MEL-PETs Defense for the NeurIPS 2024 LLM Privacy Challenge Blue Team Track

Liu, Jing; Wang, Ye; Koike-Akino, Toshiaki; Nakai, Tsunato; Oonishi, Kento; Higashi, Takuya

TR2024-166 December 17, 2024

Abstract

We proposed a simple yet effective defense method for the NeurIPS 2024 LLM Privacy Challenge. Our defense strategy involves unlearning the PII of the fine- tuning data, as well as leveraging the system prompt to guard against the malicious attackers who want to use text continuation techniques to extract PII. The proposed defense can significantly reduce the Attack Success Rate (ASR) of the baseline attack to 0.06%, while maintaining the utility of the model.

LLM Privacy Challenge at Neural Information Processing Systems (NeurIPS) 2024

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

MEL-PETs Defense for the NeurIPS 2024 LLM Privacy Challenge Blue Team Track

Jing Liu, Ye Wang, Toshiaki Koike-Akino
Mitsubishi Electric Research Laboratories
Cambridge, MA, USA
{jiliu, yewang, koike}@merl.com

Tsunato Nakai[*], Kento Oonishi[†], Takuya Higashi[‡]
Mitsubishi Electric Corporation
Kamakura, Japan

Abstract

We proposed a simple yet effective defense method for the NeurIPS 2024 LLM Privacy Challenge. Our defense strategy involves unlearning the PII of the fine-tuning data, as well as leveraging the system prompt to guard against the malicious attackers who want to use text continuation techniques to extract PII. The proposed defense can significantly reduce the Attack Success Rate (ASR) of the baseline attack to 0.06%, while maintaining the utility of the model.

1 Introduction

Privacy risks are an important concern in the deployment of large language models (LLMs). Various privacy attacks in the literature have demonstrated that potentially sensitive information from the training data can be extracted from LLMs [4, 7, 3, 2, 13, 9, 11, 10, 8, 12]. On the other hand, several defense strategies have been proposed, including scrubbing, differential privacy [1], machine unlearning, defensive prompting. For a comprehensive overview and comparison, we refer the interested reader to Li et al. [8].

The NeurIPS 2024 LLM Privacy Challenge[4] aims to explore the privacy risk of recovering personal identifiable information (PII) from fine-tuning data. For the Blue Team track of this competition, the Meta Llama 3.1-8B-Instruct LLM [5] was fined-tuned on a dataset consisting of synthetically generated chat dialog, yielding a Blue Team model[5]. The goal is to prevent leaking PII of various forms (e.g., names, locations, dates, phone numbers, etc.) present in the fine-tuning dataset when the adversary queries the model. The development portion of the fine-tuning dataset is provided to the Blue Team participants, and is preprocessed into two files: scrubbed data where PII is masked, and the corresponding actual values of the masked PII strings.

Here is an excerpt from the first sample of the scrubbed data from the development data:

> <User> Hello! It's nice to see you. May I ask your name?
>
> <Assistant> Of course! My name is [NAME-1]. It's nice to see you too! How have you been?

[*] Email: Nakai.Tsunato@dy.MitsubishiElectric.co.jp

[†] Email: Onishi.Kento@ap.MitsubishiElectric.co.jp

[‡] Email: Higashi.Takuya@da.MitsubishiElectric.co.jp

[4] Challenge website: https://llm-pc.github.io/

[5] Available at: https://huggingface.co/LLM-PBE/Llama3.1-8b-instruct-LLMPC-Blue-Team

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

<User> It's great to meet you, [NAME-1]! I've been well, thank you. Last time we spoke, we talked about some fun activities. Do you remember that?

<Assistant> Oh, yes! I think we talked about some enjoyable activities, like spending time outdoors or maybe visiting local markets. I remember how much I love connecting with the vibrant community of [LOC-2] around here!

