Zero-shot vs. Fine-tuned LLMs in Political Ideology Classification Across Textual Corpora

Assignee Research

June 11, 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 do zero-shot LLMs compare to fine-tuned models in accurately classifying political ideology across diverse textual corpora (e.g., news articles, social media) when evaluated on benchmarks like MNLI or HANS?.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.2/10.

3 Results

11 papers retrieved. 7 claims extracted; 5 independently verified. Quality review score: 7.2/10.

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.

5 Extracted Claims

Claim	Verified	Confidence
The study analyzes performance metrics jointly with inference latency and monetary cost through Pareto frontier projecti	✓	0.21
The released artifacts are intended to function as living reference points that facilitate sustainable system evolution.	✓	0.23
The study considers three primary operational constraints: latency budgets, throughput requirements, and budget constrai	×	0.12
Model selection in production-grade NLP systems is a knowledge-based decision process that evaluates empirical performan	✓	0.21
Benchmark results serve as reusable decision evidence, not merely as leaderboard rankings.	✓	0.18
The objective of the work is to construct an empirical knowledge base that supports principled model selection under rea	✓	0.21
The study jointly quantifies predictive quality, inference latency, and economic cost across representative datasets.	×	0.10

References

- http://arxiv.org/abs/2410.15308v2
- http://arxiv.org/abs/2602.06370v1
- http://arxiv.org/abs/2308.10783v2