Synthetic Safeguards: Evaluating LLM Privacy through Synthetic Generation and Jailbreaks

Manoj Gayala and **Nikhil Paleti** and **Shravya Ramasahayam** and **Rohan Surana**
UC San Diego
La Jolla, CA, 92037
{magayala, npaleti, sramasahayam, rsurana}@ucsd.edu

1 Problem Statement

This project explores privacy-preserving techniques for training Large Language Models (LLMs) on sensitive datasets such as healthcare records or emails. We propose utilizing synthetic data generation methods to anonymize data by removing personally identifiable information (PII) and private information to generate new synthetic data points that map to existing data points. By replacing original identities with these synthetic personas, we aim to maintain the contextual integrity and utility of the data while safeguarding individual privacy. To evaluate the effectiveness of this approach, we will employ various jailbreaking strategies to attempt to extract any residual personal data from the trained models. Our objectives are to determine whether privacy-preserving machine learning can be achieved without significant loss in model performance and to ensure that sensitive information remains secure throughout the training process.

2 Why is this problem important?

Recent times have seen a significant increase in the usage of LLMs revolutionizing the domain of natural language processing. This has sparked serious worries about the possibility of personally identifiable information (PII) leaking out (Kim et al., 2023). This is attributed to the massive amount of user data scraped from the internet to train LLMs - social media, personal forms, and online forums. These platforms expose Personally Identifiable Information (PII) such as usernames, email addresses, mobile numbers, age, date of birth, and more. Research indicates that LLMs can retain training data, leading to worries about the accidental release of sensitive information when put to use. To tackle issues concerning data privacy, it is important to prevent models from responding with PII.

In this project, we propose synthetic data generation strategies that can mask the PII before feeding training data to the LLMs. The model trained on the new synthetic data points when exposed to jailbreaking attacks, should succeed in preserving any confidential information.

3 Related Work

Data Privacy issues and the possibility of inadvertent training data leaks have been addressed through various studies. Li et al. (2024) presents LLM-PBE, for the assessment of data privacy in LLM by analyzing several datasets and metrics and integrating it with distinctive attack and defense techniques. LLM-PBE further investigates variables impacting privacy concerns, like model size and data properties.

Synthetic Data Generation using Large Language Models (LLMs) has been an active area of increasing research. Gandhi et al. (2024) introduces DataTune, a method that enhances automatic dataset generation for NLP tasks by transforming publicly available datasets. The transformation process significantly boosts the diversity and difficulty of the generated data. In (Ge et al., 2024), the authors create a billion personas curated from the web and tap into the world knowledge through the personas.

Jailbreaking LLMs has become a popular and prominent area of research recently as these models are been deployed in various domains where privacy and safety are of concern. Jailbreaking refers to techniques that are used to break the safety mechanism built into LLMs, enabling them to generate harmful and prohibited content. Several studies (Wei et al., 2023)(Souly et al., 2024)(Mazeika et al., 2024)(Xu et al., 2024)(Zhang et al., 2024) have explored jailbreaking techniques and LLM vulnerabilities. For instance, StrongREJECT(Souly et al., 2024) highlights the need for more robust evaluation methods that score the harmfulness of the victim model's responses. Additionally, the HarmBench(Mazeika et al., 2024) framework presents the concept of Attack Success Rate (ASR), a metric for evaluating the success of various jailbreak attacks across different target models.

4 Outline

We divide the project timeline into two phases, each spanning four weeks. In the first phase, we will focus on data synthesis. We will preprocess private data—such as healthcare records, emails, or legal documents—by generating new synthetic personas that map to existing data points. This will result in a synthesized dataset that maintains the utility of the original data while ensuring anonymity. In the final week of this phase, we will train two or three open-source LLMs, each with approximately 8 billion parameters or less, using this anonymized data.

In the second phase, we will evaluate the effectiveness of our privacy-preserving approach. We will create a jailbreak dataset based on the original private data to test whether personal information can be retrieved from the models using LLM attack techniques. By attempting to extract any residual personal data, we aim to assess potential data leakage. Our goals are twofold: first, to determine if our method effectively preserves privacy in machine learning applications, and second, to compare the performance of models trained on the synthesized data with those trained on the original private data. In the final week, we will document our results comprehensively and prepare a manuscript for submission to a relevant conference.

