SOVEREIGN: What is the performance gap trend in visual reasoning accuracy between SMoES-based MoE-VLMs and dense models a

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

With the increasing data volume, there is a trend of using large-scale pre-trained models to store the knowledge into an enormous number of model parameters. The training of these models is composed of lots of dense algebras, requiring a huge amount of hardware resources. Recently, sparsely-gated Mixture-of-Experts (MoEs) are becoming more popular and have demonstrated impressive pretraining scalability in various downstream tasks. However, such a sparse conditional computation may not be effective as expected in practical systems due to the routing imbalance and fluctuation problems. Generall

1 Introduction

Analysis of: FlexMoE: Scaling Large-scale Sparse Pre-trained Model Training via Dynamic Device Placement. Research goal: What is the performance gap trend in visual reasoning accuracy between SMoES-based MoE-VLMs and dense models across 7B, 13B, and 34B parameter scales on the MMMU benchmark under varying expert activation ratios?.

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

8 papers retrieved. 6 claims extracted, 6 verified. Tribunal: 7.7/10 → 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
FlexMoE is a novel DNN training framework that systematically and transparently addresses the inefficiency caused by dyn	✓	0.25
Sparsely-gated Mixture-of-Experts (MoEs) have demonstrated impressive pretraining scalability in various downstream task	✓	0.24
MoE models suffer from routing imbalance and fluctuation problems that make sparse conditional computation less effectiv	✓	0.24
FlexMoE overcomes routing imbalance and fluctuation problems by a dynamic expert management and device placement mechani	✓	0.30
FlexMoE introduces a novel scheduling module over the existing DNN runtime to monitor data flow, make scheduling plans,	✓	0.38
A simple but efficient heuristic algorithm is exploited to dynamically optimize device placement during training.	✓	0.26

References

- https://doi.org/10.1145/3588964
- https://doi.org/10.48550/arxiv.2404.16821
- https://doi.org/10.48550/arxiv.2404.15045