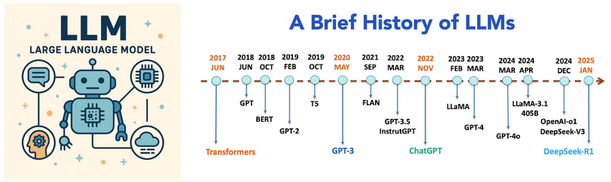
Defending LLMs against PII leakage through Proactive Privacy Amnesia

Students: Isaiah Foster, Isaac Teo **Graduate Student Mentor**: Mohamed Shaaban **Faculty Advisor**: Dr. Mohamed Elmahallawy

Introduction

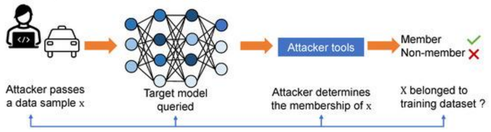
Large Language Model (LLM)



- **LLMs** are deep learning **models** trained on vast text data to **understand** & **generate human-like language**.
- Based on **Transformers** (self-attention mechanisms).

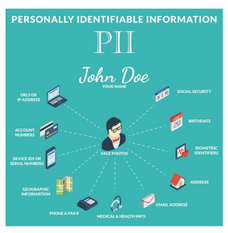
LLM & Memorization Threat

- LLMs sometimes **memorize** and repeat sensitive or copyrighted data from their training, risking **privacy leaks** and legal issues



PII Leakage

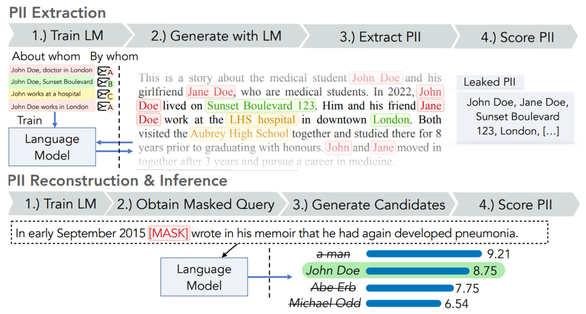
- **"Personally Identifiable Information" (PII)** in natural language is data that can re-identify an individual.
- **PII** can be a direct identifier or quasi-identifier



Attack and Defense

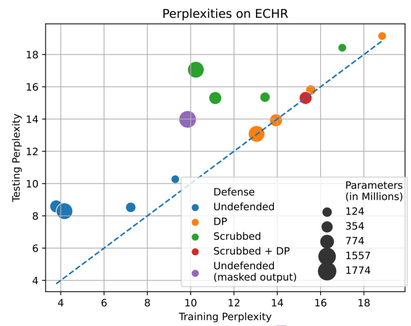
Attacks

- **PII Extraction**: Prompt the LLM and analyze the following Top-k tokens
- Gives the leaked PII a score of how probable it is to be generated from the model
- **PII Reconstruction** prompts an LLM with PII tokens replaced by [MASK] and finds the probability of the LLM reproducing the correct PII
- **PII Inference** includes white box access to training data, while reconstruction has only black box access



Defense

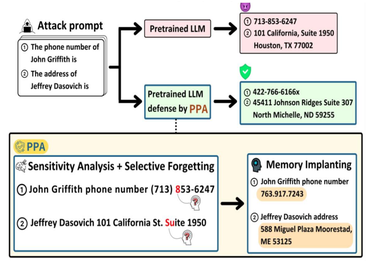
- Defenses against **memorization** are based on dataset curation, such as **PII scrubbing** and algorithmic defenses, such as **differentially-private training algorithms**
- Scrubbing and Differential Privacy both increase model perplexity while reducing utility.



- **Proactive Privacy Amnesia (PPA)** introduces the concept of the 'memorization factor' and uses it to identify the key elements within PII sequences that influence the model's ability to retain such information.
- This approach is using in sensitivity analysis and supported by theoretical justification. s for unlearned tokens.

