Preserving Privacy in the Age of Large Language Models

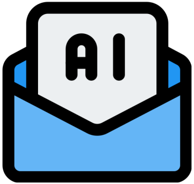
Slides for PhD Course on Generative AI

Francesco L. De Faveri

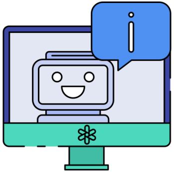
Department of Information Engineering, University of Padova, Italy

Introduction

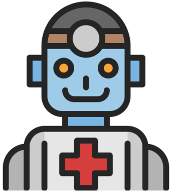
Large Language Models (LLMs) are nowadays widely used by society:



Gmail text

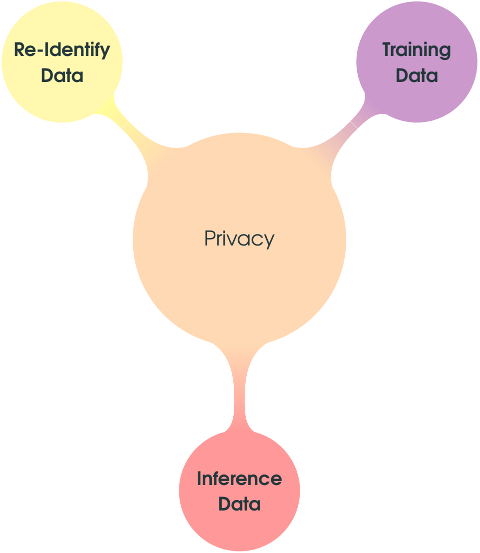


Search Assistant



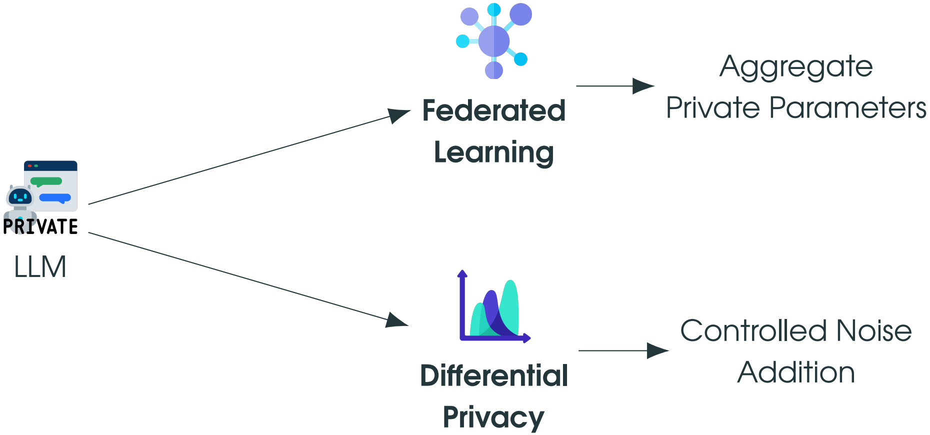
Healthcare

The “Y” for Privacy in LLMs

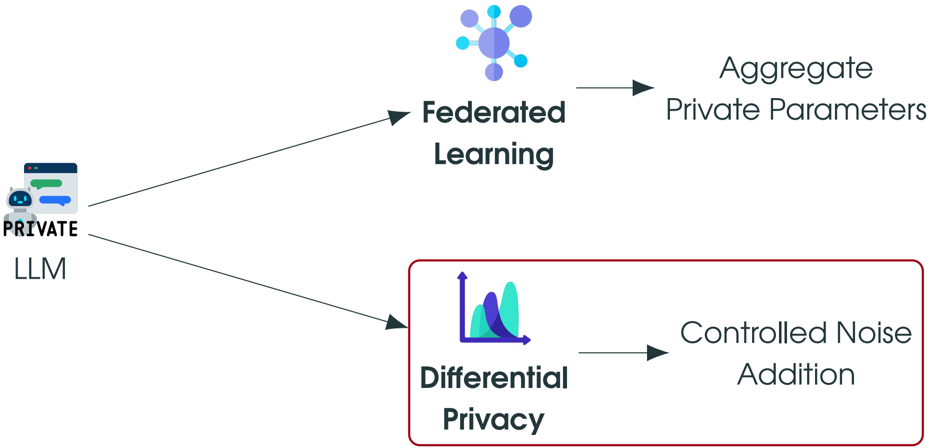


Related Works & Background

How to provide Privacy



How to provide Privacy



Differential Privacy (DP)

