Research Article

SafeSynthDP: Leveraging Large Language Models for Privacy-Preserving Synthetic Data Generation Using Differential Privacy

Md Mahadi Hasan Nahid[1], Sadid Bin Hasan[1]

1. University of Alberta, Canada

Machine learning (ML) models frequently rely on training data that may include sensitive or personal information, raising substantial privacy concerns. Legislative frameworks such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) have necessitated the development of strategies that preserve privacy while maintaining the utility of data. In this paper, we investigate the capability of Large Language Models (LLMs) to generate synthetic datasets integrated with Differential Privacy (DP) mechanisms, thereby enabling data-driven research and model training without direct exposure of sensitive information. Our approach incorporates DP-based noise injection methods, including Laplace and Gaussian distributions, into the data generation process. We then evaluate the utility of these DP-enhanced synthetic datasets by comparing the performance of ML models trained on them against models trained on the original data. To substantiate privacy guarantees, we assess the resilience of the generated synthetic data to membership inference attacks and related threats. The experimental results demonstrate that integrating DP within LLM-driven synthetic data generation offers a viable balance between privacy protection and data utility. This study provides a foundational methodology and insight into the privacy-preserving capabilities of LLMs, paving the way for compliant and effective ML research and applications.

 doi.org/10.32388/PJIL3E

1. Introduction

The increasing reliance on machine learning (ML) models for data-driven decision-making across sectors—ranging from healthcare and finance to social media content moderation—has amplified concerns surrounding data privacy. Models often require training on sensitive, personal information, heightening the risk of inadvertently exposing confidential details. Legislative frameworks, such as the General Data Protection Regulation (GDPR) [1][2] and the California Consumer Privacy Act (CCPA) [3], strictly regulate the handling of personal data, making it imperative for researchers and practitioners to explore strategies that protect individual privacy without compromising model performance. These stringent standards motivate our exploration of privacy-preserving synthetic data generation, where the goal is to produce artificial datasets that mirror the statistical characteristics of real data but do not reveal identifiable information.

Recent advances in Large Language Models (LLMs) have opened novel avenues for synthetic data generation. LLMs excel at capturing complex distributions, patterns, and linguistic structures from diverse corpora. Beyond traditional ML tasks, LLMs are increasingly used for processes like *fine-tuning*—adjusting pretrained models with domain-specific data—and *in-context learning* (ICL)—guiding model behavior by providing a few annotated examples as prompts at inference time [4][5][6]. However, incorporating sensitive data during either fine-tuning or ICL introduces privacy risks, as LLMs may memorize and subsequently disclose private information [7][8]. To mitigate this risk, we propose integrating Differential Privacy (DP) [9] directly into the synthetic data creation pipeline. By adding mathematically calibrated noise to the generation process, we ensure that no single individual's information unduly influences the model's output, significantly reducing the probability of re-identifying or inferring private details.

A practical example of the necessity for such techniques can be found in the healthcare domain. Medical researchers often train ML models on electronic health records (EHRs) to predict patient outcomes or recommend treatments. Suppose an ML model, trained directly on real patient data, learns to identify rare conditions. If prompted (either directly or inadvertently during inference), it could reveal sensitive patient attributes, thereby violating privacy regulations and ethical standards. Instead, consider generating a DP-enhanced synthetic dataset that statistically resembles the EHRs but omits any identifying details. This synthetic dataset can still support tasks like text classification

(e.g., categorizing clinical notes or predicting risk levels) and improve models through ICL without endangering patient privacy [10].

Synthetic datasets play a crucial role in this context, proving invaluable for the development of domain-specific language models, such as those designed for educational purposes like a Language Model for Kids [11][12]. A vast majority of data are stored in textual formats, either structured or semi-structured, like tables [4][13]. Large Language Models also exhibit significant capabilities in textual comprehension, allowing them to process and reason over various text-based data, including structured tabular data in various downstream tasks such as processing, prediction, and question-answering [14][15].

Differential Privacy (DP) provides a principled framework for protecting individual-level information in datasets. By adding carefully calibrated noise to computations, DP ensures that the inclusion or exclusion of a single data point does not significantly affect the aggregate output of a process, thereby limiting the risk of inference attacks [16]. While DP has been extensively studied in the context of traditional data analysis and model training, its direct integration into the synthetic data generation process, guided by LLMs, remains a relatively unexplored avenue. This integration enables the production of synthetic datasets that resemble the statistical properties of real data while obscuring individual-level details, ensuring compliance with privacy regulations and ethical mandates.

LLMs, which have demonstrated remarkable capabilities in capturing complex distributions and linguistic structures, present a unique opportunity for generating DP-enhanced synthetic data. By employing LLMs, it is possible to produce realistic, domain-relevant data surrogates without repeatedly exposing the original sensitive datasets. When combined with DP mechanisms, this approach aims to safeguard privacy from the inception of data generation. The resulting synthetic datasets can be used for downstream ML tasks, such as classification, without sacrificing performance or privacy integrity. This research is structured around three key questions that guide our exploration of integrating DP within LLM-driven synthetic data generation:

1. **Preservation of Model Utility:** To what extent can DP-enhanced synthetic datasets produced by LLMs preserve the predictive accuracy and robustness of ML models relative to models trained on real data?
2. **Balancing Privacy and Utility:** How do different DP noise parameters and distributions affect the trade-off between privacy guarantees and data utility, and under which conditions can

acceptable performance be maintained?

3. **Generality Across ML Architectures:** Can DP-enhanced synthetic datasets support a wide range of ML, Deep Learning (DL), and advanced LLM-based methods, thereby generalizing the applicability of this privacy-preserving data generation paradigm?

This study makes several noteworthy contributions to the field of privacy-preserving ML:

- **Integration of DP into LLM-Based Synthetic Data Generation:** We present a novel framework, SafeSynthDP that merges DP principles with LLM-driven data generation, enabling the creation of synthetic datasets that emulate sensitive information distributions without revealing individual-level details.
- **Quantitative Evaluation of Privacy-Utility Trade-offs:** We provide a rigorous empirical assessment of various DP parameters, including privacy budgets and noise distributions, to establish guidelines for achieving an optimal balance between privacy and data utility.
- **Broad Architectural Validation:** Our methodology is validated across multiple ML architectures—from traditional algorithms to sophisticated DL models (e.g., GRU, LSTM) and state-of-the-art LLMs (e.g., gpt-4o-mini[1], gemini-1.5-flash[2]) demonstrating the generality and practical relevance of our approach.

