Preserving Privacy in Fine-Tuned LLMs based on Domain-Specific Data

DJ Ranade[1], John Pugh[2]
1. Technological University, Dublin, Ireland | 2. Nathean Analytics, Ireland

Abstract. This study explores the balance between data utility and privacy in Large Language Models (LLMs) used in healthcare—a privacy-sensitive sector. We evaluate the effectiveness of Anonymization and Differential Privacy in preserving privacy, including re-identification on LLMs fine-tuned with treated MIMIC-IV health dataset. Further, we aim to identify optimal configurations that maintain robust privacy without compromising data utility. The results will inform strategies for deploying LLMs in environments where protecting sensitive information is paramount.

INTRODUCTION

The integration of LLMs in specialised sectors such as healthcare, legal, finance, insurance, etc. is increasing rapidly. While effective at general data processing, LLMs face challenges with domain-specific datasets, often requiring adaptations like Fine-tuning and Retrieval-Augmented Generation (RAG). These modifications aim to improve model performance on specialised tasks but may inadvertently increase the risk of privacy breaches, e.g., re-identification—even when privacy preservation measures are implemented, given the inherent capabilities of memorisation and inferencing of LLMs.

This study examines the effectiveness of Anonymization and Differential Privacy (DP) —considered the gold standard in privacy protection—on fine-tuned LLMs using the MIMIC-IV health dataset. We aim to assess the balance between privacy level and the utility of LLMs in handling sensitive data.

LITERATURE REVIEW

Information leakage and privacy concerns in LLMs are well-documented issues [2] and remain at the forefront of research priorities. The preservation of privacy in pre-trained LLMs has seen significant developments. For instance, in 2021, Horry et al. [3] at *Google* showcased the implementation of DP in constructing a privacy-preserved pre-trained model (BERT), albeit noting a trade-off between model performance and privacy assurance.

