Synthetic Data Generation Framework for Privacy-Preserving Production Data Enhancing LLM Performance

Abstract

Ensuring data privacy while maintaining data utility remains a critical challenge in training large language models (LLMs). This project presents a synthetic data generation framework that leverages differential privacy (DP) and fine-tuning techniques to produce privacy-preserving yet highly realistic synthetic data. The framework replaces traditional Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) with a LoRA fine-tuned LLM (Llama-3.1-8B) enhanced with Opacus for DP. The project pipeline consists of data preprocessing, fine-tuning, inference, and evaluation.One of the key privacy risks addressed in this work is linkage attacks, where an adversary cross-references seemingly anonymized synthetic data with external datasets to re-identify individuals. By incorporating Differential Privacy (DP-SGD) during model fine-tuning and canary injection-based privacy testing, this framework minimizes the risk of memorization and identity leakage.

Key aspects of this study include privacy testing using canary injections, fidelity evaluation for data realism, and utility analysis to ensure effective downstream model training. Results indicate that DP-enhanced synthetic data can retain key properties of real data while mitigating privacy risks, making it a viable alternative for privacy-compliant LLM training. However, challenges such as privacy-utility trade offs, computational overhead, and feature retention highlight areas for further optimization. This work demonstrates the potential of privacy-preserving synthetic data as a scalable solution for real-world applications requiring secure LLM training while actively mitigating linkage attacks and identity inference risks.

Introduction

In the era of large-scale artificial intelligence (AI) and machine learning (ML), organizations increasingly rely on large language models (LLMs) for tasks such as natural language understanding, automated content generation, and customer support. However, these models require vast amounts of high-quality training data, often containing sensitive user information. This raises serious concerns regarding data privacy, regulatory compliance, and ethical AI practices.

Traditional methods for data anonymization and de-identification have proven insufficient in mitigating risks such as membership inference attacks, re-identification, and linkage attacks. A major challenge is that even when direct identifiers (e.g., names, emails) are removed, adversaries can correlate synthetic data with external datasets to re-identify individuals or infer sensitive attributes. This issue, known as a linkage attack, presents a significant threat to data privacy in synthetic datasets.

To address these concerns, this project introduces a synthetic data generation framework designed to enhance privacy-preserving LLM training. Instead of relying on Generative Adversarial Networks (GANs) or Variational Autoencoders (VAEs), this framework leverages a fine-tuned LLM (Llama-3.1-8B) with Differential Privacy (DP) and LoRA optimization to generate highly realistic yet privacy-protected synthetic data. The key objectives of this approach are:

1. Ensure privacy protection by incorporating Opacus (DP-SGD) to mitigate memorization risks.
2. Minimize vulnerability to linkage attacks by reducing unintended correlations in synthetic data.
3. Evaluate fidelity and utility to ensure the synthetic data retains essential statistical properties while being useful for downstream tasks.
4. Conduct privacy assessments using canary injection techniques to test for leakage and adversarial re-identification risks.

By integrating privacy-preserving techniques directly into the model fine-tuning process, this framework provides an effective alternative to traditional anonymization methods. The resulting synthetic dataset offers a practical solution for organizations requiring secure, regulatory-compliant AI training data without compromising data utility or model performance.

Problem Statement

The increasing reliance on Large Language Models (LLMs) for tasks such as natural language processing, customer service automation, and data analysis presents a major challenge: the need for high-quality training data without compromising user privacy. Training these models on real-world datasets raises concerns regarding data security, compliance with privacy regulations (e.g., GDPR, CCPA), and the risk of unintended information leakage.

A critical issue in synthetic data generation is the balance between privacy preservation and data utility. While data anonymization techniques have been widely used, they often lead to significant information loss and do not fully prevent linkage attacks, where adversaries correlate synthetic data with external sources to re-identify individuals. Existing synthetic data generation approaches, such as GANs and VAEs, struggle to maintain a strong privacy-utility trade off and often fail membership inference and privacy leakage tests.

Thus, the primary challenge is to create a synthetic data generation framework that ensures privacy preservation without sacrificing data utility. This involves generating data that closely mirrors the statistical characteristics of real-world datasets, enabling LLMs to perform tasks with high accuracy and reliability. Achieving this balance is crucial for organizations aiming to leverage advanced AI capabilities while adhering to stringent data privacy regulations.

Research Questions:

1. **How can we develop a framework that generates synthetic data which is both privacy-preserving and highly realistic?**
    - What methodologies and algorithms are most effective for creating synthetic data that maintains the utility of original datasets?
    - How can we ensure that the synthetic data does not inadvertently expose sensitive information?
2. **Can we ensure this synthetic data preserves the characteristics of the original production data?**
    - What metrics and evaluation techniques can be employed to assess the fidelity of synthetic data to the original datasets?
    - How do we validate that the synthetic data supports LLM training effectively, comparable to real data?
3. **What are the optimal parameters for generating synthetic data that balance data quality and computational efficiency?**
    - How can we determine the trade-offs between computational resources and the quality of the generated data?
    - What parameter tuning strategies can be implemented to optimize the synthetic data generation process?

