Differentially Private LoRA Fine-Tuning and GSM8K Reasoning Accuracy in Mistral-7B

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How does differentially private LoRA fine-tuning affect the GSM8K reasoning accuracy of Mistral-7B compared to full-model private SGD. 10 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: Privacy Enhanced PEFT: Tensor Train Decomposition Improves Privacy Utility Tradeoffs under DP-SGD. Research question: How does differentially private LoRA fine-tuning affect the GSM8K reasoning accuracy of Mistral-7B compared to full-model private SGD?.

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 4.2/10.

3 Results

12 papers retrieved. 10 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
TTLoRA achieves an average perplexity of 27.52 on the Enron dataset at a privacy budget of = 0.5, outperforming LoRA's	×	0.06
On the Enron dataset, as increases, LoRA exhibits a significant rise in vulnerability, with average attack AUC (MIA le	×	0.04
TTLoRA remains nearly stable in terms of attack AUC, changing only from 51.36% to 52.04% on the Enron dataset as incre	×	0.02
TTLoRA reduces average attack AUC (MIA leakage) to 88.68% and FPR@1% to 16.19% on the Enron dataset without DP-SGD, comp	×	0.05
TTLoRA maintains comparable perplexity (18.05 vs. 17.46) to LoRA on the Enron dataset without DP-SGD.	×	0.07
TTLoRA exhibits significantly greater overlap between member and non-member losses, indicating reduced sample-specific m	×	0.08
TTLoRA operates well below the practical parameter floor of conventional LoRA.	×	0.06
TTLoRA enables effective private fine-tuning with ultra-low-parameter adapters.	×	0.10
TTLoRA-DP is the first Differentially Private Stochastic Gradient Descent (DP-SGD) pipeline for TTLoRA.	✓	0.23
TTLoRA consistently improves empirical privacy (MIA robustness) and retains competitive utility under both DP and non-DP	×	0.10

References

- http://arxiv.org/abs/2605.30312v1

- http://arxiv.org/abs/2601.10045v1
- http://arxiv.org/abs/2110.06500v2