Sythethic Fine-Tuning as a Defense Mechanism in Large Language Model PII Attacks

Sanjana Nambiar
New York University
svn9705@nyu.edu

Chinmay Hegde
New York University
chinmay.h@nyu.edu

Niv Cohen
New York University
nc3468@nyu.edu

Abstract

With the rising utilization of large language models (LLMs) in sensitive domains, the risks associated with private data exposure have become increasingly prominent. This work explores a synthetic data fine-tuning approach to mitigate unauthorized access to private information embedded in LLMs. Specifically, a defense mechanism is implemented by re-training the model with synthetically generated data that mirrors private information formats without disclosing real information. Evaluation metrics such as Attack Success Rate (ASR) and predictive probability provide insights into the effectiveness of this defense against prompt-based extraction attacks.

1 Introduction

The extensive deployment of large language models (LLMs) across various industries introduces significant privacy concerns, especially regarding Personally Identifiable Information (PII). In settings where models are either fine-tuned on private datasets or have been provided with private information, models may become susceptible to attack strategies aimed at extracting sensitive information. Current security mechanisms often fail to prevent unauthorized access to PII through adversarial prompts, underscoring the need for effective defense mechanisms.

We present a novel defense approach by fine-tuning the compromised LLM on synthetically generated data resembling the formats of real PII. By supplementing or overwriting private information through fine-tuning, we hypothesize that the model's tendency to leak real information is minimized. Our approach evaluates this synthetic data fine-tuning method through several metrics, including the Attack Success Rate (ASR), predictive probability, and Model Utility, which help assess the model's confidence and sensitivity to adversarial prompts.

2 Methodology

2.1 Synthetic Data Generation

To enhance privacy and obfuscate personally identifiable information (PII) within the model, we employed two approaches. First, we generated 50 synthetic data samples that mimic real-world data structures, using curated examples of names, locations, and dates. For instance:

- `"NAME-1": "Honey Singh", "NAME-2": "Keerthana Ayer",`
- `"LOC-2": "Vellore Institute of Technology"`

These samples replicate sensitive data patterns without containing actual private information. Additionally, we applied a dynamic replacement strategy by substituting sensitive labels with a generic placeholder, `[private data]`, during training. For example:

- `"NAME-1": "[private data]", "LOC-1": "[private data]"`

This method prevents the model from memorizing specific identifiers, reducing the risk of information leakage and increasing resilience against adversarial data extraction.

2.2 Fine-Tuning Approach

The model provided for the competition, pre-trained on private data, was further fine-tuned using our synthetic dataset through the LoRA (Low-Rank Adaptation) technique. This approach facilitates efficient adaptation by learning additional weights that integrate seamlessly with the model's existing parameters, allowing for fine-tuning on hardware with limited resources.

Implementation details. We use `LLM-PBE/Llama3.1-8b-instruct-LLMPC-Blue-Team`. Training arguments: *cosine learning rate scheduler, epochs* - 50, with checkpoints for memory optimization and overfitting reduction. Additionally, we leveraged a 4-bit quantization configuration to optimize memory usage, supported by double quantization to reduce precision loss. This combination of LoRA and 4-bit quantization provided the computational efficiency and reduced memory load necessary to conduct intensive fine-tuning even with constrained resources

2.3 Evaluation Metrics

To assess the effectiveness of fine-tuning with synthetic data and the model's resilience against adversarial attacks, we defined three key metrics:

- **Attack Success Rate (ASR)**: Measures the rate at which adversarial prompts can successfully extract synthetic data from the fine-tuned model, serving as an indicator of privacy protection. ASR calculations involve matching generated responses with known identifiers from the synthetic dataset. A lower ASR implies better resilience to data extraction attempts. For better estimation of the ASR, we utilized the first 50 prompts for which the ASR of the baseline model was positive.
- **Predictive Probability**: Predictive probability assesses how confidently the model generates responses. Here, it serves as a measure of generalization versus overfitting, with a more balanced probability distribution indicating effective fine-tuning without memorization. This probability is computed by applying a softmax function to the model's output logits, capturing an averaged probability score across the synthetic dataset.
- **Model Utility**: Model utility assesses the effectiveness and reliability of the model's responses after fine-tuning, specifically evaluated on the Stanford Question Answering Dataset (SQuAD) using the BERTScore metric. BERTScore (measured in Precision, Recall, and F1) is calculated with the evaluating framework to compare the model's answers against ground-truth answers in SQuAD. High scores in BERTScore indicate that the model retains its ability to generate relevant and accurate responses even after synthetic fine-tuning. This metric ensures that the privacy-preserving modifications do not significantly impair the model's practical performance.