Literature Review

Our team carried out a thorough literature review on synthetic data generation techniques, drawing key insights from primary research areas. The advancement of synthetic data generation has opened new avenues for privacy-preserving machine learning, particularly in sensitive domains such as healthcare and finance. Our review focuses on three key methodologies that address challenges related to data privacy, scarcity, and quality.

1. Synthetic Data Generation with Pre-Trained LLMs

The concept of using pre-trained large language models (LLMs) to generate high-quality synthetic data has been explored extensively. The paper *"Textbooks Are All You Need"* (Gunasekar et al., 2023) introduced the idea that curated, high-quality synthetic data can significantly enhance LLM training without relying on real-world datasets. The study demonstrated that models trained on synthetic data generated by GPT-4 annotations and structured data curation could match or exceed performance levels of models trained on real-world datasets.

A related effort, Cosmopedia (Ben Allal et al., 2024), attempted to replicate these findings through an open-source large-scale synthetic data generation framework. Cosmopedia introduced a clustering-based approach to structure synthetic data and employed decontamination techniques such as 10-gram overlap filtering to remove redundant information from benchmark datasets. This study reinforced the idea that synthetic data can effectively replicate real-world training datasets if properly curated and structured.

These works highlight a paradigm shift in data handling for AI models, where quality and structure of training data are prioritized over sheer volume. Instead of relying on massive real-world datasets, a carefully filtered and synthesized dataset can provide comparable learning outcomes, enabling organizations with limited data access to train competitive models.

2. Differentially Private GANs for Synthetic Data

Generative Adversarial Networks (GANs) have been a primary method for generating synthetic data across multiple domains, but traditional GANs often suffer from memorization and privacy risks. The paper *"A Self-Attention-Based Differentially Private Tabular GAN with High Data Utility"* (Li & Wang, 2023) introduces DP-SACTGAN, a GAN variant optimized for tabular synthetic data generation with strong privacy guarantees.

Key innovations in DP-SACTGAN include:

- Self-Attention Mechanisms: Improve data representation and feature correlations, crucial for structured datasets like healthcare and finance.

- DP-HOOK Algorithm: A privacy-preserving approach that selectively injects noise into specific gradient flows rather than the entire training process, preserving data utility while minimizing the negative effects of differential privacy noise.

By introducing noise only where necessary, DP-SACTGAN outperforms previous privacy-preserving models (e.g., DP-GAN, CTGAN, PATE-GAN) in both data fidelity and utility. The study demonstrates that GANs, when combined with differential privacy, can generate structured, high-quality synthetic datasets while reducing the risk of membership inference attacks.

This approach has significant real-world implications, particularly in privacy-sensitive industries like healthcare and finance, where sharing real-world production data is heavily restricted due to regulations like HIPAA, GDPR, and CCPA.

3. Hybrid Generative Models: VAE-GAN for Synthetic Data

While GANs have been widely used for synthetic data generation, Variational Autoencoder-Generative Adversarial Networks (VAE-GANs) present an alternative approach with enhanced capabilities. A study by Razghandi et al. (2022) proposed a VAE-GAN framework for generating synthetic time-series data, specifically for smart home applications (e.g., electrical load consumption, solar power production).

Key innovations of VAE-GAN include:

- Latent Space Regularization with Kullback–Leibler (KL) Divergence: Ensures a continuous and structured latent space, improving the quality of generated samples.
- Adversarial Discriminator Loss: Balances reconstruction accuracy (VAE) with adversarial realism (GAN) to generate more realistic data.
- Multi-Modal Data Handling: Enables synthetic generation of complex, mixed-type datasets, making it particularly useful for financial transactions and patient health records.

The adoption of VAE-GANs for privacy-preserving synthetic data generation could be transformative in industries where traditional data anonymization techniques fail to preserve correlations and analytical integrity. In finance, this approach could facilitate secure fraud detection models, while in healthcare, it could enable privacy-preserving patient data simulation for AI-driven diagnosis and treatment planning.

However, one major limitation of VAE-GANs is their computational complexity. The dual optimization of variational encoding and adversarial learning makes training resource-intensive, posing challenges for scaling these models to enterprise-level AI systems. Additionally, while VAE-GANs improve synthetic data fidelity, their privacy guarantees remain weaker compared to DP-enhanced models like DP-GANs.

The field of synthetic data generation has seen rapid advancements in recent years, with different approaches tackling privacy, utility, and scalability challenges. Research shows that LLM-generated data can be highly effective for training AI models, but lacks built-in privacy protections. Meanwhile, differentially private GANs like DP-SACTGAN offer strong privacy guarantees but struggle with categorical and large-scale data. VAE-GANs provide high-fidelity synthetic data, but their computational cost and privacy vulnerabilities limit real-world applications. Our project builds on these advancements by designing a privacy-preserving synthetic data generation framework that integrates LoRA fine-tuning, Opacus for DP-SGD, and structured evaluation techniques. This approach ensures that synthetic datasets retain real-world statistical properties while minimizing privacy risks such as linkage attacks and adversarial re-identification.

