Privacy Budget Effects on Convergence and Throughput in CausalMixFT Fine-Tuning

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: What is the impact of varying privacy budgets on the convergence speed and training throughput of CausalMixFT versus standard differentially private optimizers during LLM fine-tuning. 15 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Revisiting Privacy, Utility, and Efficiency Trade-offs when Fine-Tuning Large Language Models. Research question: What is the impact of varying privacy budgets on the convergence speed and training throughput of CausalMixFT versus standard differentially private optimizers during LLM fine-tuning?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.7/10.

3 Results

14 papers retrieved. 15 claims extracted; 3 independently verified. Quality review score: 6.7/10.

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
Membership inference attacks (MIA) aim to identify if a certain datapoint was in the training dataset using model confid	×	0.03
Training data extraction attacks attempt to extract specific parts of data from the training dataset using different pro	×	0.06
A recent study [60] indicates that membership inference attacks may be unreliable due to the practical limitation of req	×	0.03
The 'exposure' metric introduced in [9] evaluates a model's vulnerability by using artificially introduced data ('canari	×	0.04
The 'exposure' metric's formulation relies on assumptions about the surrounding knowledge of other canaries that do not	×	0.05
Existing studies focusing on privacy attacks in LLMs often consider sensitive and non-sensitive counterparts during trai	×	0.07
Differential privacy (DP-SGD) adds theoretical privacy guarantees by clipping gradients of each datapoint and adding noi	×	0.14
The primary goal of DP-SGD is to reduce the influence of individual datapoints in the training procedure to prevent leak	×	0.05
Existing work utilizing DP-SGD [57] demonstrates a trade-off between privacy (measured as theoretical guarantee) and uti	×	0.08
Efficient fine-tuning methods like LoRA mitigate privacy risks similarly to private fine-tuning methods like DP-SGD.	✓	0.37
The finding that LoRA mitigates privacy risks contradicts the prevailing wisdom that privacy and efficiency objectives a	✓	0.29
DP-SGD incurs significant additional computational overhead.	×	0.05
LoRA significantly reduces computational costs compared to full fine-tuning or DP-SGD.	×	0.10
The study's findings are established using evaluations on open-source language models from the Pythia, Gemma, Llama, and	✓	0.23
Prior to this work, no studies had investigated the privacy risks associated with efficient training methods.	×	0.13

References

- http://arxiv.org/abs/2110.06500v2
- http://arxiv.org/abs/2601.10237v2
- http://arxiv.org/abs/2502.13313v2