SOVEREIGN: What is the comparative performance drop in Video-MME accuracy for MoE models (e.g., Mixtral 8x22B) versus den

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

We introduce phi-3-mini, a 3.8 billion parameter language model trained on 3.3 trillion tokens, whose overall performance, as measured by both academic benchmarks and internal testing, rivals that of models such as Mixtral 8x7B and GPT-3.5 (e.g., phi-3-mini achieves 69% on MMLU and 8.38 on MT-bench), despite being small enough to be deployed on a phone. Our training dataset is a scaled-up version of the one used for phi-2, composed of heavily filtered publicly available web data and synthetic data. The model is also further aligned for robustness, safety, and chat format. We also provide param

1 Introduction

Analysis of: Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. Research goal: What is the comparative performance drop in Video-MME accuracy for MoE models (e.g., Mixtral 8x22B) versus dense models (e.g., Llama 3 70B) as video context length increases from 128K to 10M tokens across different temporal granularities?.

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: 1.7/10 $\rightarrow$ 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/10.48550/arxiv.2404.14219
- https://doi.org/10.48550/arxiv.2307.06435
- https://doi.org/10.1007/s10462-023-10466-8