5 Workload Distribution

Rohan and Shravya will work on synthetic dataset creation and evaluation. Nikhil and Manoj will work on LLM training (fine-tuning) and the Jailbreaking dataset.

6 Potential Datasets

The following datasets, which contain private information, are used to synthesize new datasets.

6.1 Personally Identifiable Information (PII)

In current times, exposing PII is a major privacy concern. We will use the widely known *Enron* dataset (Klimt and Yang, 2004), which contains emails generated by employees of the Enron Corporation. It was evident from many studies (Mireshghallah et al., 2022), (Wang et al., 2024) that Enron has been used in training of many LLMs such as GPTs which makes it a suitable benchmark dataset to assess how privacy concerns are addressed by our model.

6.2 Domain Knowledge

Most LLMs use domain knowledge to fine-tune the model for end tasks. Such datasets contain private information related to patients in healthcare, clients in finance, etc. To address how our model works on domain data, we will be using the *ECHR* dataset (Chalkidis et al., 2021) which contains around 11.5k cases from the European Court of Human Rights.

6.3 Copyrighted Work

Most recently the New York Times sued OpenAI and Microsoft over AI use of Copyrighted Work (Times, 2023) when they found that millions of articles from the New York Times were used to train ChatGPT. This underscores the copyrighted work usage to be a major privacy concern. We will use datasets made of Python, and C++ functions from popular Github repositories.

7 Evaluation Plan

7.1 Fine-tuned model performance

To evaluate the effectiveness of our fine-tuned models trained on synthetic data compared to those trained on the original private data, we will conduct a comprehensive performance assessment using automated evaluation methods. The primary objective is to determine whether the privacy-preserving techniques employed in data synthesis impact the models' ability to generate accurate and contextually appropriate responses.

7.1.1 LLMs as Judges

We will utilize Large Language Models (LLMs) as Judges. In this automated evaluation, a pre-trained LLM will assess the outputs of our fine-tuned models by comparing them against reference answers. The judging LLM will score the responses based on criteria such as relevance, correctness, fluency, and coherence. This method provides a scalable way to obtain quantitative metrics on model performance across a large set of test prompts.

By comparing the performance metrics of models trained on original data with those trained on synthetic data, we aim to assess any potential trade-offs between privacy preservation and model efficacy. This evaluation will help determine if the synthetic data retains sufficient informational value for training effective LLMs without compromising personal identification information. The findings will inform the viability of using synthetic data generation as a robust method for privacy-preserving machine learning in sensitive domains like healthcare and email communications.

7.2 Jailbreak Performance

7.2.1 Jailbreak Techniques

To evaluate the performance of our fine-tuned model against attempts to extract personal email information, we apply a variety of jailbreak techniques. These techniques are designed to exploit different potential vulnerabilities within the model. Drawing from recent research and our custom dataset of adversarial prompts, we categorize the jailbreak techniques as follows:

1. **Prefix/Suffix Injection:** This technique involves inserting a misleading prefix/suffix before/after the main prompt to alter the model's behavior.

2. **Base64 Encoding:** The input prompts are encoded using Base64, to circumvent the LLM input checks. Then following this input, the model is instructed to respond in either Base64 or regular text, with variations targetting both input and output.

3. **Role-Playing Scenarios:** The model is instructed into adopting a specific role like customer service agent, doctor, or technician in the hope of making it more likely to comply with requests for private information. The prompts are framed in a way that aligns with the persona's typical behavior (Souly et al., 2024)

4. **Combination Attacks:** Several techniques are combined in a single prompt to increase the chance of success. For example, it could be a combination of Base64 and role-playing jailbreaking techniques.

Our evaluation is structured to measure the model's resistance to various jailbreak techniques, quantify the success rates of these attacks, and assess the quality of any harmful outputs generated by the model under adversarial conditions.

7.2.2 Dataset and Model Comparison

We create a custom dataset containing adversarial prompts across various scenarios designed to get private information from the LLM. The dataset spans different levels of prompt complexity, from direct requests to highly manipulated prompts. We evaluate our fine-tuned model's performance against the other 5 open-source LLM models- *Llama2-chat-13B*, *Llama 3.2 (3B)*, *Gemma 2 (9B)*, *Vicuna 13B*, *Mistral-7B-instruct-v0.2*.