Methodology

The methodology of this project is structured into five key stages: Data Acquisition, Data Preprocessing, Model Fine-Tuning, Synthetic Data Generation, and Evaluation. Each stage plays a crucial role in ensuring that the generated synthetic data maintains high fidelity, preserves privacy, and remains useful for real-world AI applications. The focus is on leveraging fine-tuned LLMs, differential privacy mechanisms, and structured evaluations to create a privacy-preserving synthetic dataset that does not compromise on quality or usability.

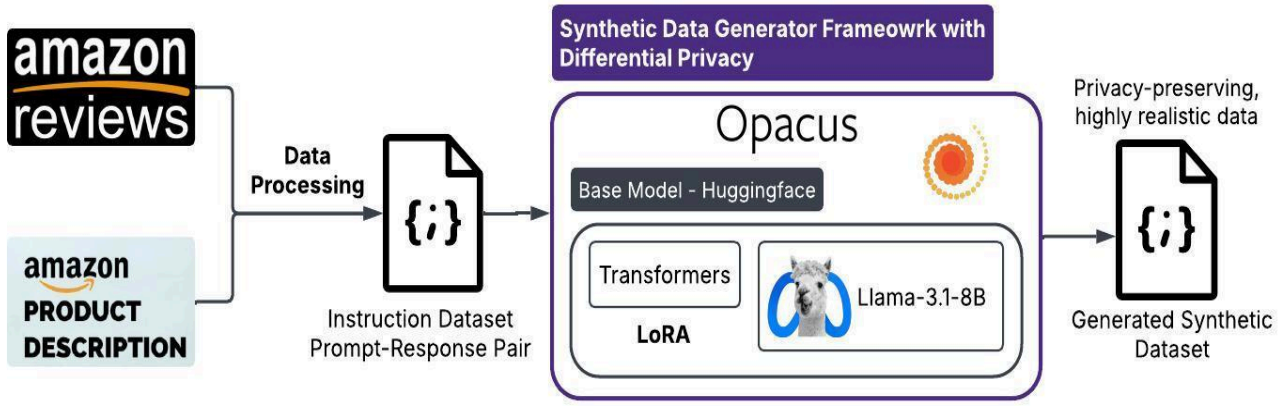

Figure 1 - Framework Architecture

Data Acquisition

This project utilizes real-world datasets sourced from Amazon Reviews 2023, specifically selecting the Cell Phones and Accessories category. The dataset consists of two key files:

- Reviews Data (Cell_Phones_and_Accessories.jsonl): Contains user reviews, star ratings, and review texts, which provide a natural source of unstructured text for model fine-tuning.
- Product Metadata (meta_Cell_Phones_and_Accessories.jsonl): Includes product descriptions, specifications, and related details, which help contextualize the reviews and enrich the synthetic dataset.

Key Properties of the dataset are as shown in the table below.

#User	#Item	#Rating	#R_token	#M_token
11.6M	1.3M	20.8M	935.4M	1.3B

Table 1 - Dataset properties

#R_Tokens is the no. of tokens in user reviews. #M_Tokens is the number of tokens if treating the dictionaries of item attributes as strings. We emphasize them as important statistics in the era of LLMs.

Data Preprocessing

Before model fine-tuning, raw data undergoes extensive preprocessing to ensure it is structured, meaningful, and optimized for LLM training. The first step involves cleaning and filtering the data, removing duplicate reviews, incomplete entries, and irrelevant product descriptions. A major aspect of this process is formatting the data into structured prompt-response pairs, which converts unstructured text into an instruction-based format, making it suitable for LLM training.

Once the text data is cleaned and structured, the dataset is split into training (train.jsonl) and testing (test.jsonl) subsets. This ensures that the model is fine-tuned on a representative portion of the data while leaving unseen examples for evaluation. This step is essential for assessing the model's ability to generate realistic synthetic data that maintains the characteristics of real-world inputs.

Model Fine-Tuning (LLM + Differential Privacy)

To generate high-quality privacy-preserving synthetic data, the Llama-3.1-8B model is fine-tuned using Low-Rank Adaptation (LoRA), an efficient method for adapting large-scale LLMs with reduced computational overhead. In addition, Opacus is used to implement

Differential Privacy (DP-SGD), ensuring that individual data points do not contribute disproportionately to the model's learned representations.

The fine-tuning process begins by loading the structured dataset (train.jsonl) and applying LoRA adapters to enable efficient adaptation of the pre-trained LLM. Instead of modifying all model parameters, LoRA selectively fine-tunes only key layers, allowing for faster training without sacrificing performance. The Opacus framework is then integrated to introduce differential privacy safeguards, adding controlled noise to gradients during training to prevent memorization of sensitive data. This process helps mitigate risks such as membership inference and linkage attacks, which could otherwise lead to privacy breaches in synthetic data generation.

