throughput-Cost Trade-offs in LLM-Based Text Classification

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: What is the impact of prompt engineering strategies on the throughput-cost trade-off for LLM-based text classification in production environments. 13 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

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: What is the impact of prompt engineering strategies on the throughput-cost trade-off for LLM-based text classification in production environments?.

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 4.5/10.

3 Results

16 papers retrieved. 13 claims extracted; 1 independently verified. Quality review score: 4.5/10.

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.

5 Extracted Claims

Claim	Verified	Confidence
Engineering teams in production environments face decision points between relying on hosted LLM APIs, deploying in-house	×	0.07
Model selection choices in production are often made under uncertainty guided by anecdotal evidence or partial assessmen	×	0.06
The study structures evaluation around deployment-relevant decision variables rather than accuracy in isolation.	×	0.03
Performance metrics in the study are analyzed jointly with inference latency and monetary cost using Pareto frontier pro	✓	0.18
The study's released artifacts allow pricing assumptions to be updated, hardware profiles to be substituted, and experim	×	0.02
The methodology grounds model evaluation in three primary operational constraints: latency budgets, throughput requireme	×	0.06
Latency budgets are defined as end-to-end per-request service-level objectives that bound acceptable response times.	×	0.02
Throughput requirements are expressed in requests per second or samples processed per day.	×	0.03
Budget constraints capture the recurring monetary cost of inference.	×	0.05
In production-grade NLP systems, model selection is rarely a single-objective optimization problem driven solely by pred	×	0.10
Model selection constitutes a knowledge-based decision process where empirical performance is evaluated alongside system	×	0.06
A model offering marginal gains in F1 score may be unsuitable if it introduces unstable latency profiles, opaque inferen	×	0.03
The work jointly quantifies predictive quality, inference latency, and economic cost across representative datasets.	×	0.07

References

- http://arxiv.org/abs/2404.08668v3
- http://arxiv.org/abs/2605.07422v1
- http://arxiv.org/abs/2602.06370v1