In the following sections, we first review the related works on prompting, in-context learning, DP, and synthetic data generation[3]. We then detail our methodology, experimental setup, results, and analysis, before concluding with a discussion of the implications and potential future directions of this research.

2. Related Work

Developing effective methods for privacy-preserving data release has long been a core challenge in machine learning research, particularly as organizations increasingly leverage real-world datasets containing sensitive information. Over time, a variety of approaches have emerged to generate synthetic datasets with Differential Privacy (DP), striving to protect individual records while retaining the statistical properties necessary for downstream analysis and modeling. This section surveys key contributions in DP-based synthetic data generation and highlights how existing methods motivate our approach, which uniquely integrates Large Language Models (LLMs) to produce utility-preserving, privacy-compliant datasets.

Balancing Privacy and Utility in Synthetic Data Generation

Early investigations into synthetic data as a privacy solution grappled with the fundamental tension between utility and protection. For instance, Stadler et al.[17] demonstrated that while synthetic data often outperforms naive anonymization methods, achieving a stable trade-off between preserving meaningful information and safeguarding individual privacy remains difficult. Building upon this premise, researchers have sought more principled solutions that employ DP mechanisms directly within the data generation process. Zhang et al.[18], in their work on PrivSyn, embedded DP into the synthesis of high-dimensional tabular data, ensuring that nuanced attribute correlations are maintained without disclosing sensitive records. Similarly, Zhang et al.[19] proposed PrivBayes, leveraging Bayesian networks to preserve essential statistical relationships and achieve strong DP guarantees, thus offering an early blueprint for balancing complexity, dimension, and privacy rigor.

Refining Techniques for Enhanced Utility and Applicability

As research advanced, emphasis shifted toward refining synthetic data methods that bolster utility and scalability. Arnold and Neunhoeffer[20] explored ensemble strategies, such as QUAIL, to combine multiple DP-based generators and enhance dataset quality for machine learning tasks like classification. Concurrently, Long et al.[21] introduced G-PATE, integrating Private Aggregation of Teacher Ensembles with GANs [22], showing that when carefully tuned, synthetic data can yield performance close to that of real datasets. These works underscore that improved privacy-utility calibrations are possible through advanced algorithmic formulations, hinting at the potential of more flexible generation techniques that adapt to different data distributions and task requirements.

Addressing Complexity and Theoretical Boundaries

While certain methods performed well in controlled settings, researchers recognized that practical constraints and intricate data distributions still pose hurdles. For instance, Cai et al.[23] introduced PrivMRF, using Markov Random Fields to model correlations in structured data. Despite strong results for counting and classification, PrivMRF struggled with highly complex dependencies. Hardt et al. [24] proposed early, more generic DP data release algorithms that were simple yet faced scalability limits. Rosenblatt et al.[25] and Torkzadehmahani et al.[26] evaluated and developed synthetic data generation techniques like DP-CGAN, calling for enhancements in both model stability and output

fidelity. Vietri et al.[27] explored oracle-efficient algorithms that achieved accuracy but depended heavily on advanced optimization tools. Collectively, these investigations illustrated that while DP-driven synthetic data methods could yield substantial gains, they often encountered bottlenecks in complexity, stability, or computational overhead.

Evolving from Model-Centric Approaches to Flexible Frameworks

Most early efforts integrated DP through model-centric approaches, focusing on gradient perturbation (e.g., DP-SGD)[9][28], or applied noise during post-processing steps. While such methods reduced privacy risks, they frequently impaired data utility and model performance[29]. Moreover, they typically required extensive training resources and struggled to adapt to rapidly changing tasks. These challenges highlight the need for approaches that minimize computational strain while maintaining rigorous privacy standards. The literature to date suggests that versatile and adaptive synthetic data methods—ones that can incorporate domain-specific constraints, handle complex patterns, and maintain high fidelity—are essential for real-world deployments.

Towards LLM-Driven Synthetic Data Generation

Despite the progress in DP-based synthetic data methods, relatively little attention has been devoted to leveraging Large Language Models (LLMs) for dataset generation. LLMs, such as GPT-4 [30], have shown remarkable capabilities in capturing contextual and linguistic patterns, but their integration into DP workflows remains nascent. Although some studies have examined DP during LLM inference [6], and prompt engineering techniques have been proposed to safeguard sensitive inputs [7] [8][31], the literature has yet to fully explore LLMs as generative engines for DP-compliant synthetic datasets. Addressing this gap is crucial, as LLMs can produce highly contextualized, linguistically coherent data that might better reflect real-world distributions without disclosing private attributes.

Motivation for Our Approach

Our work draws direct inspiration from these bodies of research. From early efforts, we learn the importance of balancing privacy and utility when producing synthetic datasets. From more recent advances, we adopt the idea that careful integration of DP can yield robust utility even in challenging settings. By leveraging LLMs, we aim to push beyond limitations encountered by conventional generative models, such as instability or excessive computational overhead, and generate synthetic

data that preserves critical patterns needed for machine learning tasks—particularly classification—while ensuring strong privacy protections. In essence, our approach strives to unite the analytical rigor of DP frameworks with the adaptability and expressiveness of LLMs, taking a decisive step toward producing practical, privacy-preserving synthetic datasets that align with stringent regulatory and ethical standards.

3. Methodology

Our methodology for generating differentially private synthetic data involves utilizing the advanced capabilities of Large Language Models (LLMs) to create datasets that closely mimic the statistical and semantic characteristics of an original, private dataset. Unlike traditional methods which might involve extensive training or fine-tuning, our approach is training-free, which significantly reduces computational demands and avoids the direct exposure of sensitive information from the source dataset. We detail below our process for generating and evaluating these synthetic datasets, focusing on integrating Differential Privacy (DP) mechanisms to bolster privacy protection.