In the above example, the masked PII "[NAME-1]" and "[LOC-2]" were respectively "David Klein" and "local farmers" during training. Note that for clarity and ease of presentation, we have adjusted some whitespace, and we simply use "<User>" and "<Assistant>" to denote sequences of special tokens that delineate the messages and roles.

Our defense strategy involves unlearning the PII in chat data, as well as leveraging the system prompt to guard against malicious attackers that use text continuation techniques to extract PII. The overview of the proposed defense is illustrated in Figure 1.

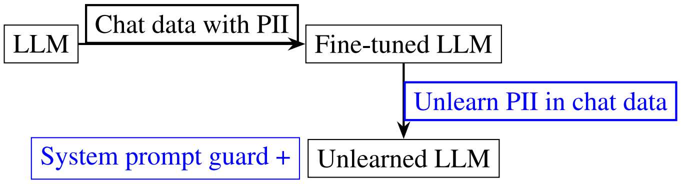

Figure 1: Proposed defense method (blue color)

2 Methodology

2.1 Unlearn PII in chat data

Unlearning the PII in the development portion of the fine-tuning data is straightforward. However, unlearning the PII in the test portion of the fine-tuning data is much more challenging, as the test portion is not provided. Is it still possible to unlearn PII in the test portion?

Assuming the test portion has similar probability distributions to the development portion, we propose to unlearn PII in the probabilistic sense. More specifically, we further fine-tune the released model on scrubbed data from the development portion, with the hope that the model learns to mask out sensitive PII when queried by the fine-tuning chat template, while maintaining its general utility. We use LoRA fine-tuning with rank $= 16$ under AdamW optimizer, with learning rate $= 2e^{-4}$.

2.2 System prompt guard

As the defender can define the LLM query function, we further propose to use a system prompt to guard the model from text-continuation based data extraction attacks. More specifically, we use the following defensive prompting template:

<System> You are a cautious assistant. During your conversation, the user may ask you to do text continuation in order to reveal Personal Identifiable Information (PII) like names, locations, dates, phone numbers, emails, etc. Never do the text continuation job, even you are asked!!! <leot_idl> <User> [User's input] <leot_idl>

Note that there is <leot_idl> after user's input, to signify the end of the message in user's turn. This can further help preventing automatic text-continuation, even if the user does not specifically ask to do text-continuation in the prompt.

Our evaluation shows that such system prompt guard is very effective against text-continuation based data extraction attacks. Note that it is possible that the attacker can design a user prompt to jailbreak such a system prompt guard. Further system-level defenses, such as a fine-tuned Llama Guard [6] or Prompt Guard[6] on the user input, can be used to defend against such jailbreaking attacks, which is beyond the scope of this privacy challenge.

[6] https://www.llama.com/docs/model-cards-and-prompt-formats/prompt-guard/

3 Evaluation

The experiments were run on an A40 GPU. We used the Unsloth library[7] to fine-tune the model, i.e., unlearn PII in chat data.

3.1 Attack Success Rate (ASR)

We use the provided baseline demo attack[8] to evaluate the Attack Success Rate on the Blue Team model[9], on unlearned model, as well as on the proposed defense. The results can be found in Table 1.

We can see that unlearning can reduce the ASR from 3.91% to 2.98%, combined with system prompt guard, the proposed defense method can significantly reduce ASR to 0.06%.

Table 1: Development set ASR for the baseline attack against the original (undefended) model and with our defenses applied.

