SOVEREIGN: What is the throughput improvement of Qwen3's thinking-mode routing over dense baselines on NLVR2 when control

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Foundation models, now powering most of the exciting applications in deep learning, are almost universally based on the Transformer architecture and its core attention module. Many subquadratic-time architectures such as linear attention, gated convolution and recurrent models, and structured state space models (SSMs) have been developed to address Transformers' computational inefficiency on long sequences, but they have not performed as well as attention on important modalities such as language. We identify that a key weakness of such models is their inability to perform content-based reasoni

1 Introduction

Analysis of: Mamba: Linear-Time Sequence Modeling with Selective State Spaces. Research goal: What is the throughput improvement of Qwen3's thinking-mode routing over dense baselines on NLVR2 when controlling for model parameter count and sequence length?.

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

5 papers retrieved. 6 claims extracted, 4 verified. Tribunal: 7.2/10 $\rightarrow$ 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.

5 Extracted Claims

Claim	Verified	Confidence
Mamba achieves 5x higher throughput than Transformers during inference.	×	0.14
Mamba scales linearly in sequence length.	×	0.12
Mamba achieves state-of-the-art performance across language, audio, and genomics modalities.	✓	0.22
A key weakness of subquadratic-time architectures like linear attention and SSMs is their inability to perform content-b	✓	0.23
Letting SSM parameters be functions of the input addresses their weakness with discrete modalities.	✓	0.24
Mamba integrates selective SSMs into a simplified end-to-end neural network architecture without attention or MLP blocks	✓	0.26

References

- https://doi.org/10.1109/jproc.2021.3067593
- https://doi.org/10.48550/arxiv.2312.00752
- https://doi.org/10.1109/comst.2023.3249835