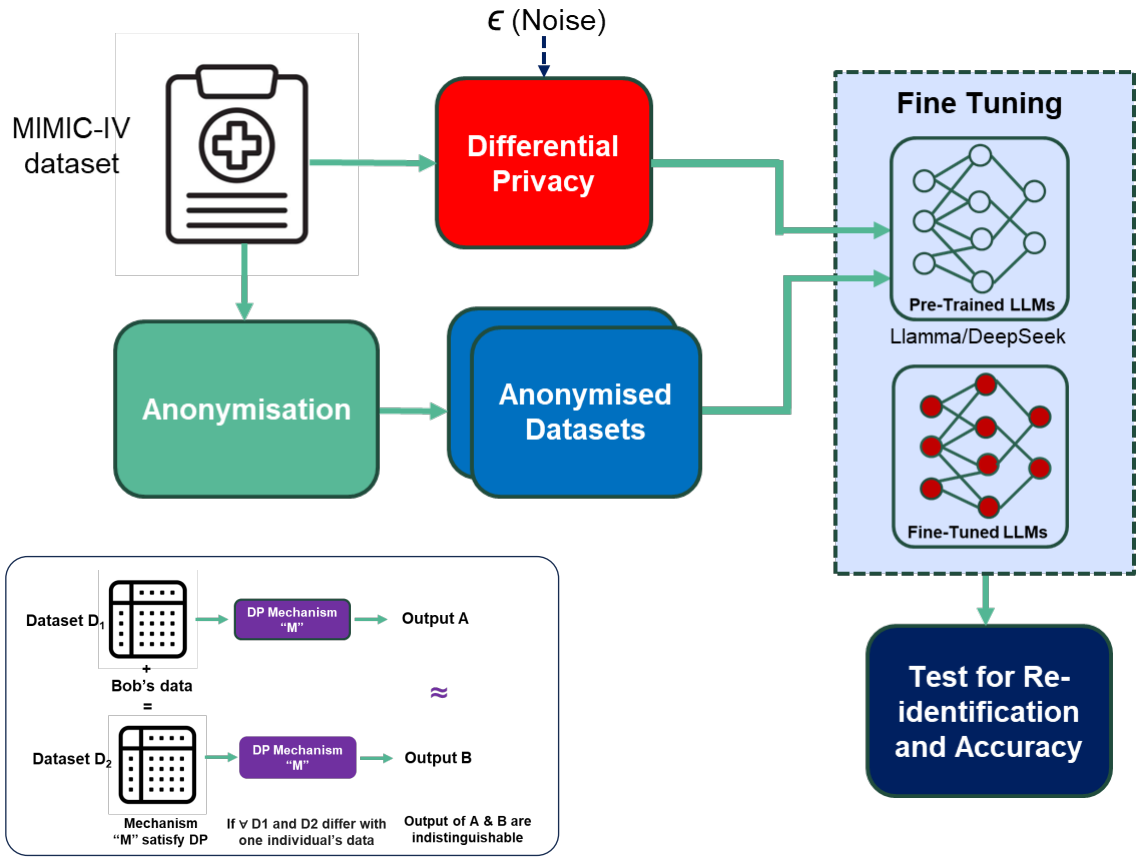

Fig. 1. Privacy preservation methodology

Additionally, the EW-Tune [4] method has demonstrated a requirement for 6% less noise, which is added to the dataset according to the DP technique, and achieved a 1.1% performance improvement for fine-tuning the General Language Understanding Evaluation (GLUE) dataset. Also, *Microsoft* has proposed a framework for Privately Fine-Tuning LLMs with DP [5] and achieved an accuracy of 83.5% using RoBERTa-Base with a privacy budget of ϵ=6.7. Despite these advancements, there is limited research on the application of these techniques across various datasets and LLMs. Our literature review reveals substantial potential for further investigation into the effectiveness of DP in mitigating re-identification risks of Personally Identifiable Information when DP LLMs are subjected to classified medical data.

RESEARCH METHODOLOGY

- Pre-trained LLMs, such as *Llamma3*, *DeepSeek*, etc., will be installed in a suitable machine with adequate computing resources.
- Pre-trained LLMs will be fine-tuned on the MIMIC-IV dataset using DP with different noise levels (ϵ).
- Fine-tuned LLMs will be tested for privacy and performance.
- A similar process will be employed on the anonymised dataset using traditional anonymisation techniques for comparison purposes. (Refer Fig.1)

PRELIMINARY CONSIDERATIONS

Recognising data privacy as a fundamental right under most jurisdictions globally, this research aligns with human-centred AI principles.

Suitable hardware for data processing and model fine-tuning will be required. The development of privacy strategies in LLMs will be guided by evaluations of re-identification risks and performance.

CONCLUSIONS

This research aims to critically examine the impact of privacy-preserving techniques on the utility and privacy of LLMs, specifically in the context of sensitive health data. Our findings will contribute to the ongoing dialogue in the AI community about balancing effective data utilisation with robust privacy protections.

References.
[1] Dwork, C. (2006). Differential privacy. In International colloquium on automata, languages, and programming (pp. 1-12). Berlin, Heidelberg: Springer Berlin Heidelberg.
[2] Buesser, B. (2024). Private Information Leakage in LLMs. In: Kucharavy, A., Plancherel, O., Mulder, V., Mermoud, A., Lenders, V. (eds) Large Language Models in Cybersecurity.
[3] Hoory, S., Feder, A., Tendler, A., Erell, S., Peled-Cohen, A., Laish, I., & Matias, Y. (2021). Learning and evaluating a differentially private pre-trained language model. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 1178-1189).
[4] R. Behnia, M. R. Ebrahimi, J. Pacheco and B. Padmanabhan. (2022). "EW-Tune: A Framework for Privately Fine-Tuning Large Language Models with Differential Privacy," *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, Orlando, FL, USA, 2022, pp. 560-566.
[5] Yu, D., Naik, S., Backurs, A., Gopi, S., Inan, H. A., Kamath, G., & Zhang, H. (2021). Differentially private fine-tuning of language models. *arXiv preprint arXiv:2110.06500*.

Acknowledgements. The HCAIM (the Human-Centred AI Master's Programme) Project is Co-Financed by the Connecting Europe Facility of the European Union Under Grant №CEF-TC-2020-1 Digital Skills 2020-EU-IA-0068. This poster was created as part of the Blended Intensive Programme organized under the Erasmus + Programme of the European Union