(ε, δ) - DP [5]

A mechanism $\mathcal{M}$ provides (ε, δ)-**DP** if for all datasets D and D' differing on at most one element, considering the privacy budget $\varepsilon, \delta \geq 0$, it holds:

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\varepsilon} \Pr[\mathcal{M}(D') \in S] + \delta \quad \forall S \subseteq \text{Range}(\mathcal{M})$$

(ε, δ) - Metric DP [3]

A mechanism $\mathcal{M}$ provides (ε, δ)-**metric DP** if for all datasets D and D' at a distance $d(D, D')$, considering the privacy budget $\varepsilon, \delta \geq 0$, it holds:

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\varepsilon d(D,D')} \Pr[\mathcal{M}(D') \in S] + \delta \quad \forall S \subseteq \text{Range}(\mathcal{M})$$

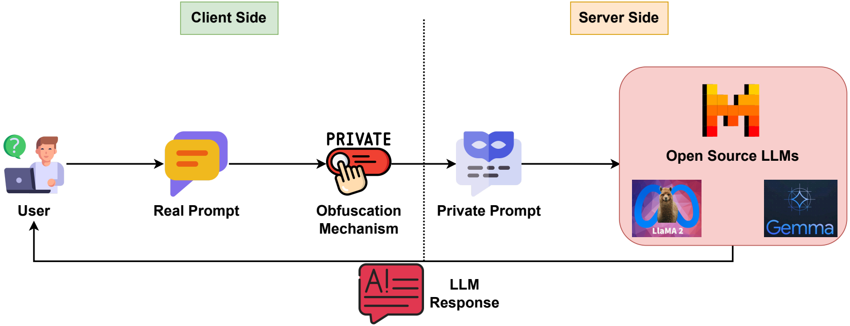
DP & LLMs

In LLMs privacy when can have three different scenarios on where to use (ε, δ)-DP:

- User **input**: Obfuscate the real user prompt
- LLM **training & fine-tuning**: DP-SGD [1, 2] or data sanitization [7]
- LLM **output**: Text generation is "noisy" [9] or sanitized [7]

Methodology & Setup

Prompt Obfuscation: Pipeline & Methods



DP for Natural Texts Obfuscation $\Rightarrow$ add noise x to the embeddings w_i:

- Spherical (CMP) [6]: $\hat{w}_i = w_i + x$ with $p_x(z) \propto \exp(-\varepsilon \cdot ||z||_2)$
- Elliptical (Mahalanobis) [10]: $\hat{w}_i = w_i + x$ with $p_x(z) \propto \exp(-\varepsilon \cdot ||z||_M)$

where $z \sim \mathcal{N}(\mathbf{0}, \mathbb{I}_n)$.

Experimental Setup

Datasets:

- BoolQ Dataset [4] $\sim 10k$ tuples: $\langle$*question*, *title*, *answer*, *passage*$\rangle$
- For the experiment $\longrightarrow$ 40 tuples ($$$ and time limitations).