To ensure that the synthetic data generated by the LLM retains both the structural integrity and thematic essence of the original data, we use a demonstration-based prompting strategy. We select a few examples from the dataset to serve as in-context demonstrations within the prompt. These examples guide the LLM to replicate the style, topics, and language patterns of the original data without directly reproducing it. By providing these exemplars, we instruct the model to generate new content that captures the statistical distribution and contextual relevance of the original dataset. For this task, we employ two recent LLMs from different model families: gpt-4o-mini [32] from the GPT series and gemini-1.5-flash [33] from Google's Gemini suite. This allows us to compare the effectiveness of synthetic data generation across different model architectures.

In our approach, we integrate Differential Privacy (DP) mechanisms directly into the data generation process for enhanced privacy. We use Laplace and Gaussian noise mechanisms to inject controlled noise into the synthetic data based on the distribution of the words. Laplace noise is applied to each data point or feature, which helps maintain the utility of the data while providing privacy. The amount of noise in Laplace mechanism or Gaussian mechanism for differential privacy is calculated based on the Laplace and Gaussian distribution equation.

The level of noise is inversely proportional to the privacy budget ϵ, where a lower ϵ indicates stronger privacy at the cost of data utility. On the other hand, Gaussian noise, known for its smoother distribution, is particularly useful for preserving the integrity of continuous data features. The noise parameters are adjusted based on the sensitivity of the data and the desired privacy level. These mechanisms are critical as they ensure that even if an adversary tries to infer membership or reconstruct original data points, the added noise significantly obscures individual entries, protecting against both membership inference attacks and attribute disclosure while maintaining data utility for machine learning tasks.

For the evaluation process, we conduct comparative experiments where we train both simple ML models, like Multinomial Naive Bayes [34] and Support Vector Machines [35], and complex deep learning models, such as Gated Recurrent Units [36] and Long Short-Term Memory networks [37], on both the original and synthetic datasets. We then evaluate their performance on a held-out test set from the original data. Additionally, we assess how well the synthetic data performs in in-context learning scenarios with the LLMs used for generation, providing insights into its ability to mimic real data for advanced learning tasks.

3.1. Synthetic Data Generation Process

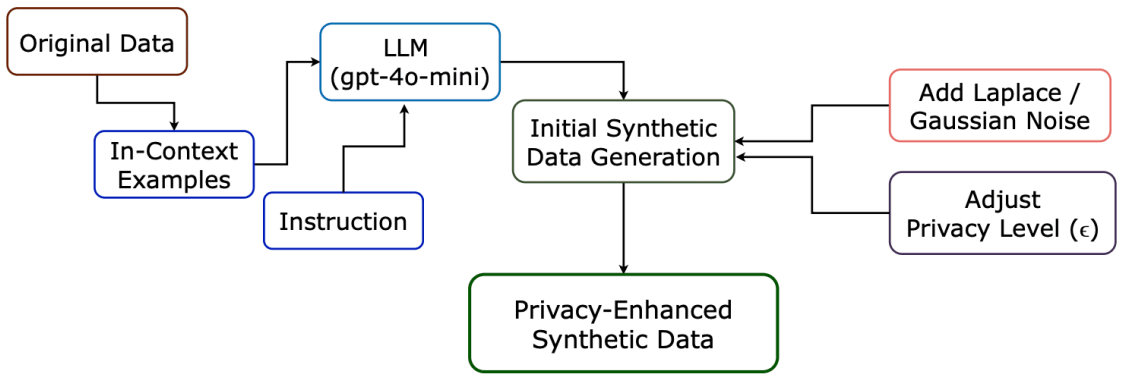

Figure 1. Workflow for Generating Privacy-Preserving Synthetic Data (SafeSynthDP). This diagram illustrates the process from selecting in-context examples from the original dataset, through the generation of initial synthetic data using the gpt-4o-mini LLM, to enhancing privacy through the addition of Laplace or Gaussian noise and adjusting the privacy level via the ϵ parameter, culminating in the evaluation of the privacy-enhanced synthetic data.

Our approach to generating synthetic data leverages the capabilities of the gpt-4o-mini model, which we use to produce datasets that emulate the structural and thematic qualities of the original data while ensuring privacy through the application of the Laplace mechanism. The process is summarized in Figure 1. The process is articulated in the following key steps:

Generating Synthetic Data with LLM

We utilize in-context learning by providing gpt-4o-mini with a curated set of demonstration prompts or examples drawn from the original dataset. For instance, when the goal is to generate synthetic news headlines, we supply the model with exemplar headlines that capture various news themes, formats, and linguistic styles. This guidance directs gpt-4o-mini to create analogous content, effectively crafting a dataset that statistically and contextually resembles the original without directly replicating any sensitive information.

Privacy Enhancement through DP Mechanisms

To further enhance privacy, we integrate Laplace and/or Gaussian mechanisms during the data generation phase with LLM. We manipulate token frequencies in the generated text by adding carefully calibrated noise based on the Laplace or Gaussian distribution. For the Laplace mechanism, we adjust the counts of common words or phrases, altering their likelihood in the synthetic text. Similarly, Gaussian noise is applied with a distribution that helps smooth the impact on the text's features. These noise additions are pivotal in masking any identifiable patterns or information that could be exploited in privacy attacks, thereby strengthening the privacy safeguards of the dataset.

Adjusting Hyperparameters for Privacy Level (ϵ)

The privacy level is governed by the hyperparameter ϵ (epsilon). A smaller ϵ provides stronger privacy guarantees but at the expense of increased noise, potentially reducing data utility. A larger ϵ allows for less noise, thus preserving more of the characteristics of original data but with weaker privacy protection. We calibrate ϵ to find the optimal balance between privacy and utility, taking into account the context of the data. For instance, with general topics, we might allow for more noise, whereas in highly sensitive contexts, we would choose a higher ϵ to maintain data fidelity while still offering privacy protection.

