ALF-LB Impact on Inference Latency and Throughput in HumanEval Code Generation

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: What is the impact of ALF-LB on inference latency and throughput during code generation tasks in the HumanEval benchmark when compared to auxiliary-loss-based MoE methods at different expert. 12 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: MoE++: Accelerating Mixture-of-Experts Methods with Zero-Computation Experts. Research question: What is the impact of ALF-LB on inference latency and throughput during code generation tasks in the HumanEval benchmark when compared to auxiliary-loss-based MoE methods at different expert utilization levels (e.g., 1, 2, or 4 experts per token)?.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.7/10.

3 Results

11 papers retrieved. 12 claims extracted; 0 independently verified. Quality review score: 3.7/10.

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
MoE++ is trained exclusively on public datasets, making it accessible for academic research settings.	×	0.04
MoE++ uses the tokenizer of LLaMA2, which contains 65,536 vocabulary tokens.	×	0.02
The hyper-parameters for MoE++ are selected based on the common practice for dense language models.	×	0.05
MoE++ replaces all FFN layers in the transformer with MoE++ layers and sets the Top-K to 2 for every layer, resulting in	×	0.05
MoE++ is evaluated using the lm-evaluation-harness package on an extensive suite of downstream tasks.	×	0.04
MoE++ reports 0-shot accuracy on ARC Easy, LAMBADA, LogiQA, PIQA, SciQ, and WinoGrande.	×	0.01
MoE++ reports the accuracy of tasks from the Open LLM Leaderboard, including 10-shot HellaSwag, 25-shot ARC Challenge, a	×	0.03
MoE++ reports the exact match score for 32-shot Natural Questions and the accuracy for 32-shot BoolQ.	×	0.03
Comparative evaluations of MoE++ against vanilla MoE models start with a modest scale of 0.6B parameters and expand up t	×	0.05
The training budget for all MoE++ and vanilla MoE models listed is 100B tokens.	×	0.07
The visualization of the number of FFN experts activated per token at the token level comes from the 'MoE++ 7B/(16+4)E'	×	0.07
The evaluation for the visualization of the number of FFN experts activated per token is done over 60,000 tokens and ave	×	0.09

References

- http://arxiv.org/abs/2506.14646v2
- http://arxiv.org/abs/2410.07348v1

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