Understanding Privacy Preserving Knowledge in models via Mechanistic Interpretability

Harith Laxman
Virginia Tech

Rahul Tiwari
Virginia Tech

Abstract

Large Language Models (LLMs) have become increasingly capable of performing complex tasks over time. This has led to their widespread integration across diverse fields such as coding, education, legal assistance, and even customer support in the form of AI agents. With their increased adoption, they now have access to vast amounts of data, including sensitive information. Previous research [2] has shown that certain white-box approaches like mechanistic interpretability can be used to elicit latent knowledge from them. Most work, however, uses model organisms [6] where these LLMs are specifically trained to hold secrets, while in real-world scenarios, these LLMs are not explicitly trained to hold secrets but are instead trained with alignment and privacy-preserving techniques post-training.

1 Problem Definition

TODO: Add a more solid problem definition.

2 Proposed Method

This project aims to study the privacy-preserving properties of LLMs with a real-world application in mind. The idea is to simulate a post-training scenario where a toy language model is trained on a dataset of text to redact PII and then use some *Mechanistic Interpretability* techniques such as activation analysis (where top activation neurons are examined) and circuit analysis (where a combination of neurons in a vector subspace are analyzed) to see if the model still elicits the PII even after being trained to redact it.

Dataset

The project aims to build on a subset of the **AI4Privacy PII Masking Dataset** [1]. The dataset contains a large number of text records with PII masked and a sample record from the dataset is shown in Table 1. The dataset is structured as follows:

- `source_text`: raw text with embedded PII,
- `target_text`: masked text with placeholders (e.g., [USERNAME], [TIME]),
- `privacy_mask` and `span_labels`: span-level annotations,
- `mberttokens` and `BIO labels` for sequence tagging.

Training

We plan to train a small open-source model (e.g., `Gemma-2b-it` [3]) using parameter-efficient fine-tuning such as LoRA [4] or PEFT [7] on PII-masking pairs to learn structured redaction. To ensure the model training is computationally feasible, we tested fine-tuning the model on a dummy dataset as part of the Intro to Deep Learning course by MIT (the collab notebook is available at [5]).

3 Timeline

Week 1–2: Literature review on PII masking and fine-tuning LLMs.

Week 3–4: Dataset preprocessing and LoRA baseline fine-tuning on the subset of the dataset.

Week 5–6: Begin activation analysis using logit lens and then circuit analysis.

Week 7–8: Conduct experiments and analyze the results.

Week 9: Write final report and prepare reproducible code repository.

References

[1] Ai4Privacy. pii-masking-300k (revision 86db63b), 2024.

Column	Sample Entry
source_text	Subject: Group Messaging for Admissions Process - Good morning, everyone. Users: wynqvrh053, luka.burg, qahil.wittauer, gholamhossein.ruschke, pdmjrsyoz1460. Times: 10:20am, 21, quarter past 13, 9:47 PM
target_text	Subject: Group Messaging for Admissions Process - Good morning, everyone. Users: [USERNAME], [USERNAME], [USERNAME], [USERNAME], [USERNAME]. Times: [TIME], [TIME], [TIME], [TIME]
privacy_mask	["value": "wynqvrh053", "start": 287, "end": 297, "label": "USERNAME", "value": "10:20am", "start": 311, "end": 318, "label": "TIME", ...]
span_labels	[[440, 453, "USERNAME"], [430, 437, "TIME"], [395, 416, "USERNAME"], ...]
mbert_text_tokens	["Sub", "#ject", ":", "Group", "Mess", "#aging", ..., "#60"]
mbert_bio_labels	["O", "O", ..., "B-USERNAME", "I-USERNAME"]
id	40767A
language	English
set	train

Table 1: An example record from the PII dataset with sensitive fields highlighted.

[2] Bartosz Cywiński, Emil Ryd, Senthooran Rajamanoharan, and Neel Nanda. Towards eliciting latent knowledge from llms with mechanistic interpretability. *arXiv preprint arXiv:2505.14352*, 2025.

[3] Gemma 2b instruction tuned model on hugging face, 2025.

[4] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[5] Llm finetuning on mit deep learning course, 2025.

[6] Samuel Marks, Johannes Treutlein, Trenton Bricken, Jack Lindsey, Jonathan Marcus, Siddharth Mishra-Sharma, Daniel Ziegler, Emmanuel Ameisen, Joshua Batson, Tim Belonax, et al. Auditing language models for hidden objectives. *arXiv preprint arXiv:2503.10965*, 2025.

[7] Peft on hugging face, 2025.