4. Experimental Evaluation

In this section, we present a comprehensive evaluation of our synthetic data generation methodology. We assess whether synthetic datasets produced via our Differential Privacy (DP)-augmented Large Language Model (LLM) approach can serve as effective substitutes for real data in both traditional machine learning (ML) tasks and in-context learning (ICL) scenarios. The evaluation involves training and testing various ML models, as well as conducting ICL experiments with two distinct LLM architectures. To mitigate the computational and temporal constraints associated with this preliminary investigation, we conducted experiments on a sampled subset of the AGNews dataset rather than employing the full dataset. This choice allows us to efficiently identify key trends and insights regarding the privacy-utility trade-off in synthetic data use while laying the groundwork for future, more extensive studies.

By employing LLM in this nuanced methodology, we aim to generate synthetic datasets that are both functional for further machine learning tasks and compliant with stringent privacy requirements, ensuring that any sensitive information from the original dataset remains confidential.

We discussed our hyperparameter settings in Apperndix B.

4.1. Dataset

Our project aims to evaluate the utility of Differential Privacy (DP)-augmented synthetic datasets for machine learning tasks, prioritizing privacy.In our experimental evaluation, we focused on reporting results for the AGNews [38] dataset.

We selected the AGNews dataset as our primary evaluation platform because it is well-studied, contains a variety of news topics, and supports straightforward text classification. In our experimental setting we used 12000 samples for trainings the models and 4000 samples for testing.

4.2. Clarifying the ML Task and the Role of the LLM

Our primary ML task is *text classification*, specifically categorizing news articles into four classes (World, Sports, Business, Sci/Tech) using the AGNews dataset [38]. In a typical ML setup, a model is trained directly on thousands of labeled examples until it learns to generalize. For instance, a model might see multiple examples of “Business” headlines and learn words and phrases associated with financial reports, corporate earnings, or market indicators. In contrast, an LLM can perform

classification through *prompting* and *in-context learning (ICL)*. Rather than training the model's parameters extensively, we present the LLM with a few carefully chosen examples at inference time. For example, we show the LLM a prompt:

```
Classify the headline into [World, Sports, Business, Sci/Tech].
Example: "Government imposes new tariffs on foreign goods" -> Business "Local team clinches
regional championship" -> Sports
Now classify: "Stock markets rally after positive economic indicators"
```

The LLM uses these provided examples to infer the correct category (likely "Business") without additional training. However, if these demonstration examples contain sensitive data, the model might memorize and reveal private details. Our DP-based synthetic data prevents such leakage by ensuring that no individual piece of sensitive information is directly represented.

4.3. Model Choices and Experimental Setup

To thoroughly assess the utility of the synthetic data , we evaluated a range of models with differing complexities and requirements:

- **Multinomial Naive Bayes (MNB):** A simple model that relies on frequency counts of words. If MNB performs reasonably on synthetic data, it suggests that core statistical properties are preserved [34].
- **Support Vector Machine (SVM):** SVM can handle high-dimensional feature spaces effectively. Its performance on synthetic data indicates whether essential discriminative features remain intact [35].
- **Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM):** These recurrent neural networks leverage sequential and contextual information. If these models show reasonable performance, it means the synthetic data retains more complex patterns, not just basic word frequencies [37][36].
- **LLM-based ICL (gpt-4o-mini and gemini-1.5-flash):** By testing in-context learning with synthetic data as demonstration examples, we assess whether the synthetic data can guide an LLM to make correct classifications, reflecting higher-level semantic coherence.

Each of these model types addresses a different layer of complexity in text classification. From basic statistical features (MNB, SVM) to advanced temporal patterns (GRU, LSTM) and finally to context-based reasoning (ICL with LLMs), this multi-tiered approach provides a comprehensive understanding of how well our synthetic data stands in for real data. For word embedding, we employ TfidfVectorizer to prepare the data for training Multinomial Naive Bayes (MNB) and Support Vector Machine (SVM) models. This choice is justified by TfidfVectorizer's ability to reflect the importance of words in the context of the entire corpus, which is particularly useful for traditional machine learning models that benefit from frequency-based features. For the GRU and LSTM models, we use GloVe [39] embeddings to preprocess the training data. GloVe embeddings capture semantic relationships between words through their co-occurrence statistics, providing a dense vector representation that is well-suited for deep learning models that thrive on understanding contextual and semantic nuances in text. This dual approach allows us to leverage the strengths of each embedding method tailored to the specific requirements of the models we are training.

4.4. Performance Metrics and Interpretation of Results

Our primary evaluation metric is *accuracy*, the percentage of test samples correctly classified. Accuracy is intuitive and sufficient for balanced classification tasks like AGNews. Although accuracy below 90% may seem disappointing, it is essential to interpret these numbers in the context of privacy goals. Achieving near-random performance (e.g., around 25% for a four-class problem) would be useless, while reaching moderate accuracy (e.g., above 50%) already indicates that the synthetic data has captured some meaningful signals. As our method evolves, we aim to improve accuracy while maintaining strong privacy guarantees.

5. Results

In this section, we assess the utility of our synthetic data across a range of machine learning models and Large Language Model (LLM)-based approaches.

5.1. Machine Learning Model Evaluation

We first assessed how the DP-augmented synthetic data influenced the performance of a range of ML models, including Multinomial Naive Bayes (MNB), Support Vector Machines (SVM), Gated Recurrent Units (GRU), and Long Short-Term Memory (LSTM) networks. In Table 1, we report the accuracy of

each model trained on either the original AGNews dataset or on our DP-based synthetic data. To establish a reference point, we began by training all models on the original dataset, then repeated the training with the synthetic data generated via our proposed LLM based approach.

Method	Accuracy Original Data	Accuracy Synthetic Data
MNB	80.73 %	77.92 %
SVM	86.75 %	76.43 %
GRU	83.62 %	65.62 %
LSTM	84.42 %	64.83 %

Table 1. Performance Comparison of Machine Learning Models on Original vs. Synthetic Data. This table evaluates the effectiveness of various machine learning models (including both simple and complex architectures) trained on synthetic data versus the original AGNews dataset. It highlights the utility of synthetic data for machine learning by comparing classification accuracy across different model types.

