Scaling ALF-LB with Model Size and Expert Count in Code Generation Benchmarks

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How does the ALF-LB method scale with increasing model size and number of experts in terms of training stability and final accuracy on the MBPP code generation benchmark compared to traditional MoE. 15 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.4/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Not All Experts are Equal: Efficient Expert Pruning and Skipping for Mixture-of-Experts Large Language Models. Research question: How does the ALF-LB method scale with increasing model size and number of experts in terms of training stability and final accuracy on the MBPP code generation benchmark compared to traditional MoE approaches?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.4/10.

3 Results

12 papers retrieved. 15 claims extracted; 1 independently verified. Quality review score: 4.4/10.

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
LLaMA-2 70B has fewer active parameters (13B) during inference.	×	0.03
MoE LLMs achieve a reduction in on-the-fly (active) parameters by choosing only top-k experts for the inference of each	×	0.12
Loading the Mixtral 8x7B model in bf16 format requires at least two A100-80G GPUs.	×	0.02
In the Mixtral 8x7B model, the eight experts constitute around 96% (45B out of 47B) of the total number of parameters.	×	0.03
Recent studies, such as (Chi et al., 2022), have demonstrated a discrepancy in expert training outcomes.	×	0.07
The proposed method significantly reduces memory usage for deploying MoE LLMs and enhances their inference speed.	×	0.10
The proposed method aims to minimize the token reconstruction loss in a layer-by-layer manner.	×	0.03
The proposed method examines expert-level pruning for both task-agnostic and task-specific models.	✓	0.18
The proposed method introduces hardware-friendly post-training methods for either permanently removing unimportant exper	×	0.14
The MoE layer of the Mixtral 8x7B model features 8 experts.	×	0.08
Each token x in the input sequence is routed to the top-2 experts based on the routing weights w.	×	0.08
The router computes routing logits l = {l0, . . . , ln-1} and routing weights w = Softmax(l) for the experts.	×	0.03
The top-k experts, where k = 2 for the Mixtral 8x7B model, are selected based on their routing weights to process the to	×	0.06
Each of the k selected experts, applying a SwiGLU transformation Ei($\cdot$) (i $\in${0, 1, . . . , n -1}), contributes to the fin	×	0.02
The final output is a weighted sum of the individual expert outputs, with weights ewi being the normalized values.	×	0.01

References

- http://arxiv.org/abs/2306.08568v2
- http://arxiv.org/abs/2402.14800v2
- http://arxiv.org/abs/2412.21199v2