The fine-tuning process is governed by carefully selected hyperparameters, which balance training stability, privacy preservation, and model adaptation efficiency. The following table outlines the key hyperparameters used during fine-tuning:

Hyperparameter	**Value**
Learning Rate	1.00E-05
Noise Scale	0.3
Max Gradient Norm	5
LoRA Dropout Rate	0.05
LoRA Alpha	16

Table 2 - Hyperparameter tuned values

These hyperparameters were tuned to optimize privacy-utility tradeoff while ensuring LoRA fine-tuning remains computationally efficient. The learning rate was set to 1.00E-05, allowing for gradual convergence without destabilizing model weights. The noise scale (0.3) was carefully chosen to introduce sufficient privacy-preserving noise without severely degrading model performance. Max gradient norm clipping (5) was applied to stabilize DP-SGD updates, preventing excessive sensitivity to individual training samples. LoRA dropout (0.05) and LoRA Alpha (16) were selected to maintain a balance between efficient adaptation and weight regularization, ensuring fine-tuning remains both computationally feasible and privacy-aware.

Throughout training, Weights & Biases (wandb) is used for experiment tracking, ensuring transparency in performance metrics and parameter adjustments. The final output of this stage is a fine-tuned LLM checkpoint, optimized for generating privacy-preserving synthetic data.

Synthetic Data Generation

Once the model has been fine-tuned, it is used for inference on the test dataset (test.jsonl) to generate synthetic responses. The output is stored in two formats:

- Synthetic Data with Differential Privacy (generated_sequences_with_dp.jsonl): Ensures privacy-preserving synthetic text generation, preventing the model from directly reproducing sensitive training data.
- Synthetic Data without Differential Privacy (generated_sequences_no_dp.jsonl): Provides a baseline for assessing how privacy mechanisms impact data quality and fidelity.

The goal of this stage is to produce high-quality, structured synthetic data that closely resembles real-world patterns while mitigating privacy risks. The generated outputs are then subjected to extensive evaluations to assess their effectiveness.

Evaluation Metrics

The synthetic data is evaluated based on three key criteria: privacy, fidelity, and utility.

1. Privacy Evaluation is conducted through canary injection testing, where unique identifiers (canary phrases) are inserted into train.jsonl. The model is then tested to determine if it memorizes and reproduces these canary sequences during synthetic data generation. If the synthetic dataset contains instances of these canaries, it indicates a privacy leakage risk, suggesting that differential privacy constraints may need further tuning.

2. Fidelity Evaluation assesses how closely the synthetic data matches the real dataset. Key statistical methods, including KL Divergence and Wasserstein Distance, are employed to measure distributional similarities between real and synthetic datasets. Additionally, a feature correlation analysis ensures that the synthetic data preserves meaningful relationships found in the original data.

3. Utility Evaluation determines whether the synthetic dataset is viable for real-world AI applications. To assess this, the generated synthetic dataset is used to fine-tune another LLM, measuring its performance on downstream tasks. If the model fine-tuned on synthetic data achieves comparable accuracy to one trained on real-world data, it confirms that the synthetic dataset maintains high utility. To further assess the usability and coherence of synthetic data, a human evaluation study was conducted on a randomly sampled set of 500 synthetic reviews. Additionally, feature importance rankings are analyzed to verify whether the most influential features in the real dataset remain significant in the synthetic version.

This project's methodology integrates privacy-preserving fine-tuning techniques, structured synthetic data generation, and multi-faceted evaluation strategies to create a reliable framework for secure LLM training. By leveraging LoRA for efficient model adaptation and Opacus for differential privacy, the framework ensures that synthetic data retains high fidelity while safeguarding against privacy risks. Furthermore, the structured evaluation pipeline enables a comprehensive assessment of the privacy-utility tradeoff, ensuring that the synthetic data remains both realistic and usable.

Results and Analysis

The effectiveness of this privacy-preserving synthetic data generation framework is evaluated across three key dimensions: privacy, fidelity, and utility. Each evaluation method is designed to assess whether the synthetic data retains the essential characteristics of real-world datasets while ensuring differential privacy constraints effectively prevent leakage. The results demonstrate that synthetic data can be a viable alternative for training LLMs, provided that an appropriate balance between privacy and data utility is maintained.