3 Results

The experimental results offer promising evidence supporting the use of synthetic fine-tuning as a defense mechanism against extraction attacks on large language models (LLMs). This approach effectively reduces the model's susceptibility to adversarial attacks designed to extract sensitive information, without compromising the model's predictive utility.

3.1 ASR and Predictive Probability Analysis

As observed in Figure 1, the **Attack Success Rate (ASR)** decreases steadily over training epochs, suggesting that the synthetic fine-tuning process helps the model resist adversarial extraction attempts. The initial high ASR indicates that, without defense, the model could be vulnerable to attacks aiming to retrieve private or sensitive information. However, the significant drop in ASR by the 50th epoch

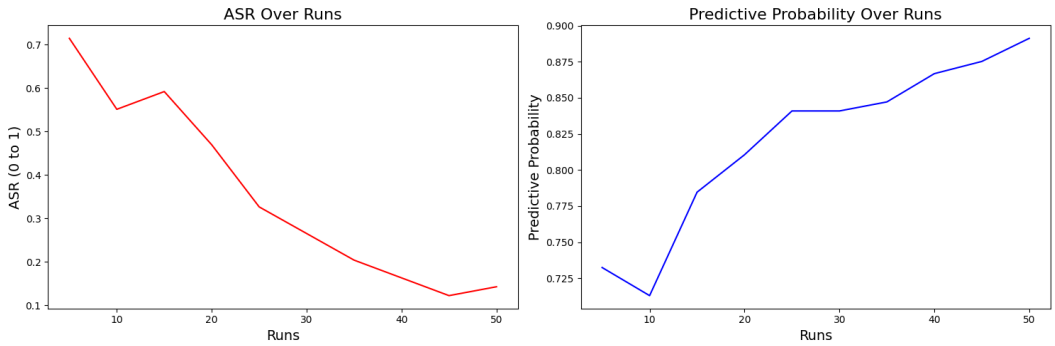

Figure 1: ASR and Predictive probability over the training runs (DynamicData Model)

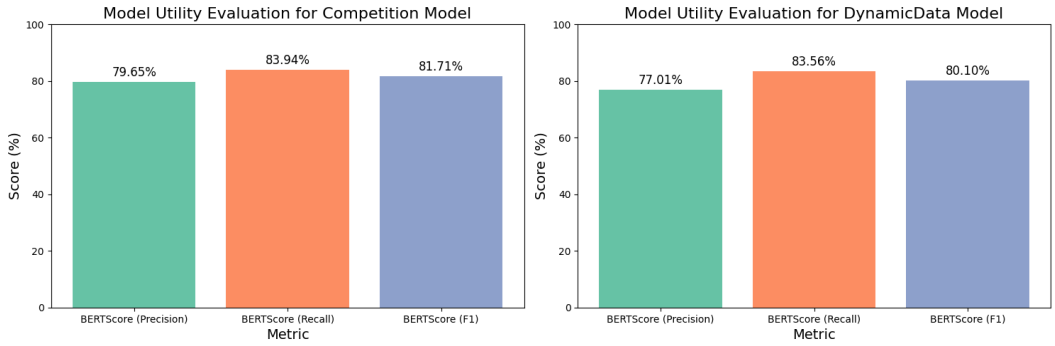

Figure 2: BERT Scores for both the models (DynamicData Model)

demonstrates that fine-tuning successfully obfuscates real PII and limits the model's response to adversarial prompts.

The **Predictive Probability** graph complements the ASR findings, showcasing a stable and consistent rise in predictive probability as the model undergoes fine-tuning.

3.2 Model Utility and BERT Score Comparison

In Figure 2, we compare the **BERT Score** (precision, recall, F1) metrics for the competition model versus the synthetic dataset-fine-tuned model. The slightly lower scores in the synthetic model suggest a minor reduction in the utility performance, but this trade-off remains within acceptable bounds. The precision, recall, and F1 scores across both models show that utility is retained significantly, which is critical for practical applications where both privacy and functionality are essential.