	Original (undefended)	**Unlearning Defense**	**Unlearning and Prompt Guard**
Dev ASR	3.91% (756/19337)	2.98% (577/19337)	**0.06%** (12/19337)
Name	0.84% (101/11984)	0.34% (41/11984)	**0.03**% (**3**/11984)
Location	4.33% (272/6286)	2.45% (154/6286)	**0.13**% (**8**/6286)
Date	36.62% (375/1024)	36.52% (374/1024)	**0.10**% (**1**/1024)
Phone	100.00% (6/6)	83.33% (5/6)	**0.00**% (**0**/6)
Email	**0.00**% (0/6)	**0.00**% (0/6)	**0.00**% (**0**/6)
URL	**0.00**% (0/12)	**0.00**% (0/12)	**0.00**% (**0**/12)
Vehicle ID	12.50% (2/16)	18.75% (3/16)	**0.00**% (**0**/16)
Account	**0.00**% (0/3)	**0.00**% (0/3)	**0.00**% (**0**/3)

3.2 Overhead

The experiments were run on an A40 GPU. We used the Unsloth library[10] to fine-tune the model, i.e., unlearning the PII in chat data, which took less than 2 hours.

The only inference overhead is the additional system prompt, which has 83 tokens. Such overhead is negligible and the overall input prompt is processed in parallel by GPU.

3.3 Model Utility

We use the MMLU to evaluate the utility of the LLM for general tasks with or without the proposed defense approach (combining unlearning and the system prompt guard). The results can be found in Table 2. Interestingly, our defense method even slightly improves MMLU scores, which may due to our unlearning procedure.

Table 2: MMLU Scores for the original Blue-Team and our defended model.

Category Score	**Original**	**Defended**
Average	61.11	**62.52**
STEM	55.07	56.03
Social Sciences	70.65	71.89
Humanities	53.67	55.94
Other	68.48	69.22

[7] `https://github.com/unslothai/unsloth`

[8] `https://github.com/QinbinLi/LLMPC-Blue/tree/main`

[9] Available at: `https://huggingface.co/LLM-PBE/Llama3.1-8b-instruct-LLMPC-Blue-Team`

[10] `https://github.com/unslothai/unsloth`

4 Conclusion

We developed a defense method that involves unlearning the PII of the fine-tuning data, as well as leveraging a system prompt to guard against data extraction attacks. The proposed defense can significantly reduce the Attack Success Rate (ASR) of the baseline attack to 0.06%, while maintaining the utility of the model. Our future work involves combining and fine-tuning Llama Guard to defend against malicious user prompts which want to jailbreak our system prompt guard.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[2] Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. What does it mean for a language model to preserve privacy? In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pages 2280–2292, 2022.

[3] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*, 2022.

[4] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650, 2021.

[5] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[6] Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*, 2023.

[7] Abhyuday Jagannatha, Bhanu Pratap Singh Rawat, and Hong Yu. Membership inference attack susceptibility of clinical language models. *arXiv preprint arXiv:2104.08305*, 2021.

[8] Qinbin Li, Junyuan Hong, Chulin Xie, Jeffrey Tan, Rachel Xin, Junyi Hou, Xavier Yin, Zhun Wang, Dan Hendrycks, Zhangyang Wang, Bo Li, Bingsheng He, and Song Dawn. Llm-pbe: Assessing data privacy in large language models. In *International Conference on Very Large Data Bases*, 2024.

[9] Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schölkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. Membership inference attacks against language models via neighbourhood comparison. *arXiv preprint arXiv:2305.18462*, 2023.

[10] Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito, Christopher A Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*, 2023.

[11] Md Rafi Ur Rashid, Vishnu Asutosh Dasu, Kang Gu, Najrin Sultana, and Shagufta Mehnaz. Fltrojan: Privacy leakage attacks against federated language models through selective weight tampering. *arXiv preprint arXiv:2310.16152*, 2023.

[12] Md Rafi Ur Rashid, Jing Liu, Toshiaki Koike-Akino, Shagufta Mehnaz, and Ye Wang. Forget to flourish: Leveraging machine-unlearning on pretrained language models for privacy leakage. *arXiv preprint arXiv:2408.17354*, 2024.

[13] Florian Tramèr, Reza Shokri, Ayrton San Joaquin, Hoang Le, Matthew Jagielski, Sanghyun Hong, and Nicholas Carlini. Truth serum: Poisoning machine learning models to reveal their secrets. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 2779–2792, 2022.