SOVEREIGN: MoEcho: Exploiting Side-Channel Attacks to Compromise User Privacy in Mixture-of

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 27, 2026

Abstract

The transformer architecture has become a cornerstone of modern AI, fueling remarkable progress across applications in natural language processing, computer vision, and multi-modal learning. As these models continue to scale explosively for performance, implementation efficiency remains a critical challenge. Mixture-of-Experts (MoE) architectures, selectively activating specialized subnetworks (experts), offer a unique balance between model accuracy and computational cost. However, the adaptive routing in MoE architectures—where input tokens are dynamically directed to specialized experts base

1 Introduction

Analysis of: MoEcho: Exploiting Side-Channel Attacks to Compromise User Privacy in Mixture-of-Experts LLMs. Research goal: How does the accuracy of SMoES-based MoE-VLMs with soft modality-guided routing compare to dense models of equivalent parameter count on the MMMU benchmark across 7B to 34B scales, and what is the performance gap trend?.

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

13 papers retrieved. 5 claims extracted, 5 verified. Tribunal: 7.8/10 → 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 transformer architecture has become a cornerstone of modern AI, fueling progress across applications in natural lang	✓	0.28
Mixture-of-Experts (MoE) architectures selectively activate specialized subnetworks (experts) to balance model accuracy	✓	0.25
Adaptive routing in MoE architectures opens up a new attack surface for privacy breaches by leaving distinctive temporal	✓	0.31
MoEcho introduces four novel architectural side-channels on different computing platforms: Cache Occupancy Channels and	✓	0.32
MoEcho proposes four attacks that effectively breach user privacy in LLMs and VLMs based on MoE architectures: Prompt In	✓	0.33

References

- https://www.semanticscholar.org/paper/62a22eb108e5dfa476461229e2afb8ff1e2a1163
- https://www.semanticscholar.org/paper/f6ab135a925626ff3946b5360b0cf71891323e40
- http://arxiv.org/abs/2604.23996v1