LLMs:

Groq Free API $\left\{ \right.$

- LLaMA2-70b-chat
- Mixtral-8x7b-Instruct-v0.1
- Gemma-7b-it

Obfuscation:

- GloVe Wiki 300d embs.
- $\varepsilon \in \{5, 10, 20, 30, 40, 50\}$

Task:

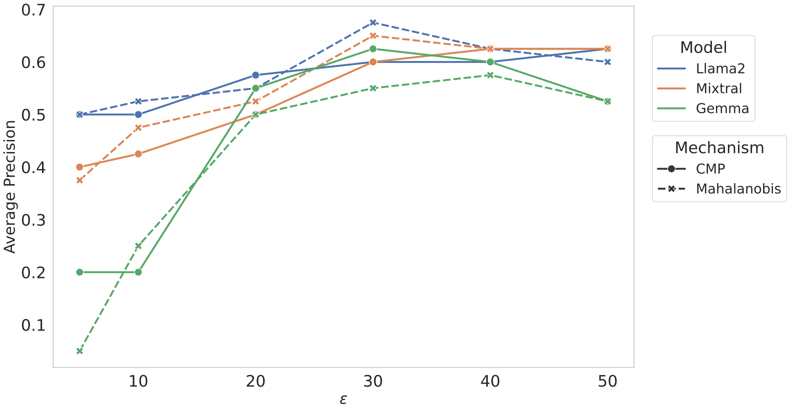
- Provide a private prompt to the LLM and evaluate the model's accuracy[1].

[1] Code available at https://github.com/Kekkodf/GenAI

Results Discussion

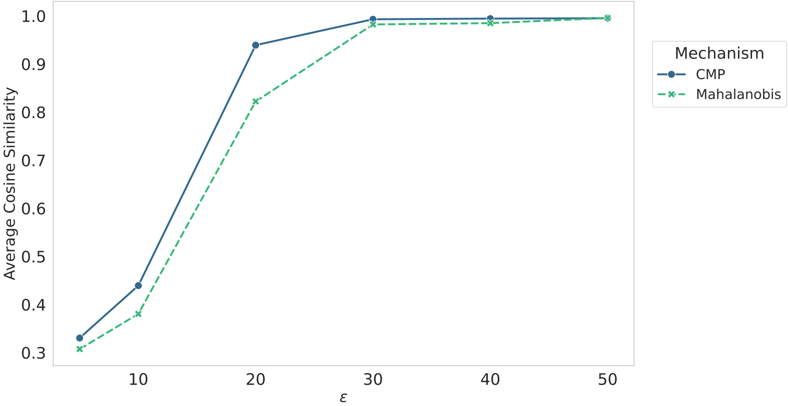
Results - Precision vs Privacy

- Measuring the precision in answering possible True/False questions varying the Privacy Budget ε. The higher, the better.



Results - Measuring Privacy

- Used Sentence-BERT [8] to compute the cosine similarity between the original prompt and the obfuscated one. The lower, the better.



Future Works

Future Works

Providing Privacy in LLMs remains an ***open problem***!

- Privacy & Hallucinations: A Possible Cocktail?
- Can a LLM Learn the Concept of Privacy?
- Is DP Enough to Preserve Privacy in the Age of LLMs?

Thanks for the attention!

Question time :)

Francesco L. De Faveri

Department of Information Engineering, University of Padova, Italy

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

References i

M. Abadi, A. Chu, I. J. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang.
Deep learning with differential privacy.
In E. R. Weippl, S. Katzenbeisser, C. Kruegel, A. C. Myers, and S. Halevi, editors, *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, October 24-28, 2016*, pages 308–318. ACM, 2016.

R. Behnia, M. Ebrahimi, J. Pacheco, and B. Padmanabhan.
Ew-tune: A framework for privately fine-tuning large language models with differential privacy.
In K. S. Candan, T. N. Dinh, M. T. Thai, and T. Washio, editors, *IEEE International Conference on Data Mining Workshops, ICDM 2022 - Workshops, Orlando, FL, USA, November 28 - Dec. 1, 2022*, pages 560–566. IEEE, 2022.