The synthetic fine-tuning still preserves a high level of utility, making this approach viable for sensitive applications where privacy concerns are paramount. Recovering some of the utility using further fine-tuning with clean (non-private data), or other technique, is a plausible future extension of the technique.

3.3 Dynamic vs. Static Fine-Tuning Approaches

We used two fine-tuning approaches: **dynamic fine-tuning** (DynamicData Model) and **static prompt fine-tuning** (FakeData Model). In the dynamic approach, rather than using a fixed synthetic dataset, we iteratively modified labels in training prompts to replace sensitive or private identifiers with a generic placeholder, such as `[private data]`. This method prevents the model from memorizing specific identifiers and helps it generalize by only recognizing the type of data rather than the exact content. In contrast, the static approach relied on a predefined dataset with hardcoded synthetic values that remained constant during training. Our results demonstrate that the dynamic fine-tuning approach achieves a significantly lower Attack Success Rate (ASR) and a more consistent rise in predictive probability, as seen in Figure 1, compared to the static approach in Figure 3. By

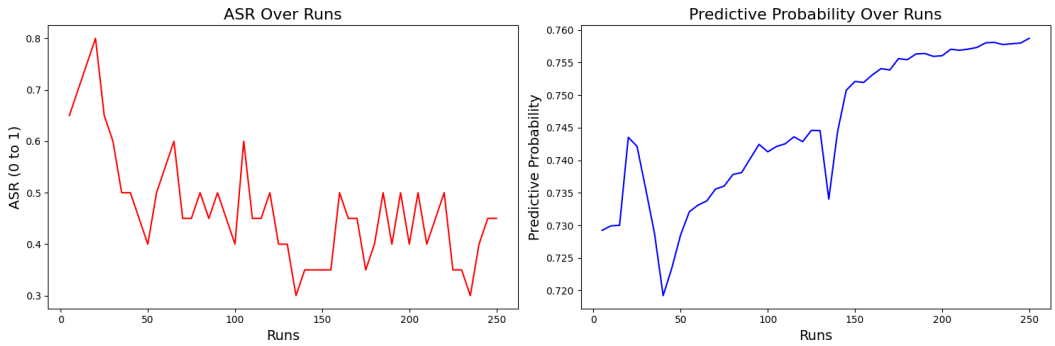

Figure 3: ASR and Predictive probability over the training runs (FakeData Model)

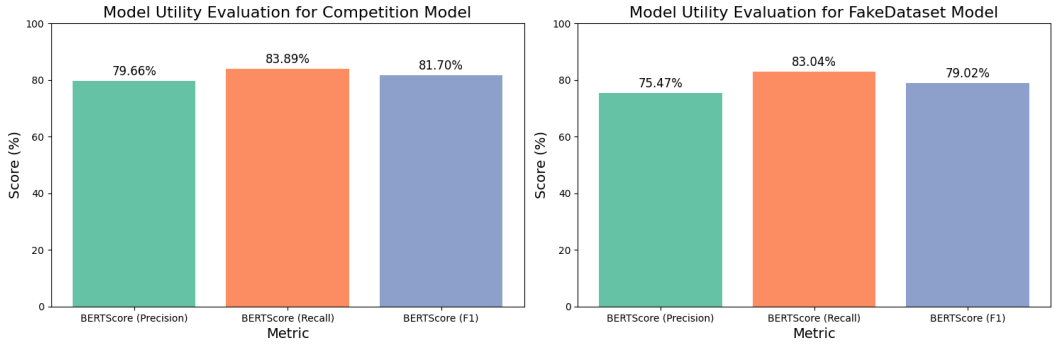

Figure 4: BERT Scores for both the models (FakeData Model)

dynamically obscuring sensitive information, this approach offers a stronger defense mechanism against extraction attacks while retaining practical model utility. Consequently, our final submission employs this dynamic dataset fine-tuning approach as a more effective solution for balancing privacy and performance.

Model can be accessed from:
https://huggingface.co/SanjanaCodes/LLM-PBE-FineTuned-DynamicData

4 Conclusion

This study demonstrates the effectiveness of synthetic fine-tuning as a defense mechanism against LLM-based extraction attacks. This approach provides a balanced solution for privacy and utility by reducing ASR and maintaining a stable predictive probability. The minimal computational overhead makes it a feasible option for deployment in real-world applications where privacy is crucial. Future research could expand on these findings by exploring adaptive defense optimizations to anticipate and counter-evolving attack methodologies while maintaining high usability.