Privacy and Ethical Risks: Large Language Models on Anonymized Data

human centred artificial intelligence masters

Saige Wu, TU Dublin

ABSTRACT

This thesis investigates the potential of Large Language Models (LLMs) trained on domain-specific data that has undergone anonymization using privacy-preserving techniques. The research aims to evaluate whether LLMs can cross-reference anonymized data and successfully reidentify sensitive information. The study adopts a human-centered perspective, exploring the ethical challenges, risks, and safeguards necessary to ensure responsible AI development and deployment.

INTRODUCTION

Recent advances in Large Language Models (LLMs) have raised critical privacy concerns, particularly their potential to reidentify anonymized data. This capability challenges traditional privacy protection methods and raises significant ethical questions.

Our research is motivated by the growing sophistication of LLMs in pattern recognition, which may render conventional anonymization techniques inadequate. This study aims to evaluate LLMs' reidentification capabilities, assess existing privacy-preserving methods, and develop ethical frameworks to address these challenges.

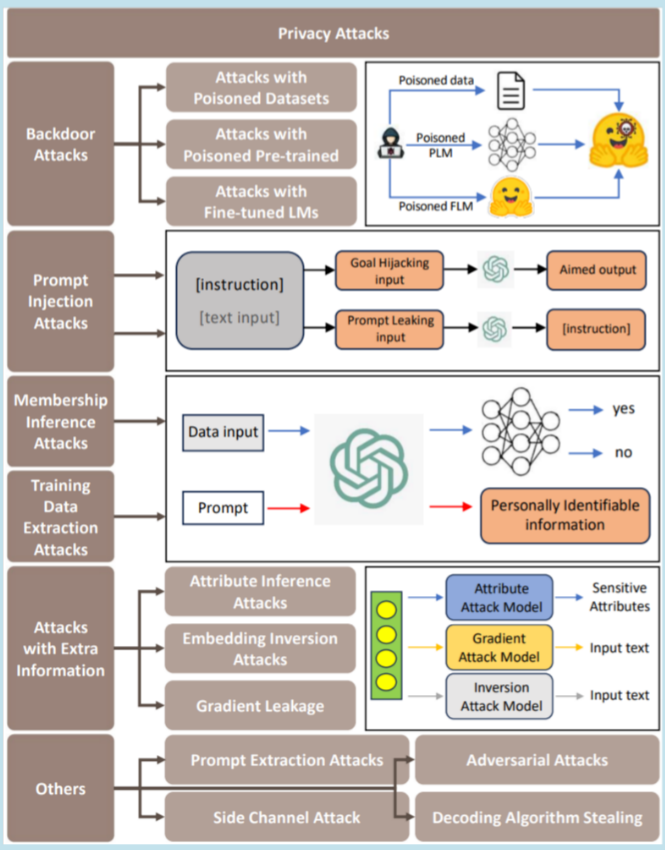

Fig. 1 An overview of existing privacy attacks on LLMs

RESEARCH METHODOLOGY

1. Data Processing
- Collect domain-specific datasets
- Apply privacy-preserving techniques

2. Model Development
- Train LLMs on anonymized data
- Evaluate pattern recognition capabilities

3. Security Assessment
- Conduct reidentification attacks
- Measure privacy breach risks

LITERATURE REVIEW

Privacy protection research has shifted focus from code security to language model safety. Early studies [1] examined basic code protection methods, while recent work [3] shows that protecting language models from privacy attacks is much more challenging. Even advanced privacy protection methods can be vulnerable to new types of attacks.

This shows there is a significant gap between what privacy protection methods promise in theory and how well they actually work in practice, especially as AI systems get better at handling specialized data.

PRELIMINARY CONSIDERATIONS

1. Technical Framework
   - Data Characteristics
   - Privacy Protection Methods
   - LLM Architecture & Capabilities

2. Evaluation Framework
   - Privacy Risk Assessment
   - Model Performance Metrics
   - Defense Effectiveness Analysis

3. Governance Framework
   - Privacy Compliance Standards
   - Regulatory Requirements
   - Stakeholder Impact Analysis

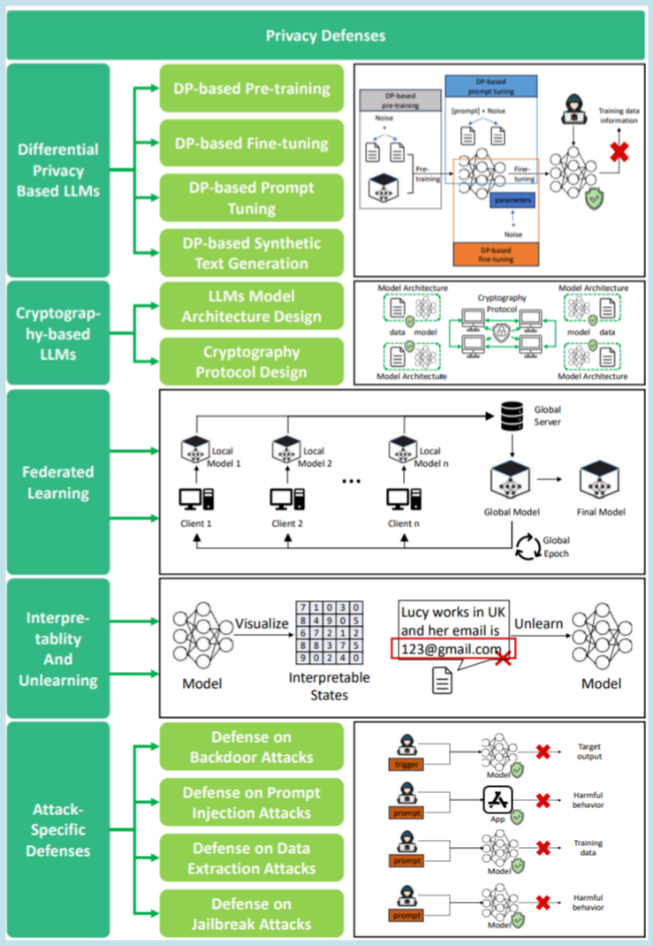

Fig. 2 An overview of existing privacy defenses on LLMs

CONCLUSIONS

This research demonstrates that while traditional anonymization techniques are insufficient for protecting privacy in Large Language Models, recent privacy-preserving methods have shown significant improvements [3,4]. As LLMs continue to evolve, future work must focus on developing robust privacy protection while maintaining model functionality.

References.
[1]. Witkowska, J. (2006). The Quality of Obfuscation and Obfuscation Techniques. In: Saeed, K., Pejaś, J., Mosdorf, R. (eds) Biometrics, Computer Security Systems and Artificial Intelligence Applications. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-36503-9_16
[2] S. A. Ebad, A. A. Darem and J. H. Abawajy, "Measuring Software Obfuscation Quality–A Systematic Literature Review," in IEEE Access, vol. 9, pp. 99024-99038, 2021, doi: 10.1109/ACCESS.2021.3094517
[3] Hu, H., Sablayrolles, A., Dupoux, E., Hsu, L., & Rainforth, T. (2023). Membership Inference Attacks Against Language Models via Neighbourhood Comparison. In 32nd USENIX Security Symposium (USENIX Security 23) (pp. 4763-4780).
[4] Li, H., Chen, Y., Luo, J., Wang, J., Peng, H., Kang, Y., Zhang, X., Hu, Q., Chan, C., Xu, Z., Hooi, B., & Song, Y. (2023). Privacy in Large Language Models: Attacks, Defenses and Future Directions. arXiv preprint arXiv:2310.10383.

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 IP Event Organized Under Erasmus + Programme of the European Union.