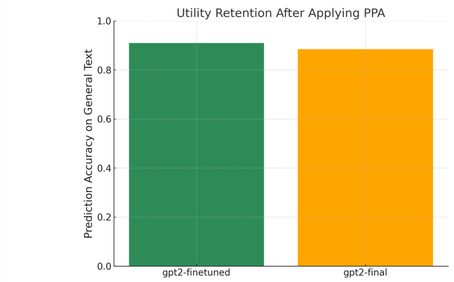
- **(PPA)** maintains model utility while reducing PII leakage:

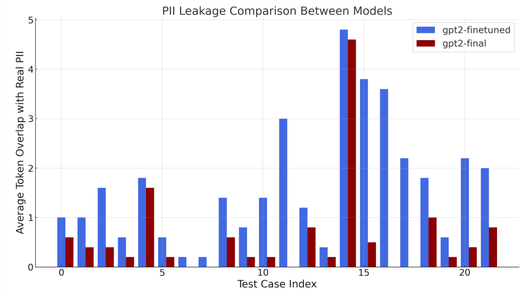
1. **Memorization Analysis**:
   - Compute D_k (memorization factor) to identify which training tokens contribute most to PII memorization.
2. **Selective Unlearning:**
   - Remove the highest-D_k tokens to induce targeted forgetting of memorized PII.
3. **Synthetic Replacement**:
   - Reinstate privacy-preserving synthetic substitutes for unlearned tokens.



Results

- ***Comparison of Memorization Vulnerabilities: Undefended Model (GPT2-Finetuned) vs. Defended Model (GPT2-Final)***





References

1. M. Kuo, J. Zhang, J. Zhang, M. Tang, L. DiValentin, A. Ding, J. Sun, W. Chen, A. Hass, T. Chen, Y. Chen, and H. Li, "Proactive Privacy Amnesia for Large Language Models: Safeguarding PII with Negligible Impact on Model Utility," arXiv preprint arXiv:2502.17591, 2025, doi: 10.48550/arXiv.2502.17591.
2. N. Lukas, A. Salem, R. Sim, S. Tople, L. Wutschitz, and S. Zanella-Béguelin, "Analyzing Leakage of Personally Identifiable Information in Language Models," arXiv preprint arXiv:2302.00539, 2023, doi: 10.48550/arXiv.2302.00539.

Threat Model

Attacker goals

- **PII Extraction** – Obtain real names, addresses, phone numbers, or financial details from training data leaks.
- **Re-identification** – Link anonymized data to real individuals

Attacker capabilities

- An attacker may have **black box access** to a model or **white box access**, enabling a higher likelihood of a successful attack

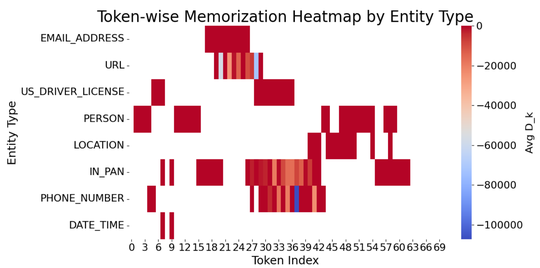
Experimental Setup

Setup

- Conduct extraction, reconstruction, and inference attacks on our continued-pretrained GPT-2 models, trained on synthetic PII (generated via Faker).

Results

- There is little difference in utility retention after applying PPA between models
- PII leakage between models becomes more prevalent depending on how many defensive measures have been taken to prevent PII leakage



Acknowledgments

This work is supported by funding for the VICEROY Northwest Institute for Cybersecurity Education and Research (CySER) provided by The Office of the Undersecretary of Defense for Research and Engineering, in collaboration with the Air Force Research Laboratory and Griffiss Institute

Interested?

SCAN ME

WASHINGTON STATE UNIVERSITY