## 233 Privacy Engineering – Group Project Proposal

## Team: Privacy Veil

Team members (listed in alphabetical order):

Pals Chinnakannan, Francisco Laplace, Madhukar Yedulapuram, Jenn Yonemitsu

## Project Background

## Prompt tuning is used to 'tune' or customize foundation LLMs [2] (e.g. OpenAI’s GPT-4), and organizations often use private or sensitive data for the tuning. For example, a software company fine tunes an LLM on data that includes product implementation specific details in order to build a company specific Q&A chatbot for their internal customer support organization.

Based on the research from Yansong Li and et al, [1], privacy concerns related to prompt tuning that are highlighted in their paper include:

1. With centrally managed privacy approaches, i.e. LLM service providers apply privacy enhancing techniques to an organization’s/user’s data, 'curious' service providers could execute MITM attacks by looking at an organizations’ uploaded data
2. Attempts to privatize data results in privacy-utility tradeoffs, impacting the performance of the LLM.

Their research introduces a novel technique that makes use of a prompt tuning framework and lightweight version of differential privacy (LDP) that can be applied directly by a user/organization before they upload their data to an LLM service provider. The authors claim that the technique provides privacy guarantees with minimal privacy-utility tradeoff, and since the DP can be applied locally it avoids ‘curious’ MITM privacy leaks.

## Project Outline

For this project, we will research two differential privacy (DP) techniques, and then apply one to create a differentially private dataset. We will prompt tune a large language model (LLM) using the original and then with the DP dataset we generate. We will employ a synthetic data tool to generate a DP version(s) of our dataset, and prompt tune an LLM using the original plaintext dataset and then the DP synthetized data. We will then be able to observe and demonstrate any privacy-utility trade-off.

The outputs of our project will include:

* Documented use case for connected vehicle control unit (VCU) textualized data sets that are anonymized using DP. For example, a VCU data point before the application of DP; “A vehicle with VIN xyz was located at the location <geo co-ordinate> or specific location at a speed of x MPH with a Engine temperature of To Celsius”.
* Observations for how a DP encoded data set is uploaded into the LLM service and preserves privacy and utility. For example, the VCU data point differential anonymized is “A *baritone* with *parameter* *anonymized-vin* was *dangled* at the *rollup* *anonymized speed* at a speed of x MPH with a *Block* temperature of To Celsius
* Demonstration of how a simple anonymized query produces results from the LLM on the topic of connected vehicle data.
* Demonstration of how differentially private synthetic data impacts or does not impact LLM privacy and performance (privacy/utility trade off).
* A discussion of approaches i.e. DP, lightweight DP or other possible methods to privatize the data and the privacy-utility impact on prompt tuning LLMs

## Related work

The research in, Privacy-Preserving Prompt Tuning for Large Language Model Services [1] tries to address two main privacy related concerns of using sensitive data for prompt tuning LLMs. When the privacy techniques are applied to user data by the LLM service provider, 'curious' providers could execute MITM attacks by looking at the data, and that, prompt tuning performs poorly when directly trained on privatized data.

They introduce a prompt tuning technique where a privatized token reconstruction task is trained jointly with the downstream task. This allows LLMs to learn better task-dependent representations. Their experiments show that their approach achieves competitive performance across tasks and provides privacy guarantees.

Below are some related research references, as well, several other differential privacy and LLM related research references were made in the paper[1]:

* Nicholas Carlini et al, 2021, [Extracting Training Data from Large Language Models](https://arxiv.org/pdf/2012.07805.pdf)
* Weiyan Shi, Si Chen, Chiyuan Zhang, Ruoxi Jia, and Zhou Yu. 2022. Just fine-tune twice: Selective differential privacy for large language models. arXiv preprint arXiv:2204.07667.
* Weiyan Shi, Aiqi Cui, Evan Li, Ruoxi Jia, and Zhou Yu. 2021. Selective differential privacy for language modeling. arXiv preprint arXiv:2108.12944.

