SOVEREIGN: How does ExpertFlow's inference efficiency (measured in tokens per second and GPU memory usage) on multimodal

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 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 ExpertFlow's inference efficiency (measured in tokens per second and GPU memory usage) on multimodal MoE models like MoE-CLIP or VL-MoE scale with increasing numbers of experts (e.g., 8 to 64) on standard VQA benchmarks, and what is the resulting accuracy-throughput Pareto frontier?.

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

11 papers retrieved. 0 claims extracted, 0 verified. Tribunal: 4.0/10 → REJECT (revision_round=0). Policy: ESCALATE_TO_OWNER.

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.

References

- https://doi.org/10.1093/cercor/bhx179
- https://doi.org/10.48550/arxiv.1712.09923
- https://doi.org/10.1007/s11704-026-60308-3