SOVEREIGN: How does the inference throughput (tokens per second) of SMoES-based MoE-VLMs compare to dense models of equal

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

In performing a Bayesian analysis of astronomical data, two difficult problems often emerge. First, in estimating the parameters of some model for the data, the resulting posterior distribution may be multimodal or exhibit pronounced (curving) degeneracies, which can cause problems for traditional MCMC sampling methods. Second, in selecting between a set of competing models, calculation of the Bayesian evidence for each model is computationally expensive. The nested sampling method introduced by Skilling (2004), has greatly reduced the computational expense of calculating evidences and also pr

1 Introduction

Analysis of: Multimodal nested sampling: an efficient and robust alternative to Markov Chain Monte Carlo methods for astronomical data analyses. Research goal: How does the inference throughput (tokens per second) of SMoES-based MoE-VLMs compare to dense models of equal total parameters on multimodal reasoning benchmarks (e.g., MMMU, MathVista) at 7B and 34B scales under varying batch sizes and sequence lengths?.

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

10 papers retrieved. 8 claims extracted, 8 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 nested sampling method introduced by Skilling (2004) has greatly reduced the computational expense of calculating ev	✓	0.31
The nested sampling method produces posterior inferences as a by-product	✓	0.24
Mukherjee et al. (2006) applied the nested sampling method successfully in cosmological applications	✓	0.26
Mukherjee et al. (2006) implementation was efficient only for unimodal distributions without pronounced degeneracies	✓	0.30
Shaw et al. (2007) introduced a clustered nested sampling method which is significantly more efficient in sampling from	✓	0.35
Shaw et al. (2007) method determines the expectation and variance of the final evidence from a single run of the algorit	✓	0.30
The new methods presented in this paper lead to a further substantial increase in efficiency for sampling and evidence e	✓	0.26
The new methods presented in this paper provide an even more efficient technique for estimating the uncertainty on the e	✓	0.20

References

- https://doi.org/10.3390/informatics11030057
- https://doi.org/10.1613/jair.5477

- https://doi.org/10.1111/j.1365-2966.2007.12353.x