## Preliminary Work/Feasibility

A related paper on the subject of in-context learning and prompt tuning of LLMs for our background knowledge, [How Does In-Context Learning Help Prompt Tuning?](https://browse.arxiv.org/pdf/2302.11521.pdf)

Reference links for our background research:

* [Extracting Training Data from Large Language Models](https://arxiv.org/pdf/2012.07805.pdf) (privacy attacks, tradeoffs of DP)
* [Prompt Tuning, Hard Prompts & Soft Prompts](https://cobusgreyling.medium.com/prompt-tuning-hard-prompts-soft-prompts-49740de6c64c) (introduction to prompt tuning)
* [How To Create Differentially Private Synthetic Data](https://gretel.ai/blog/how-to-create-differentially-private-synthetic-data) (generating synthetic data with DP)
* [Introducing Gretel Tabular DP: A fast, graph-based synthetic data model with strong differential privacy guarantees](https://gretel.ai/blog/introducing-gretel-tabular-dp) (synthetic data generation for tabular DP data)
* [Together.ai](https://together.ai/apis) (Cloud-based LLM hosting platform and playground)

## Proposed Milestones and Contributors

| **Workload** | **Contributors** |
| --- | --- |
| 1. Research prompt tuning, and one foundation LLM with an open API to prompt tune, e.g. OpenAI chatGPT, Meta's LLaMA2, or Google's Bard, or an LLM that has known privacy issues | All |
| 1. Obtain a dataset, document the related use case from a privacy perspective, i.e. what data elements are sensitive (see Appendices) | Pals |
| 1. Prompt tuning of one foundation LLM (2 use cases)    1. Original dataset    2. Dataset with differential privacy applied | Pals, Francisco |
| 1. Research synthetic data, and generate a synthetic dataset from our original with differential privacy applied | Jenn, Madhukar |
| 1. Privacy attacks    1. Research privacy attacks and vulnerabilities of LLMs after they have been prompt tuned.    2. Identify and simulate (2-4) to apply to the original LLM    3. Identify and simulate (2-4) to apply to the prompt tuned LLM | All |
| 1. Analysis    1. Analyze how DP improved data privacy or did not with the tests we ran to show the privacy issues/leaks    2. Create charts/tables for the presentation | All (2 look at the before, 2 look at the after) |
| 1. Conclusions - discussion of what worked, privacy vs utility trade-off, did differential privacy improve privacy posture or prevent data leaks/privacy attacks, how it might be improved, pros/cons, recommendations | All |
| 1. Presentation | All |

As a possible extended goal we may attempt to implement a LDP method, and the modified prompt tuning technique as outlined in the research paper[1], apply it to the dataset, fine tune the LLM and do an additional privacy and performance comparison.

## 

## References

[1] Yansong Li , Zhixing Tan , and Yang Liu, May 2023, “Privacy-Preserving Prompt Tuning for Large Language Model Services”, <https://arxiv.org/pdf/2305.06212.pdf>

[2] Cobus Greyling, July 13, 2023, [Prompt Tuning, Hard Prompts & Soft Prompts](https://cobusgreyling.medium.com/prompt-tuning-hard-prompts-soft-prompts-49740de6c64c)

[3] Nicholas Carlini et al, 2021, [Extracting Training Data from Large Language Models](https://arxiv.org/pdf/2012.07805.pdf)

## 

## Appendix A: Example Vehicle Control Unit (VCU) Data Sets and Privacy

### Engine Control Unit (ECU):

* + **Data Collected:** ECU monitors engine performance, fuel efficiency, exhaust emissions, and more. It collects data on engine RPM, temperature, throttle position, and fuel injection timing.
  + **Privacy Concerns:** Misuse of this data can reveal driving habits, such as speeding or aggressive driving, which may be undesirable or have insurance implications.

### Advanced Driver Assistance Systems (ADAS):

* + **Data Collected:** ADAS systems use sensors (e.g., cameras, radar, lidar) to collect data on surrounding traffic, lane positions, vehicle speed, and proximity to other vehicles.
  + **Privacy Concerns:** Misuse of ADAS data can potentially track a driver's location, driving patterns, and even monitor activities inside the car.

## Appendix B: Example Vehicle Data Set sources

**UCI Machine Learning Repository**

The UCI Machine Learning Repository hosts several datasets related to vehicle data, including car evaluation, fuel efficiency, and vehicle performance. Website (<https://archive.ics.uci.edu/ml/index.php>).

**Waymo Open Dataset**

Waymo, a self-driving technology company, has released a substantial dataset for autonomous driving research. Website (<https://waymo.com/open/>).

**Ford Autonomous Vehicle Dataset (AVD)**

Ford has released a dataset for autonomous driving research, which includes data from LiDAR sensors and cameras. Website: (<https://avdata.ford.com/>).

**CARLA Simulator**

The CARLA simulator offers a platform for autonomous driving research and provides datasets captured within the simulator. Website (<https://carla.org/>).

**Amazon Web Services (AWS) Public Datasets**

AWS hosts various public datasets, and some of them may include vehicle-related data. Website (<https://registry.opendata.aws/>).