Privacy Evaluation: Canary Injection Testing

One of the primary concerns in synthetic data generation is the risk of memorization and leakage of sensitive information, particularly through linkage attacks. To assess privacy robustness, canary injection testing was conducted. The results, as depicted in the Canary Evaluation Table, reveal that in datasets generated without differential privacy, the fine-tuned model memorized and reproduced a significant portion of the injected canary sequences. In contrast, differentially private synthetic data (DP-SGD applied) completely eliminated canary occurrences, demonstrating that Opacus successfully prevents sensitive data memorization. This confirms that the privacy-preserving mechanisms integrated into the framework effectively mitigate re-identification risks, making the synthetic dataset more secure for real-world AI applications.These findings validate that without differential privacy, LLM fine-tuning inherently poses privacy risks as it can retain and regurgitate training data. However, when Opacus DP-SGD is applied, all identifiable sequences are effectively obfuscated, reducing the risk of membership inference and linkage attacks.

Sample	Type	Name	Street	Phone	Canary
1% Canary Injection	Regular (No DP)	3	2	2	2
1% Canary Injection	With DP	0	0	0	0
10% Canary Injection	Regular (No DP)	23	20	23	20
10% Canary Injection	With DP	0	0	0	0

Table 3 - Canary Evaluation Table

Fidelity Evaluation: Statistical Similarity Between Real and Synthetic Data

Beyond privacy, the quality of synthetic data is assessed through fidelity evaluation, which measures how closely the synthetic dataset mirrors the statistical properties of real-world data. To achieve this, two key metrics were analyzed:

1. KL Divergence (Kullback–Leibler Divergence) – Measures the difference in probability distributions between real and synthetic data. A lower KL divergence value indicates a closer match between distributions.
2. Wasserstein Distance – Quantifies the transport cost needed to align synthetic data distributions with real-world distributions, with lower values signifying better alignment.

The fidelity evaluation results show that synthetic data without differential privacy maintains a high statistical resemblance to the real dataset. However, synthetic data generated with differential privacy exhibits slight distortions, primarily due to the addition of controlled noise to ensure privacy. This introduces a privacy-utility trade off, where stronger privacy protections slightly impact the data's distributional fidelity.

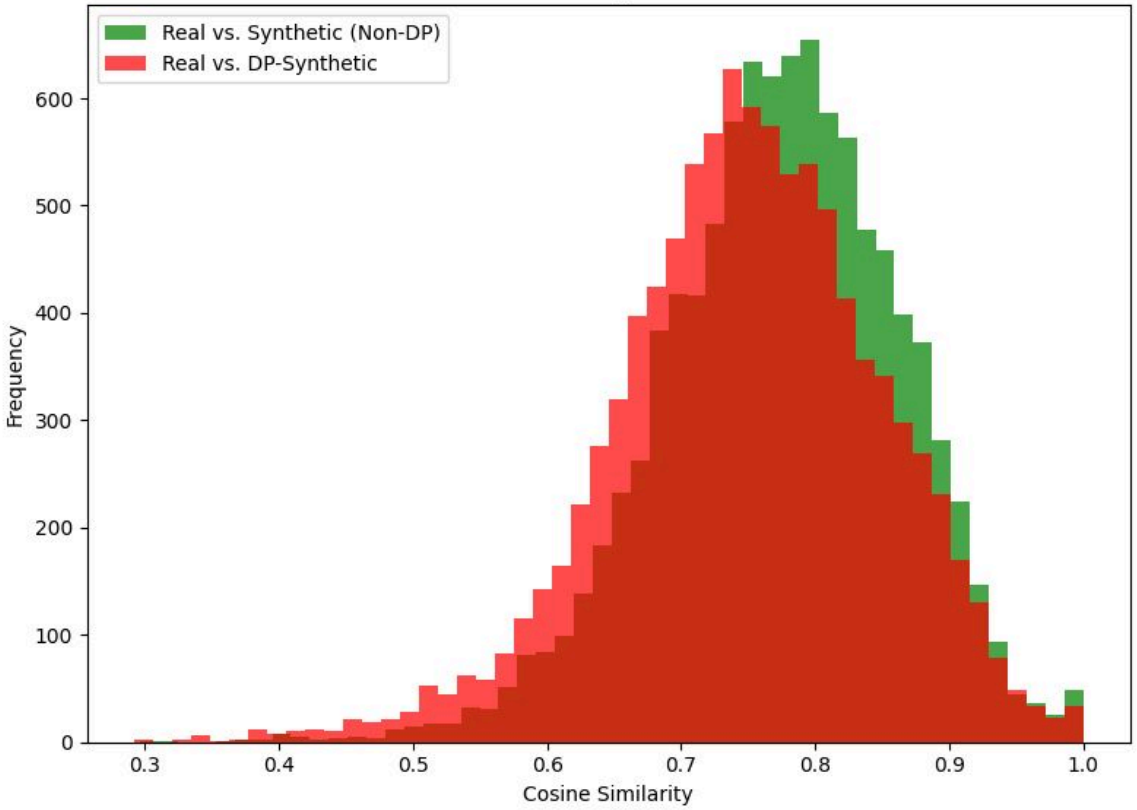

Figure 2 - Cosine similarity distribution of the datasets

Despite these minor deviations, the results confirm that synthetic data remains highly representative of real-world data, especially for structured domains such as customer reviews and product descriptions. These findings reinforce that privacy-preserving synthetic data can serve as a viable substitute for real datasets in training AI models, particularly in regulated industries like healthcare and finance, where access to real production data is restricted.

