SOVEREIGN: Does the accuracy gap on long-context multimodal benchmarks (e.g., Video-MME, Needle-in-a-Haystack) between Mo

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

We study the continual pretraining recipe for scaling language models' context lengths to 128K, with a focus on data engineering. We hypothesize that long context modeling, in particular \textit{the ability to utilize information at arbitrary input locations}, is a capability that is mostly already acquired through large-scale pretraining, and that this capability can be readily extended to contexts substantially longer than seen during training~(e.g., 4K to 128K) through lightweight continual pretraining on appropriate data mixture. We investigate the \textit{quantity} and \textit{quality} of

1 Introduction

Analysis of: Data Engineering for Scaling Language Models to 128K Context. Research goal: Does the accuracy gap on long-context multimodal benchmarks (e.g., Video-MME, Needle-in-a-Haystack) between MoE and dense models widen or narrow as context length increases from 128K to 10M tokens?.

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

3 papers retrieved. 0 claims extracted, 0 verified. Tribunal: 2.2/10 → 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

- http://arxiv.org/abs/2502.17129v2
- http://arxiv.org/abs/2505.09561v2
- http://arxiv.org/abs/2402.10171v1