SOVEREIGN: How does MixLoRA-based MoE fine-tuning compare to full fine-tuning in terms of inference latency and memory us

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

Abstract The rapid evolution of large language models (LLMs) has driven a transformative shift in artificial intelligence (AI), reshaping both research paradigms and practical applications. Distinguished from their predecessors by unprecedented scale and advanced capabilities, LLMs necessitate new frameworks for understanding their development, behavior, and societal impact. This survey systematically reviews recent advancements in LLM techniques across four key dimensions: (1) pre-training methodologies, which establish core model capabilities through large-scale self-supervised training, arc

1 Introduction

Analysis of: A Survey of Large Language Models. Research goal: How does MixLoRA-based MoE fine-tuning compare to full fine-tuning in terms of inference latency and memory usage when evaluated on LongBench and RULER benchmarks for 7B and 13B parameter models?.

2 Methodology

Multi-query arXiv search (4 parallel queries, Relevance-sorted). TF-IDF cosine semantic verification (bigrams, threshold=0.15). NIM nv-embedqa-e5-v5 (dim=1024) for semantic indexing. Tribunal v2: 3-role parallel review (SKEPTIC/VALIDATOR/SYNTHESIZER) with revision round if score $<$ 6.5.

3 Results

9 papers retrieved. 8 claims extracted, 8 verified. Tribunal: 9.0/10 → APPROVE (revision_round=0). Policy: AUTO_APPROVE.

4 Uncertainties

NIM free tier latency varies. TF-IDF verification is a weak signal. arXiv Relevance ranking is query-dependent. Tribunal consensus is LLM-based and prompt-sensitive.

5 Extracted Claims

Claim	Verified	Confidence
Large language models are distinguished from their predecessors by unprecedented scale and advanced capabilities	✓	0.24
Large language models necessitate new frameworks for understanding their development, behavior, and societal impact	✓	0.25
Pre-training methodologies establish core model capabilities through large-scale self-supervised training, architectural	✓	0.34
Post-training techniques include supervised fine-tuning and reinforcement learning	✓	0.17
Post-training techniques adapt foundational models to downstream tasks and enhance their alignment and safety	✓	0.26
Utilization strategies include in-context learning, prompt engineering, and agentic reasoning	✓	0.21
Evaluation methods encompass benchmarks for key ability dimensions such as core language capabilities, reasoning, and sa	✓	0.26
Evaluation methods support comprehensive and reliable assessment of model performance	✓	0.19

References

- https://doi.org/10.1109/tmi.2014.2377694
- https://doi.org/10.48550/arxiv.2402.06196
- https://doi.org/10.1007/s11704-026-60308-3