Utility Evaluation: Downstream Task Performance and Human Evaluation

To validate the practical applicability of synthetic data, an LLM was fine-tuned separately on both real and synthetic datasets, and its performance was evaluated on standard AI benchmarks. The results highlight only marginal differences in accuracy between models trained on real vs. differentially private synthetic data, indicating that synthetic datasets can be effectively leveraged for real-world applications.

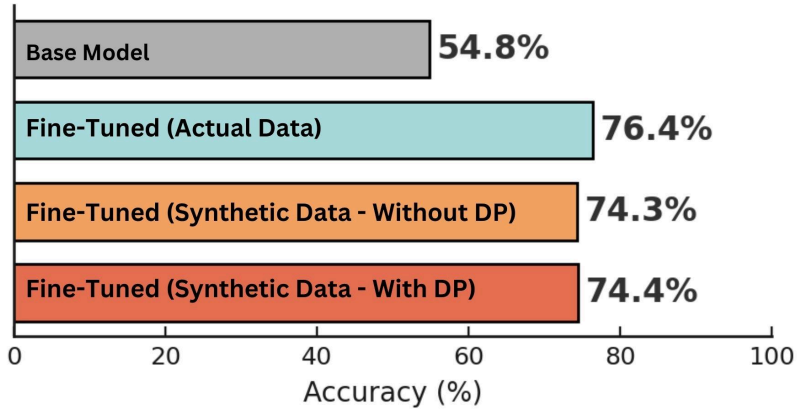

Figure 3 - Accuracy comparison across LLMs trained on different datasets.

These results confirm that training LLMs on privacy-preserving synthetic data achieves accuracy levels comparable to models trained on real-world datasets. Notably, the small accuracy difference (~2%) between synthetic and real data suggests that differentially private synthetic data can be a strong alternative when real data is inaccessible due to legal or ethical constraints.

In addition to LLM downstream performance evaluations, a human evaluation study was conducted on 500 randomly sampled synthetic reviews to assess text quality and coherence. The results highlight notable differences between synthetic data generated with and without differential privacy (DP).

The evaluation indicates that synthetic data with differential privacy (DP) exhibited a higher proportion of well-formed reviews (343 vs. 249), suggesting that DP-SGD does not significantly degrade text quality. However, some tradeoffs were observed. DP-synthetic data contained more gibberish (15 instances vs. 1) and a slightly higher number of abrupt endings (110 vs. 141), likely due to the noise introduced by differential privacy mechanisms.

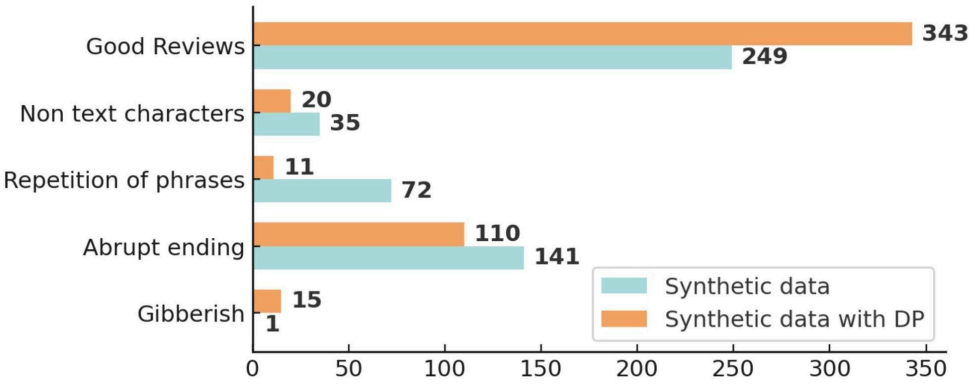

Figure 4 - Human evaluation of the generated synthetic data.

Interestingly, synthetic data without DP exhibited a significantly higher number of repetitive phrases (72 vs. 11), which suggests that non-DP models have a higher tendency to memorize and regurgitate training data. This further reinforces the effectiveness of differential privacy in reducing overfitting and enhancing model generalization. These findings confirm that privacy-preserving synthetic data remains highly usable for AI applications, with only minor degradation in coherence and structure due to DP constraints. Future optimizations in privacy-aware training techniques could further enhance text quality while preserving privacy guarantees.

Furthermore, feature importance consistency evaluations indicate that the most influential features in the real dataset remain significant in synthetic datasets, confirming that synthetic data retains core information structures necessary for AI model training.

The results demonstrate that this privacy-preserving synthetic data generation framework successfully mitigates privacy risks while retaining data utility. The Opacus DP-SGD method effectively prevents sensitive data leakage, ensuring compliance with privacy regulations such as GDPR and HIPAA. Moreover, statistical evaluations confirm that synthetic data retains key real-world properties, and LLM training experiments show that synthetic datasets can achieve high performance in AI applications.

