SOVEREIGN: Uni-MoE-2.0-Omni: Scaling Language-Centric Omnimodal Large Model with Advanced M

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 27, 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 omnimodal 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 computational efficiency (FLOPs per forward pass) of SMoES scale with the number of active experts and input modality composition relative to dense and hard-routed MoE baselines across different batch sizes on multimodal QA tasks?.

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

14 papers retrieved. 10 claims extracted, 8 verified. Tribunal: 7.5/10 → APPROVE (revision_round=1). 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
Uni-MoE-2.0-Omni is a fully open-source omni-modal large model.	✓	0.31
The model is built from scratch based on a dense LLM.	×	0.09
The model uses a dynamic-capacity Mixture-of-Experts (MoE) design.	×	0.13
The training strategy includes a progressive supervised fine-tuning strategy enhanced with an iterative reinforcement st	✓	0.24
The model is capable of omnimodal understanding and generating images, text, and speech.	✓	0.22
The MoE framework uses shared, routed, and null experts for 10 cross-modal inputs.	✓	0.21
The Omni-Modality 3D RoPE ensures spatio-temporal cross-modality alignment in the self-attention layer.	✓	0.28
The base model was trained on approximately 75B tokens of open-source multimodal data.	✓	0.27
The model is equipped with special speech and image generation tokens.	✓	0.17
The model achieves SOTA or highly competitive performance across 85 benchmarks.	✓	0.17

References

- http://arxiv.org/abs/2605.15484v1
- https://www.semanticscholar.org/paper/3c75c242a8a2d15448d0911da4349791f224a776

- https://www.semanticscholar.org/paper/0a945649b37e4970f56a66b76204f6fb0e6de590