For the simpler ML models, MNB and SVM, the accuracy decline is comparatively modest (around 3% and 10%, respectively), indicating that the synthetic data preserves sufficient statistical features to enable effective classification. This suggests that our generation method can produce datasets that remain useful for tasks where interpretability and efficiency are paramount, even if the ultimate performance is somewhat reduced.

In contrast, the sequence-based GRU and LSTM architectures exhibit substantially larger accuracy reductions. These models rely heavily on subtle semantic and temporal patterns, which may be partially obscured by the DP noise and the synthetic generation process. This shortfall highlights a current limitation in our approach for tasks requiring sophisticated contextual understanding. Future refinements in noise calibration and generation strategies may be necessary to better capture the nuanced structures these models depend upon.

5.2. LLM In-Context Learning Performance

The Table 2 provides a detailed comparison of how Large Language Models (LLMs), specifically gpt-4o-mini and gemini-1.5-flash, perform across zero-shot, 2-shot, and 4-shot in-context learning scenarios, with both original and synthetic data from the AGNews dataset. In zero-shot learning, where no specific examples are provided, the models still manage to achieve accuracies around 58.67% and 54.76% respectively, demonstrating a foundational understanding derived from their pre-training. This baseline performance, however, is markedly lower than when contextual examples are provided, underscoring the effectiveness of in-context learning.

Method	**Zero-shot**	**Accuracy (2-shot)**		**Accuracy (4-shot)**	
		Original	**Synthetic**	**Original**	**Synthetic**
LLM-ICL (gpt-4o-mini)	58.67%	77.28 %	72.42 %	81.37 %	75.43 %
LLM-ICL (gemini-1.5-flash)	54.76%	65.71 %	61.21 %	72.83 %	69.28 %

Table 2. Comparison of LLM In-Context Learning Accuracy with Original vs. Our generated Synthetic Data. This table showcases the accuracy of Large Language Models (LLMs) when performing In-Context Learning (ICL) tasks using both the original AGNews dataset and our privacy-preserving synthetic data. It quantifies how well synthetic data can substitute for real data in terms of model performance for zero-shot, 2-shot, and 4-shot learning scenarios.

As we move from 2-shot to 4-shot learning, there is a notable enhancement in accuracy for both LLMs, on both types of data. This suggests that the more context we provide, the better these models can adapt their responses to mimic the desired task output. However, there is an evident performance gap between models trained on original versus synthetic data, with synthetic data consistently yielding lower accuracies. This gap becomes less significant as we increase the number of shots, indicating that additional context can compensate somewhat for the synthetic data's limitations in capturing the real data's nuances.

Comparing the two models, gpt-4o-mini generally outperforms gemini-1.5-flash, which could be due to differences in their pre-training, model architecture, or fine-tuning. Despite this, both models

demonstrate that synthetic data can be effectively utilized in in-context learning, particularly when privacy concerns limit the use of real data.

The trade-off between privacy and accuracy is clear; synthetic data introduces a performance penalty, but this can be mitigated to an extent by providing more contextual examples. This analysis confirms that while synthetic data might not replicate the performance of original data, it offers significant utility in privacy-sensitive applications. Future research could look into refining synthetic data generation or exploring further optimizations in in-context learning to narrow the performance gap further.

5.3. Evaluating Utility and Privacy Trade-Off

To further investigate the relationship between privacy and utility, we experimented with different privacy budgets (ϵ) controlling the intensity of DP noise. As shown in Table 3, lowering ϵ (stronger privacy) results in more pronounced accuracy degradation, while increasing ϵ (weaker privacy) improves performance. This outcome aligns with the fundamental DP trade-off: protecting individuals more thoroughly reduces the clarity of patterns in the data, thereby hampering ML performance.

Method	Accuracy
LLM ICL ($\epsilon = 0$)	69.83 %
LLM ICL ($\epsilon = 0.5$)	72.43 %
LLM ICL ($\epsilon = 1$)	73.64 %
LLM ICL ($\epsilon = 10$)	75.42 %

Table 3. Impact of Differential Privacy Levels on LLM ICL Accuracy Using Synthetic Data. This table explores how varying levels of Differential Privacy (DP), with various ϵ, affect the accuracy of Large Language Models (LLMs) in in-context learning tasks ($\epsilon = 0, 0.5, 1,$ and 10). It demonstrates the trade-off between privacy and utility, showing how accuracy varies as privacy protections are either intensified or reduced. Notably, even with privacy constraints, LLM-based in-context learning with synthetic data achieves performance comparable to that using the original data.

Though these accuracies may not match those of models trained directly on real data, they must be considered in the context of privacy preservation. Achieving perfect accuracy is not the sole objective in scenarios where personal information must remain confidential. Instead, identifying a balance where the ML model offers practical utility while meeting strict privacy requirements remains the overarching goal.

5.4. Privacy Assurance in Synthetic Data Generation

Our method for generating synthetic data is designed to ensure privacy by fundamentally avoiding the use of original data in the training phase. Instead, we employ a Large Language Model (LLM) to generate new data based on a few in-context examples, which serve only as guidance rather than direct sources of information. This approach means the LLM does not learn from or memorize the sensitive content of the original dataset during the generation process.

To further safeguard privacy, we incorporate noise into the synthetic data through Differential Privacy mechanisms like Laplace or Gaussian noise. This noise ensures that even if the synthetic data structurally or thematically resembles the original data, it is textually distant enough to prevent any direct linkage back to individual entries in the original dataset. By doing so, we maintain the statistical and semantic properties necessary for machine learning tasks while ensuring that the generated data does not compromise the privacy of the source data. This dual strategy of not training on original data and adding noise effectively protects privacy, making our synthetic dataset a privacy-preserving alternative for data-driven applications.

5.5. Interpreting the Findings and Future Directions

These results confirm that while DP-augmented synthetic data does not fully replicate the richness of the original dataset, it can still enable non-trivial classification performance. Simpler models and tasks are more forgiving of the noise-induced distortions, whereas models requiring complex semantic or temporal understanding face greater challenges.