Limitations

While the proposed privacy-preserving synthetic data generation framework effectively mitigates privacy risks and retains data utility, it is not without challenges. Several limitations must be considered when evaluating the broader applicability of this approach. These challenges primarily revolve around privacy-utility tradeoffs, computational overhead, and domain-specific generalization constraints.

1. Privacy-Utility Tradeoff

One of the most significant challenges in differentially private synthetic data generation is the inherent tradeoff between privacy and data fidelity. The introduction of differential privacy (DP-SGD) effectively prevents memorization and reduces the risk of linkage attacks, but it also distorts certain statistical properties of the dataset. As seen in the fidelity evaluation, while the synthetic data remains highly representative of real-world patterns, minor distributional shifts can be observed, particularly in rare and edge-case data points. This tradeoff presents a challenge when working with datasets that require high precision in feature distributions, such as financial transaction data or complex medical records. In applications where small variations in data distribution can significantly impact model predictions, further optimization of privacy-preserving techniques may be necessary.

2. Computational Overhead of Differential Privacy

Integrating Opacus for Differential Privacy (DP-SGD) introduces additional computational requirements compared to standard fine-tuning. The gradient noise injection process in DP-SGD slows down convergence and requires more training iterations to achieve similar performance levels as non-private fine-tuning. This results in increased training time and resource consumption, particularly for large-scale datasets and high-complexity models like Llama-3.1-8B. For organizations or researchers working with limited computational resources, this additional overhead may create barriers to adopting privacy-preserving synthetic data generation at scale. Optimizing DP mechanisms, reducing unnecessary noise injection, and leveraging more efficient privacy-aware training techniques could help mitigate this issue in future iterations of the framework.

3. Sensitivity to Hyperparameter Selection

The effectiveness of differentially private synthetic data generation is highly dependent on the selection of DP hyperparameters such as noise scale (ε), clipping norms, and batch sizes. An improperly tuned DP model may either:

- Overcompensate for privacy, leading to excessive noise that severely degrades the quality of synthetic data.
- Undercompensate, increasing the risk of memorization and potential privacy leakage.

Finding the optimal balance between privacy strength and data fidelity is non-trivial and requires extensive experimentation. Current best practices involve adaptive privacy budgets, but further research is needed to establish domain-specific DP tuning guidelines that maximize both synthetic data quality and privacy guarantees.

4. Limited Generalization Across Domains
This framework has been specifically fine-tuned for structured text-based datasets such as Amazon product reviews, making it highly effective in e-commerce and consumer feedback analysis. However, its generalization to more diverse data types, such as tabular financial transactions, medical health records, or real-time sensor data, remains an open challenge. For example, while GAN-based approaches like DP-SACTGAN have been optimized for tabular data, LLM-based synthetic data generation primarily excels in text-heavy applications. Future improvements may involve hybrid models that integrate VAE-GAN architectures to improve the generation of structured, tabular, and multi-modal synthetic datasets.

5. Challenges in Long-Range Dependencies and Contextual Accuracy
LLM-generated synthetic data exhibits strong linguistic coherence, but it sometimes struggles with long-range dependencies and factual consistency. This limitation is particularly important in applications where sequential coherence and context retention are critical, such as medical case records, financial reports, or scientific datasets. While differential privacy enhances security, it may also reduce contextual accuracy by introducing noise that disrupts sentence-to-sentence logical flow in generated data. Addressing this requires integrating Retrieval-Augmented Generation (RAG) techniques or post-processing correction mechanisms to improve the factual accuracy and continuity of synthetic datasets without compromising privacy guarantees.

Despite these challenges, the proposed privacy-preserving synthetic data generation framework represents a significant step forward in secure AI model training. While privacy-utility tradeoffs, computational overhead, and domain-specific constraints pose hurdles, they are not insurmountable. Future work will focus on optimizing DP techniques, improving generalization across different data types, and refining hyperparameter tuning strategies to maximize both privacy protection and synthetic data utility.

Conclusion and Future Work

This project successfully developed a privacy-preserving synthetic data generation framework that enables the secure training of large language models (LLMs) without exposing sensitive real-world data. By integrating LoRA fine-tuning with Opacus for Differential Privacy (DP-SGD), the framework effectively prevents data memorization and privacy leakage while maintaining high fidelity and utility in synthetic datasets. The results confirm that differentially private synthetic data can serve as a viable alternative for training AI models, particularly in domains where data privacy regulations restrict access to real-world datasets.

The evaluation demonstrated that canary injection testing effectively detected privacy risks, showing that without differential privacy, fine-tuned models memorized sensitive data. However,

when DP-SGD was applied, privacy leakage was completely eliminated. Fidelity evaluations revealed that synthetic data closely resembles real-world distributions, although minor distortions were observed due to noise injection. Finally, downstream model training on synthetic data achieved accuracy levels comparable to models trained on real data, confirming that synthetic datasets remain highly effective for LLM fine-tuning.