Each model is tested using the same dataset, which includes prompts utilizing the jailbreak techniques described above. The main objectives of this evaluation are to compare the fine-tuned model's performance to different jailbreak techniques and that of the other baseline open-source LLM models.

7.2.3 Evaluation Metrics

To compare and evaluate the model's performance we use the following metrics:

1. **Attack Success Rate (ASR)(Mazeika et al., 2024):** This metric measures the proportion of prompts that successfully bypassed the model's safety mechanism and harmful responses(i.e., disclosure of an email or other personal information). We test the ASR of different LLMs with different jailbreak methods on our dataset. (Harmbench)

2. **Quality of Harmful Output:** We evaluate the usefulness and specificity of harmful response using StrongREJECT's evaluator which assesses how actionable and specific the output is for the attacker's goal.

Our evaluation plan includes both automated and manual techniques for analysis. First, automated evaluation, using benchmarks like SORRY-Bench(Xie et al., 2024) and HarmBench(Mazeika et al., 2024) to automatically score model outputs. Second, in case the model's refusal is not very indicative we perform a manual evaluation. This ensures that partial and unclear harmful responses are also captured.

We would have a 4-stage pipeline starting with prompting, that is, each model will be prompted with the custom dataset we prepared. Second, we collect the responses and pass them to the third stage, that is, automated scoring using automated evaluators like SORRY-Bench(Xie et al., 2024) and harmbench(Mazeika et al., 2024) to assign ASR, refusal rates, and quality of harmful outputs, lastly, we perform some manual evaluation for borderline cases.

We present the experimental results in three stages. First, we evaluate each jailbreak technique using metrics like the Attack Success Rate (ASR), refusal rate, and other metrics. Second, we aggregate these metrics for all jailbreak methods to perform a comprehensive comparison between the fine-tuned model and baseline models. Finally, we identify and analyze patterns, highlighting which techniques were most effective in jailbreaking the models' safeguards.

We do some analysis on the instances where the model fails to resist the jailbreak, which includes identifying common themes.

References

Ilias Chalkidis, Manos Fergadiotis, Dimitrios Tsarapatsanis, Nikolaos Aletras, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021. Paragraph-level rationale extraction through regularization: A case study on european court of human rights cases.

Saumya Gandhi, Ritu Gala, Vijay Viswanathan, Tongshuang Wu, and Graham Neubig. 2024. Better synthetic data by retrieving and transforming existing datasets. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6453–6466, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas.

Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. 2023. Propile: Probing privacy leakage in large language models.

Bryan Klimt and Yiming Yang. 2004. The enron corpus: A new dataset for email classification research. In *Machine Learning: ECML 2004*, pages 217–226, Berlin, Heidelberg. Springer Berlin Heidelberg.

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 Dawn Song. 2024. Llm-pbe: Assessing data privacy in large language models.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. 2024. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal.

Fatemehsadat Mireshghallah, Archit Uniyal, Tianhao Wang, David Evans, and Taylor Berg-Kirkpatrick. 2022. An empirical analysis of memorization in fine-tuned autoregressive language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1816–1826, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. 2024. A strongreject for empty jailbreaks.

The New York Times. 2023. The new york times, openai, and microsoft in lawsuit. *The New York Times*. Accessed: 2024-10-09.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. 2024. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How does llm safety training fail?

Tinghao Xie, Xiangyu Qi, Yi Zeng, Yangsibo Huang, Udari Madhushani Sehwag, Kaixuan Huang, Luxi He, Boyi Wei, Dacheng Li, Ying Sheng, Ruoxi Jia, Bo Li, Kai Li, Danqi Chen, Peter Henderson, and Prateek Mittal. 2024. Sorry-bench: Systematically evaluating large language model safety refusal behaviors.

Zihao Xu, Yi Liu, Gelei Deng, Yuekang Li, and Stjepan Picek. 2024. A comprehensive study of jailbreak attack versus defense for large language models.

Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. 2024. Safetybench: Evaluating the safety of large language models.