- K. Chatzikokolakis, M. E. Andrés, N. E. Bordenabe, and C. Palamidessi. **Broadening the scope of differential privacy using metrics.** In E. D. Cristofaro and M. K. Wright, editors, *Privacy Enhancing Technologies - 13th International Symposium, PETS 2013, Bloomington, IN, USA, July 10-12, 2013. Proceedings*, volume 7981 of *Lecture Notes in Computer Science*, pages 82–102. Springer, 2013.
- C. Clark, K. Lee, M. Chang, T. Kwiatkowski, M. Collins, and K. Toutanova. **Boolq: Exploring the surprising difficulty of natural yes/no questions.** In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*

2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (*Long and Short Papers*), pages 2924–2936. Association for Computational Linguistics, 2019.

C. Dwork, F. McSherry, K. Nissim, and A. D. Smith.
Calibrating noise to sensitivity in private data analysis.
In S. Halevi and T. Rabin, editors, *Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006, Proceedings*, volume 3876 of *Lecture Notes in Computer Science*, pages 265–284. Springer, 2006.

References v

O. Feyisetan, B. Balle, T. Drake, and T. Diethe.
Privacy- and utility-preserving textual analysis via calibrated multivariate perturbations.
In J. Caverlee, X. B. Hu, M. Lalmas, and W. Wang, editors, *WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA, February 3-7, 2020*, pages 178–186. ACM, 2020.

Y. Ishibashi and H. Shimodaira.
Knowledge sanitization of large language models.
CoRR, abs/2309.11852, 2023.

- N. Reimers and I. Gurevych.
  Sentence-bert: Sentence embeddings using siamese bert-networks.
  In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.
- X. Tang, R. Shin, H. A. Inan, A. Manoel, F. Mireshghallah, Z. Lin, S. Gopi, J. Kulkarni, and R. Sim.
  Privacy-preserving in-context learning with differentially private few-shot generation.
  CoRR, abs/2309.11765, 2023.

Z. Xu, A. Aggarwal, O. Feyisetan, and N. Teissier.
A differentially private text perturbation method using regularized mahalanobis metric.
In O. Feyisetan, S. Ghanavati, S. Malmasi, and P. Thaine, editors, *Proceedings of the Second Workshop on Privacy in NLP*, pages 7–17, Online, Nov. 2020. Association for Computational Linguistics.

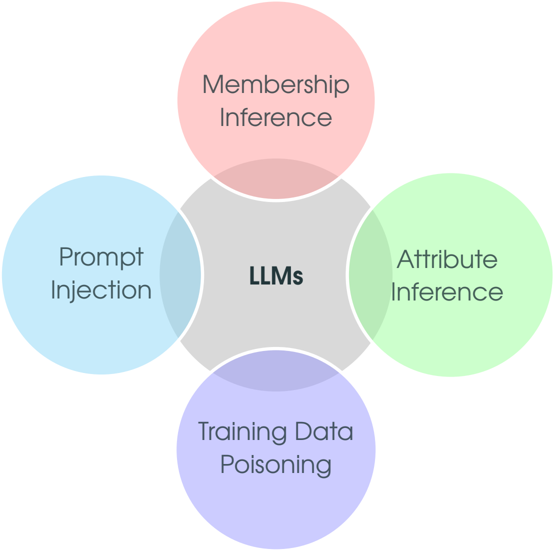
Backup Slides

Privacy & LLMs - Slides for PhD Course on Generative AI

Francesco L. De Faveri

Department of Information Engineering, University of Padova, Italy

Backup 1 - Privacy Related Attacks



Backup 2 - Federated Learning in LLMs

Privacy in Federated Learning

Federated Learning preserves data privacy by design as the data remains on users' devices and only model updates are shared with the server.

Important works:

- "Federated Large Language Model : A Position Paper", Chen et al., 2023.
- "FATE-LLM: A Industrial Grade Federated Learning Framework for Large Language Models", Fan et al., 2023.
- "Federated Learning of Large Language Models with Parameter-Efficient Prompt Tuning and Adaptive Optimization", Che et al., 2024.