SOVEREIGN: How does the S* hybrid test-time scaling framework affect the inference efficiency (measured in average latenc

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

We introduce MiniMax-01 series, including MiniMax-Text-01 and MiniMax-VL-01, which are comparable to top-tier models while offering superior capabilities in processing longer contexts. The core lies in lightning attention and its efficient scaling. To maximize computational capacity, we integrate it with Mixture of Experts (MoE), creating a model with 32 experts and 456 billion total parameters, of which 45.9 billion are activated for each token. We develop an optimized parallel strategy and highly efficient computation-communication overlap techniques for MoE and lightning attention. This app

1 Introduction

Analysis of: MiniMax-01: Scaling Foundation Models with Lightning Attention. Research goal: How does the S* hybrid test-time scaling framework affect the inference efficiency (measured in average latency and FLOPs per correct solution) relative to standard parallel scaling when evaluated on self-invoking code generation tasks from HumanEval Pro and MBPP Pro?.

2 Methodology

Multi-query arXiv search (1 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

1 papers retrieved. 6 claims extracted, 6 verified. Tribunal: 6.3/10 → REVISE (revision_round=1). Policy: SOFT_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
The MiniMax-01 series includes MiniMax-Text-01 and MiniMax-VL-01 models.	✓	0.25
MiniMax-Text-01 can process up to 1 million tokens during training and extrapolate to 4 million tokens during inference.	✓	0.27
The model has 32 experts and 456 billion total parameters, with 45.9 billion activated per token.	✓	0.27
The model achieves 20-32 times longer context window than state-of-the-art models.	✓	0.23
MiniMax-VL-01 was built through continued training with 512 billion vision-language tokens.	✓	0.33
The models match the performance of state-of-the-art models like GPT-4o and Claude-3.5-Sonnet.	✓	0.24

References

- https://doi.org/10.48550/arxiv.2501.08313