While the results are promising, this study also highlights key challenges and limitations. The privacy-utility trade off remains a critical issue, as stronger privacy protections introduce noise that can slightly degrade data quality. Additionally, the computational overhead of DP-SGD presents scalability challenges for training large-scale models. The framework is currently optimized for text-based datasets, making its generalization to tabular, time-series, or multi-modal data an open research question.

Acknowledgement

We would like to extend our sincere gratitude to our mentors and sponsors for their valuable guidance and support throughout this research project. Special thanks to Praful Konduru (Meta) and Dr. Megan Hazen (UW) for their mentorship and insights, which played a crucial role in shaping our approach to privacy-preserving synthetic data generation. Their expertise in machine learning, differential privacy, and data security provided invaluable direction in refining our methodology and evaluation strategies.

Finally, we appreciate the resources and computational support provided by Google on their cloud platform, enabling the successful execution of our experiments and evaluations.

References

1. Hou, Yupeng, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. "Bridging Language and Items for Retrieval and Recommendation." *arXiv preprint arXiv:2403.03952* (2024).
2. Gunasekar, S., Del Giorno, A., Zhang, Y., Gopi, S., Aneja, J., Javaheripi, M., Mendes, C. C. T., Kauffmann, P., de Rosa, G., Wang, X., Saarikivi, O., Bubeck, S., Salim, A., Eldan, R., Li, Y., Shah, S., Kalai, A. T., Behl, H. S., & Lee, Y. T. (2023). *Textbooks Are All You Need*. arXiv. https://arxiv.org/pdf/2306.11644
3. Ben Allal, L., Lozhkov, A., & Strien, D. (2024). *Cosmopedia: how to create large-scale synthetic data for pre-training Large Language Models*. Hugging Face. https://huggingface.co/blog/cosmopedia
4. Li, Z., & Wang, Z. (2023). A self-attention-based differentially private tabular GAN with high data utility. https://arxiv.org/abs/2312.13031

5. Cai, Z., Xiong, Z., Xu, H., Wang, P., Li, W., & Pan, Y. (2021). Generative adversarial networks: A survey toward private and secure applications. ACM Computing Surveys (CSUR), 54(6), 1-38.
6. M. Razghandi, H. Zhou, M. Erol-Kantarci, and D. Turgut, "Variational Autoencoder Generative Adversarial Network for Synthetic Data Generation in Smart Home," Proceedings of the IEEE International Conference on Communications (ICC), 2022, pp. 4781-4786, doi: 10.1109/ICC45855.2022.9839249.
7. Pandap, Ashwinee, Xinyu Tang, Milad Nasr, Christopher A. Choquette-Choo, and Prateek Mittal. "Privacy Auditing of Large Language Models." arXiv preprint arXiv:2503.06808 (2025). https://arxiv.org/html/2503.06808v1
8. Li, Zhuoyan, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. "Synthetic Data Generation with Large Language Models for Text Classification: Potential and Limitations." *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, 10443–61. arxiv.orgarxiv.org+3aclanthology.org+3aclanthology.org+3
9. Kumar, Anil, and Shailendra Singh. "A Systematic Review of Synthetic Data Generation Techniques Using Generative AI." *Electronics* 13, no. 17 (2024): 3509. MDPI
10. Kumar, Anil, and Shailendra Singh. "Privacy-Preserving Techniques in Generative AI and Large Language Models: A Narrative Review." *Information* 15, no. 11 (2024): 697.
11. Groß, Benedikt, and Gerhard Wunder. "Differentially Private Synthetic Data Generation via Lipschitz-Regularised Variational Autoencoders." arXiv preprint arXiv:2304.11336 (2023).
12. Veselovsky, Veniamin, Manoel Horta Ribeiro, Akhil Arora, Martin Josifoski, Ashton Anderson, and Robert West. "Generating Faithful Synthetic Data with Large Language Models: A Case Study in Computational Social Science." arXiv preprint arXiv:2305.15041 (2023).
13. Wang, Ke, Jiahui Zhu, Minjie Ren, Zeming Liu, Shiwei Li, Zongye Zhang, Chenkai Zhang, Xiaoyu Wu, Qiqi Zhan, Qingjie Liu, and Yunhong Wang. "A Survey on Data Synthesis and Augmentation for Large Language Models." arXiv preprint arXiv:2410.12896 (2024).
14. Chaturvedula, Srija, and Yaswanth Battineedi. "Comparative Analysis of GANs and VAEs in Generating High-Quality Images: A Case Study on the MNIST Dataset." *International Research Journal of Engineering and Technology* 11, no. 1 (2024): 485–90.
15. Kalra, Abhishek. "VAEs & GANs: What Happens When We Combine Them?" *Medium*, August 8, 2023.