Instance-Level Dynamic Loss Adaptation in Few-Shot LLM Reasoning Benchmarks

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: To what extent does instance-level dynamic loss adaptation improve the few-shot learning performance of LLMs on reasoning benchmarks compared to static group-level reweighting strategies. 12 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Benchmarking Few-shot Transferability of Pre-trained Models with Improved Evaluation Protocols. Research question: To what extent does instance-level dynamic loss adaptation improve the few-shot learning performance of LLMs on reasoning benchmarks compared to static group-level reweighting strategies?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.2/10.

3 Results

15 papers retrieved. 12 claims extracted; 1 independently verified. Quality review score: 4.2/10.

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
There is a positive correlation (r = 0.38) between sensitivity and ensemble penalty in the Hyperparameter Ensemble (HPE)	×	0.10
Paired statistical tests were conducted across 6000 tasks.	×	0.03
The practical advantages of sophisticated transfer algorithms over simple fine-tuning are often negligible according to	×	0.13
Simple all-parameter fine-tuning (Full-FT) avoids expected overfitting in few-shot scenarios.	×	0.10
The performance collapse of multimodal models on specialized domains is quantified as a consequence of linguistic rarity	✓	0.24
For Task ID 3 in Table (p3), the optimal Backbone learning rate is 40.	×	0.03
For the Plant Disease dataset in Table (p4), the optimal Head learning rate is 0.005.	×	0.02
In configuration (2,2) on the Aircraft dataset, the accuracy is 71.01.	×	0.02
The DINOv2-base model pre-trained on LVD-142M achieves an average accuracy of 55.4$\pm$1.4 across the benchmark datasets.	×	0.07
The LoRA method achieves an average accuracy of 70.7$\pm$1.3 across the benchmark datasets.	×	0.04
The VPT method achieves an accuracy of 96.8$\pm$0.4 on the Flowers dataset.	×	0.04
The BiT-R101 model pre-trained on IN-14M achieves an accuracy of 88.2$\pm$0.8 on the UCF dataset.	×	0.06

References

- http://arxiv.org/abs/2208.01009v2
- http://arxiv.org/abs/2508.04848v1
- http://arxiv.org/abs/2603.00478v1