For LLM-driven ICL, increasing the number of examples mitigates some of the losses, indicating potential strategies to refine prompting techniques. Moreover, systematically tuning ϵ values and experimenting with alternative noise distributions (e.g., Laplace) offers pathways to optimize this trade-off. As the field progresses, integrating more sophisticated generative techniques, exploring

data augmentation, or leveraging domain-specific priors may help retain more subtle patterns while maintaining robust privacy guarantees.

Our experiments indicate that the DP augmented LLM approach can generate synthetic data that, provides meaningful support for various ML and ICL tasks. Although improvements are needed to better capture complex dependencies and raise accuracy levels, the fundamental promise of privacy-preserving synthetic data as a safer alternative to using real, sensitive datasets is evident. The key is in balancing privacy needs with the required data utility, understanding that some performance degradation is a trade-off for enhanced privacy protection. This balance will inform future synthetic data strategies in machine learning and data science.

6. Conclusion

This work demonstrates that integrating Differential Privacy (DP) directly into Large Language Model (LLM)-based data generation can produce privacy-preserving synthetic datasets that retain sufficient utility for certain ML and in-context learning tasks. By tuning parameters such as the privacy budget ϵ, stakeholders can dynamically balance data utility with strict privacy requirements. Although the current approach does not fully capture the semantic and temporal nuances of original data, our results establish a promising baseline for privacy-preserving synthetic data generation in sensitive domains. Future directions include refining DP strategies to better preserve complex data patterns, employing more diverse evaluation metrics, and exploring adaptive or data-driven noise mechanisms. Additionally, advancing prompt engineering and conducting targeted evaluations against a broader array of privacy attacks will further solidify the robustness of this method. As privacy regulations evolve, ensuring compliance and maintaining ethical standards remain paramount. By pursuing these enhancements, this research moves toward a framework where data-driven insights can be safely harnessed without compromising individual privacy.

Limitations

While our study introduces a novel method for generating privacy-preserving synthetic data, it is crucial to acknowledge several limitations that might impact how our findings are interpreted and applied.

The quality and utility of the synthetic data generated by LLMs, such as gpt-4o-mini, are contingent upon the performance of model and training. Furthermore, there is the issue of scalability and

computational resources. Our approach bypasses the need for traditional model training for each new dataset, yet generating synthetic data still demands significant computational power, particularly with larger datasets or for tasks requiring intricate data simulation. This might restrict the practical implementation of our method in settings with limited resources.

Another limitation is the trade-off between privacy and data utility. We have demonstrated how adjusting the ϵ parameter can balance these aspects, but determining the optimal privacy level for various applications remains a challenge. Increasing privacy protections often comes at the cost of reduced data utility, which could render synthetic data less effective for applications needing high accuracy or dealing with detailed data nuances.

Our evaluation was confined to the AGNews dataset, which, while suitable for news classification, might not reflect the behavior of our method across all data types or domains. This specificity to one dataset could limit the generalizability of our findings.

There is also a theoretical risk of information leakage despite employing differential privacy. If the noise added is not appropriately calibrated for the dataset or if new, more sophisticated attack methods emerge, the privacy protections might be undermined.

Appendix A. Background

This section provides an overview of key concepts necessary for understanding our approach. We begin by explaining how Large Language Models (LLMs) adapt to various tasks through prompting and in-context learning. We then define Differential Privacy (DP) as a framework for formalizing privacy guarantees, and finally discuss synthetic data generation as a method to protect sensitive information while preserving data utility for machine learning (ML) tasks.

A.1. Prompting and In-Context Learning

Large Language Models (LLMs) are advanced neural network-based models trained on vast amounts of text data. Traditional ML models typically require explicit retraining or fine-tuning when adapting to new tasks, such as text classification (i.e., assigning a category like *Sports* or *Business* to a news article). In contrast, LLMs can often perform new tasks without extensive parameter updates, using techniques known as *prompting* and *in-context learning (ICL)*. *Prompting* involves providing an LLM with carefully crafted instructions or examples at inference time, guiding it to produce outputs aligned with a given task [4]. For instance, if the task is text classification, rather than retraining the entire

model, we can supply a prompt that outlines the categories and shows a few labeled examples. The LLM then generates answers consistent with these categories. *In-context learning* (ICL) further reduces training overhead by embedding a small set of demonstrations—usually just a few labeled examples—directly into the query fed to the model [5]. The LLM uses these demonstrations as context to infer the desired output format and reasoning steps for the task at hand. In practical terms, if we want the model to classify a piece of text, we insert a few examples of text-label pairs into the prompt. The model observes these examples and tries to classify the new, unlabeled text following the same pattern, often with surprisingly high accuracy given no explicit retraining. While prompting and ICL can save computational resources and simplify workflows, they also introduce new privacy challenges. If the prompt or the demonstrations contain sensitive information—such as personal identifiers or medical details—the LLM might memorize and reveal them later. This necessitates effective privacy safeguards that prevent inadvertent data leakage.

A.2. Differential Privacy (DP)

Differential Privacy (DP) provides a mathematically rigorous framework for preserving individual privacy when analyzing or generating data [9]. In simple terms, DP ensures that the presence or absence of any single data record in a dataset only slightly affects the outcome of a computation. By controlling the degree to which any individual record influences the final output, DP protects against adversaries attempting to infer whether a specific person's data was included. The key mechanism that enables DP is the deliberate addition of controlled noise to computations. For example, when querying a dataset to compute statistical summaries or when training a model on sensitive information, small amounts of random noise can be injected into intermediate steps. This ensures that outputs do not depend too heavily on any single record, making it difficult to "reverse-engineer" the original data. A crucial parameter in DP is the *privacy budget*, often denoted by ϵ. A smaller ϵ implies stronger privacy guarantees but typically introduces more noise and thus may reduce the utility (i.e., accuracy or performance) of the resulting model or dataset. Practitioners must choose ϵ values that strike a suitable balance between privacy protection and task effectiveness.

