Dynamic Expert Capacity Allocation and Its Impact on HumanEval Accuracy in MoE Models

Assignee Research

May 29, 2026

Abstract

Mixture-of-Experts (MoE) architectures enable conditional computation by routing inputs to multiple expert subnetworks and are often motivated as a mechanism for scaling large language models. In this project, we instead study MoE behavior in an image classification setting, focusing on predictive performance, expert utilization, and generalization. We compare dense, SoftMoE, and SparseMoE classifier heads on the CIFAR10 dataset under comparable model capacity. Both MoE variants achieve slightly higher validation accuracy than the dense baseline while maintaining balanced expert utilization

1 Introduction

This paper examines: Mixture-of-Experts Models in Vision: Routing, Optimization, and Generalization. Research question: How does dynamic expert capacity allocation affect pass@1 accuracy on HumanEval compared to fixed-capacity MoE models with equivalent parameter counts?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.7/10.

3 Results

13 papers retrieved. 5 claims extracted; 1 independently verified. Quality review score: 4.7/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
Mixture-of-Experts (MoE) architectures introduce conditional computation, routing inputs to a subset of specialized expe	×	0.13
In large language models, MoEs are primarily motivated by efficiency and scalability.	×	0.06
The study uses the CIFAR-10 dataset to compare dense classifier heads with SoftMoE and SparseMoE variants built on a sha	×	0.13
The controlled setup allows isolating the effects of expert routing, load balancing, and sparsity on predictive performa	×	0.06
The study analyzes generalization-related behavior through the geometry of the loss landscape, specifically computing He	✓	0.18

References

- http://arxiv.org/abs/2506.14646v2
- http://arxiv.org/abs/2601.15021v1
- http://arxiv.org/abs/2603.01697v1