SOVEREIGN: How does the scaling efficiency of soft modality-guided routing in SMoES compare to dense and hard MoE baselin

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 the scaling efficiency of soft modality-guided routing in SMoES compare to dense and hard MoE baselines when scaling from 7B to 13B parameters on multimodal reasoning benchmarks (e.g., MMMU, MathVista), measured by accuracy-per-parameter and FLOPs per inference step?.

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

6 papers retrieved. 7 claims extracted, 7 verified. Tribunal: 7.5/10 → 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 (LLMs) have driven a transformative shift in artificial intelligence (AI), reshaping both research	✓	0.32
LLMs are distinguished from their predecessors by unprecedented scale and advanced capabilities.	✓	0.20
Pre-training methodologies establish core model capabilities through large-scale self-supervised training, architectural	✓	0.36
Post-training techniques include supervised fine-tuning and reinforcement learning, which adapt foundational models to d	✓	0.32
Utilization strategies include in-context learning, prompt engineering, and agentic reasoning, that optimize real-world	✓	0.33
Evaluation methods encompass benchmarks for key ability dimensions such as core language capabilities, reasoning, and sa	✓	0.34
Critical research issues include those concerning theoretical foundations, efficient scaling, alignment, and agentic cap	✓	0.25

References

- https://doi.org/10.48550/arxiv.2403.14608
- https://doi.org/10.48550/arxiv.2307.06435
- https://doi.org/10.1007/s11704-026-60308-3