A.3. Synthetic Data Generation

Many ML tasks, including text classification, rely on large, representative datasets. However, using real-world data may pose significant privacy, ethical, or legal challenges—particularly when the data

includes sensitive personal information. *Synthetic data generation* addresses this issue by producing artificial datasets that resemble real data in their statistical properties but do not contain actual individual-level records [7]. Traditional synthetic data generation methods include probabilistic modeling or employing complex generative models, such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) [22][40]. More recently, LLMs have emerged as powerful tools for producing synthetic text data that captures linguistic patterns and topic distributions [8][7]. By guiding an LLM with representative examples, it is possible to generate new text instances that appear similar to the original data without reproducing any specific confidential entries. When combined with Differential Privacy, synthetic data generation can offer both strong privacy guarantees and practical utility. Specifically, one can integrate DP noise mechanisms directly into the synthetic data creation process, ensuring that each generated data point is influenced by many original records, but never strongly enough to reveal an individual's sensitive information. The resulting DP-augmented synthetic datasets can then be used for ML tasks—such as text classification—without exposing the original private data. This approach is particularly valuable in fields like healthcare or finance, where regulations and ethical considerations prohibit the use of raw sensitive data for model training or evaluation.

A.4. Linking Concepts to Our Approach

In our research, we combine these three concepts—prompting and in-context learning with LLMs, Differential Privacy, and synthetic data generation—to tackle the challenge of producing privacy-preserving datasets for ML tasks. By understanding how LLMs can be guided through prompting and ICL, we reduce computational overhead. By applying DP, we ensure that no individual's data is compromised. By generating synthetic data that mimics real samples without revealing them, we maintain data utility while adhering to strict privacy standards. This integrated approach aims to produce synthetic datasets that are both useful and compliant with rigorous privacy requirements, ultimately supporting safer, more responsible data-driven innovation in sensitive domains.

Appendix B. Implementation

B.1. Hyperparameters Settings for LLMs

We configured the hyperparameters for in-context learning as per the specifications provided in Table 4.

Parameter	Synthetic Data Generation
temperature	0.7
top_p	1
sample_n	1
max_tokens	200
num_shots	4

Table 4. The hyper-parameters we set in Synthetic Data Generation

The average text length in the dataset is approximately 200 words. We configured the maximum tokens to 200. To promote diversity in token generation by the LLM, we set a high temperature of 0.7.

B.2. ML Model configuration

In our study, we utilized an SVM classifier with a range of parameters for optimization, defined as:

```
svm_parameters = { 'clf__C': [0.1, 1, 10], 'clf__gamma': [1, 0.1, 0.01], 'clf__kernel':
                                  ['rbf', 'linear'] }
```

This configuration allows us to assess performance on both non-linear and linear data distributions by adjusting the regularization parameter C, the kernel coefficient gamma, and the kernel type. For the Multinomial Naive Bayes (MNB) classifier, we employed a pipeline defined as:

```
mnb_parameters = Pipeline([ ('tfidf', TfidfVectorizer()), ('clf', MultinomialNB()) ])
```

This pipeline incorporates TF-IDF vectorization for feature extraction followed by the MNB classifier. For our dataset split, we allocated 70% for training and 30% for validation to compare the performance of both classifiers.

For training the GRU and LSTM classifiers, we utilized the settings summarized in Table 5.

Layer	Output Shape	Param #
Embedding	(None, 975, 100)	9,308,600
GRU or LSTM	(None, 32)	12,864
Dense	(None, 4)	132

Table 5. Parameters set for GRU and LSTM classifiers

For the embedding layer, we utilized GloVe embeddings to convert input text into numerical vectors, which are then fed into the GRU or LSTM layer. The embedding layer outputs sequences of length 975 with each token represented by a 100-dimensional vector, capturing semantic relationships. The subsequent GRU or LSTM layer reduces the dimensionality to 32 units, allowing the model to capture temporal dependencies. We also incorporated a dropout rate of 0.2 to prevent overfitting by randomly setting 20% of input units to 0 during training. Finally, a dense layer with 4 units corresponds to our four-class classification task, providing the final output predictions. Additionally, we implemented early stopping to halt training when the validation loss ceased to decrease, thus avoiding unnecessary epochs and potential overfitting.

B.3. Sample Prompt for Synthetic Data Generation

For the generation of synthetic data for the news classification task, we used the following prompt to guide the Large Language Model (LLM):

```
Your task is to generate synthetic dataset for news classification task. Here is some example:

Title: Wall St. Bears Claw Back Into the Black (Reuters) Description: Reuters - Short-sellers, Wall Street's dwindling\band of ultra-cynics, are seeing green again. Class Label: “Sci/Tech”

Title: Singh Leads, but Leonard Is Following Description: Avoiding the late trouble that knocked other contenders off track, Vijay Singh held a one-stroke lead over Justin Leonard heading into the final round of the P.G.A. Championship. Class Label: “Sports”

Title: Two visions of Iraq struggle to take hold Description: Fighting in Najaf threatened to undermine a conference to choose a national assembly. Class Label: “World”

Title: Dollar Falls to Fresh Low Vs Euro (Reuters) Description: Reuters - The dollar fell to a fresh four-week low\versus the euro on Monday after a widening of the U.S. trade\gap to record levels raised worries about capital inflows in\the United States and a possible slowdown in the economy. Class Label: “Bussiness”

You need to generate synthetic dataset for these classes: "World", "Bussiness", "Sports" and "Sci/Tech".

Now generate ###<NUMBER>### different synthetic data without any explanation. Output should be in json format containing "Title", "Description", "Class_Label".
```

B.4. In-Context Learning Prompt for LLM Classifier

The following prompt was used for in-context learning with the Large Language Model (LLM) classifier to predict news class:

```
You are a helpful assistant. Your task is to predict news class for a given news. The classes are: classes: "World", "Bussiness", "Sports" and "Sci/Tech".

Here are some demonstrations:

Title: Breakthrough in Renewable Energy Technology Description: Innovative new technology in renewable energy could lead to more efficient solar panels and wind turbines. Class Label: “Sci/Tech”

Title: College Basketball Tournament Kicks Off Description: The much-anticipated college basketball tournament has begun, with teams vying for the championship title. Class Label: “Sports”

Title: Cultural Heritage Sites Under Threat Description: Several cultural heritage sites around the world are facing threats due to climate change and urban development. Class Label: “World”

Title: Tech Stocks Rally After Positive Earnings Description: Tech stocks saw a significant rally today following a series of positive earnings reports from major companies. Class Label: “Bussiness”

Now predict only the class label for the follwoing news: ###<NEW SAMPLE>###
```

Footnotes

[1] https://platform.openai.com/docs/models/gpt-4o-mini

[2] https://deepmind.google/technologies/gemini/flash/

[3] For background see Appendix A.

