Differential Privacy Noise Scales and Few-Shot Reasoning in LLaMA-2 Models

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: What is the impact of differential privacy noise scales on the few-shot reasoning capabilities of LLaMA-2 models across arithmetic and logical deduction tasks. 13 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Privacy Preserving In-Context-Learning Framework for Large Language Models. Research question: What is the impact of differential privacy noise scales on the few-shot reasoning capabilities of LLaMA-2 models across arithmetic and logical deduction tasks?.

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 4.8/10.

3 Results

15 papers retrieved. 13 claims extracted; 3 independently verified. Quality review score: 4.8/10.

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 model with ε = 0 shows zero-shot performance uniformly poor across tasks (e.g., 24.8% on AGNews).	×	0.04
The model with ε = ∞ performs better than the current SOTA Amin et al. (2024)'s baseline on Gemma-1.12B in four tasks: D	×	0.02
The model with ε = ∞ outperforms Tang et al. (2023)'s baseline with GPT3 Babbage that uses only top-100 logprobs on most	×	0.02
The model with ε = 1 outperforms both baselines by Tang et al. (2023) and Amin et al. (2024) with Gemma-1.12B on three o	×	0.02
The model with ε = 1 outperforms Tang et al. (2023)'s baseline with GPT3 Babbage that uses only top-100 logprobs on AGNe	×	0.02
Performance on DBPedia generally improves with larger values of ε, indicating that relaxing privacy constraints allows f	×	0.05
The best performance for 4-shot (80.94), 8-shot (81.44), and 12-shot (75.48) on DBPedia appears in the ε = ∞ (non-privat	×	0.05
For TREC, ε = 8 outperforms all other settings for 1-shot, 2-shot, and 4-shot configurations.	×	0.04
The proposed method performs inference on private records and aggregates the resulting per-token output distributions, e	✓	0.39
The proposed method outperforms previous state-of-the-art methods on in-context-learning (ICL) tasks.	✓	0.26
Current approaches to creating differentially private text with large language models fall into two main groups: private	✓	0.18
Fine-tuning methods update model weights on the private data using a DP-SGD style algorithm and often produce high-quali	×	0.07
Private prediction based methods only rely on test-time inference instead of fine-tuning the model.	×	0.08

References

- http://arxiv.org/abs/2509.13625v4
- http://arxiv.org/abs/2407.04973v1
- http://arxiv.org/abs/2110.06500v2