SOVEREIGN: MambaFormer: Token-Level Guided Routing Mixture-of-Experts for Accurate and Effi

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 27, 2026

Abstract

The deployment of large language models (LLMs) in real-world clinical applications is constrained by the fundamental trade-off between computational cost and the efficiency of linear-time models. To address this, we propose an LLM-based MambaFormer hybrid Mixture-of-Experts (MoE) framework for efficient medical question-answering (QA) and clinical assistance. The MambaFormer employs a lightweight gating mechanism that performs token-level dynamic routing to a customized Transformer expert (ET5) for short, complex queries or to a State Space Model expert (EMamba) for long, high-throughput seque

1 Introduction

Analysis of: MambaFormer: Token-Level Guided Routing Mixture-of-Experts for Accurate and Efficient Clinical Assistance. Research goal: How does the throughput-accuracy trade-off of dynamic expert specialization in MoE-VLMs compare to fixed top-2 routing on VQA v2 and GQA benchmarks when scaling active parameters from 1B to 10B?.

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

15 papers retrieved. 6 claims extracted, 6 verified. Tribunal: 7.5/10 → APPROVE (revision_round=0). 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
The MambaFormer framework employs a lightweight gating mechanism that performs token-level dynamic routing to a customiz	✓	0.40
The customized EMamba and ET5 models are tailored to accommodate input sequence dimensionality, embedding structure, seq	✓	0.34
The EMamba and ET5 models are fine-tuned through transfer learning on a new, custom-designed DentalQA dataset.	✓	0.29
Intelligent routing decisions are driven by the contextual complexity of token embeddings, normalized sequence length, a	✓	0.31
A novel utility-guided multi-objective loss jointly optimizes decisions, router parameters, routing behavior, expert uti	✓	0.40
The proposed MambaFormer is cross-validated (holdout) for medical QA on the new, custom-designed DentalQA and PubMedQA d	✓	0.34

References

- http://arxiv.org/abs/2605.15484v1
- https://www.semanticscholar.org/paper/8a3ee6b06695a444b63e79d9ff542d1c7c7b947a
- https://www.semanticscholar.org/paper/f9355312c476dc79257303c37811f5fb2786d30c