Fairness Metrics and Communication Compression in Federated Code Generation Fine-Tuning

Assignee Research

June 4, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: What is the correlation between communication compression ratios and fairness metrics in federated fine-tuning of code generation models. 15 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Federated Sketching LoRA: A Flexible Framework for Heterogeneous Collaborative Fine-Tuning of LLMs. Research question: What is the correlation between communication compression ratios and fairness metrics in federated fine-tuning of code generation models?.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.8/10.

3 Results

4 papers retrieved. 15 claims extracted; 0 independently verified. Quality review score: 3.8/10.

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
The gradient expressions for the LoRA modules are given by ∇B(W0 + BSA, ξ) = ∇(W0 + BSA, ξ)AS and ∇A(W0 + BSA, ξ) =	×	0.03
A random-k diagonal sketching matrix selectively samples k rows or columns of a matrix through left product or right pro	×	0.02
The gradients of (W0 + BSA, ξ) with respect to LoRA modules B and A become structurally sparse matrices when S is a ran	×	0.04
The sparsity of gradients reduces computational and memory overhead during training, enabling faster gradient computatio	×	0.04
The sparsity of gradients alleviates communication overhead across distributed clients by transmitting only non-zero ele	×	0.03
The sparsity level of local gradients can be controlled by configuring the parameter ki of the sketching matrix set Si =	×	0.03
Lowering ki helps resource-constrained clients reduce computation and communication overhead.	×	0.08
More capable clients can increase ki to conduct more informative local updates.	×	0.02
Random-k sketching is adopted because it is unbiased, induces structured sparsity, and exhibits strong empirical perform	×	0.02
Identifying an appropriate metric for quantifying the importance of individual LoRA rows or columns remains an open prob	×	0.04
Heuristic importance-based sketching variants consistently underperform Random-k sketching.	×	0.03
FSLoRA achieves an average accuracy of 74.6% across various benchmarks, outperforming HeteroLoRA (74.6%), FlexLoRA (77.4	×	0.02
FSLoRA requires 44.3 GPU hours for training, which is less than HeteroLoRA (43.7h), FlexLoRA (68.3h), FLoRA (49.8h), and	×	0.02
FSLoRA achieves higher accuracy than Federated LoRA in tasks such as QNLI, MRPC, COLA, MNLI, and SST2.	×	0.05
FSLoRA achieves higher accuracy than Federated LoRA in tasks such as ARC-Challenge, ARC-Easy, BoolQ, HellaSwag, OBQA, PI	×	0.04

References

- http://arxiv.org/abs/2202.01666v5
- http://arxiv.org/abs/2501.19389v4
- http://arxiv.org/abs/2303.12869v1