SOVEREIGN: How does the inference throughput (tokens per second) of Uni-MoE-2.0-Omni's dynamic-capacity MoE compare to de

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

We present Uni-MoE 2.0 from the Lychee family. As a fully open-source omnimodal large model (OLM), it substantially advances Lychee's Uni-MoE series in language-centric multimodal understanding, reasoning, and generating. Based on the dense LLM, we build Uni-MoE-2.0-Omni from scratch through three core contributions: dynamic-capacity Mixture-of-Experts (MoE) design, a progressive training strategy enhanced with an iterative reinforcement strategy, and a carefully curated multimodal data matching technique. It is capable of omni-modal understanding, as well as generating images, text, and speech

1 Introduction

Analysis of: Uni-MoE-2.0-Omni: Scaling Language-Centric Omnimodal Large Model with Advanced MoE, Training and Data. Research goal: How does the inference throughput (tokens per second) of Uni-MoE-2.0-Omni's dynamic-capacity MoE compare to dense baselines and fixed-top-k hard-routed MoE on multimodal QA benchmarks (e.g., MMBench, MMMU) when varying the number of active experts from 2 to 8 under batch sizes of 1, 16, and 64?.

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. 0 claims extracted, 0 verified. Tribunal: 4.3/10 → REVISE (revision_round=1). 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

- http://arxiv.org/abs/2410.17954v2
- http://arxiv.org/abs/2604.23996v1
- http://arxiv.org/abs/2511.12609v2