References

1. ^ *Marjanov T, Konstantinou M, J\u00f3\u017awiak M, Spagnuelo D (2023). "Data security on the groun d: Investigating technical and legal requirements under the GDPR". Proceedings on Privacy Enhancing Technologies.*
2. ^ *Nouwens M, Liccardi I, Veale M, Karger D, Kagal L. Dark patterns after the GDPR: Scraping consent pop -ups and demonstrating their influence. In: Proceedings of the 2020 CHI conference on human factors i*

n computing systems. 2020. p. 1–13.

3. ^*Samarin N, Kothari S, Siyed Z, Bjorkman O, Yuan R, Wijesekera P, Alomar N, Fischer J, Hoofnagle C, Eg elman S (2023). "Lessons in VCR Repair: Compliance of Android App Developers with the California Con sumer Privacy Act (CCPA)". Proceedings on Privacy Enhancing Technologies. 3: 103––121.*
4. a, b, c*Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Wint er C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Su tskever I, Amodei D (2020). "Language Models are Few-Shot Learners". Preprint, arXiv:2005.14165. Av ailable from: https://arxiv.org/abs/2005.14165.*
5. a, b*Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, Le QV, Zhou D, et al. (2022). "Chain-of-thoug ht prompting elicits reasoning in large language models". Advances in neural information processing sy stems. 35: 24824–24837.*
6. a, b*Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, As kell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler D, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskev er I, Amodei D. "Language Models are Few-Shot Learners". In: Larochelle H, Ranzato M, Hadsell R, Balc an MF, Lin H, editors. Advances in Neural Information Processing Systems. Curran Associates, Inc.; 202 0. p. 1877-1901. Available from: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bf cb4967418bfb8ac142f64a-Paper.pdf.*
7. a, b, c, d*Tang X, Shin R, Inan HA, Manoel A, Mireshghallah F, Lin Z, Gopi S, Kulkarni J, Sim R (2024). "Pr ivacy-Preserving In-Context Learning with Differentially Private Few-Shot Generation". In: The Twelft h International Conference on Learning Representations.*
8. a, b, c*Hong J, Wang JT, Zhang C, LI Z, Li B, Wang Z (2024). "DP-OPT: Make Large Language Model Your Privacy-Preserving Prompt Engineer". In: The Twelfth International Conference on Learning Represent ations.*
9. a, b, c*Dwork C, Roth A, et al. (2014). "The algorithmic foundations of differential privacy". Foundations and Trends® in Theoretical Computer Science. 9 (3–4): 211–407.*
10. ^*Keshta I, Odeh A (2021). "Security and privacy of electronic health records: Concerns and challenges". Egyptian Informatics Journal. 22 (2): 177–183.*
11. ^*Nayeem MT, Rafiei D. 2024. "KidLM: Advancing Language Models for Children -- Early Insights and F uture Directions". Proceedings of the 2024 Conference on Empirical Methods in Natural Language Proc*

essing. Miami, Florida, USA: Association for Computational Linguistics. p. 4813–4836. doi:10.18653/v1/2024.emnlp-main.277.

12. ^*Liu R, Wei J, Liu F, Si C, Zhang Y, Rao J, Zheng S, Peng D, Yang D, Zhou D, et al. (2024). "Best practices and lessons learned on synthetic data for language models". arXiv e-prints. pages arXiv--2404. Available from: https://arxiv.org/abs/2404.07503.*
13. ^*Nahid MMH (2024). "Improving Table Reasoning through Table Decomposition and Normalization". Univerity of Alberta. doi:10.7939/r3-ckmh-a783.*
14. ^*Nahid MMH, Rafiei D. 2024. "NormTab: Improving symbolic reasoning in LLMs through tabular data normalization". In: Findings of the Association for Computational Linguistics: EMNLP 2024. Miami, Florida, USA: Association for Computational Linguistics. p. 3569-3585. doi:10.18653/v1/2024.findings-emnlp.203. Available from: https://aclanthology.org/2024.findings-emnlp.203.*
15. ^*Nahid MMH, Rafiei D. TabSQLify: Enhancing reasoning capabilities of LLMs through table decomposition. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Mexico City, Mexico: Association for Computational Linguistics; 2024. p. 5725-5737. doi:10.18653/v1/2024.naacl-long.320. Available from: https://aclanthology.org/2024.naacl-long.320.*
16. ^*Shokri R, Stronati M, Song C, Shmatikov V (2017). "Membership inference attacks against machine learning models". In: 2017 IEEE symposium on security and privacy (SP). IEEE. pp. 3–18.*
17. ^*Stadler T, Oprisanu B, Troncoso C (2022). "Synthetic Data \\u2013 Anonymisation Groundhog Day". In: 31st USENIX Security Symposium (USENIX Security 22). Boston, MA: USENIX Association. p. 1451--1468. Available from: https://www.usenix.org/conference/usenixsecurity22/presentation/stadler.*
18. ^*Zhang Z, Wang T, Li N, Honorio J, Backes M, He S, Chen J, Zhang Y (2021). "PrivSyn: Differentially Private Data Synthesis". In: 30th USENIX Security Symposium (USENIX Security 21). Boston, MA: USENIX Association; p. 929-946.*
19. ^*Zhang J, Cormode G, Procopiuc CM, Srivastava D, Xiao X (2017). "PrivBayes: Private Data Release via Bayesian Networks". In: Proceedings of the 2017 ACM SIGMOD International Conference on Management of Data. New York, NY: ACM. pp. 1423–1434. doi:10.1145/3035918.3035919.*
20. ^*Arnold Q, Neunhoeffer T (2020). "Differentially Private Synthetic Data: Applied Evaluations and Enhancements". arXiv preprint arXiv:2011.05537. Available from: https://arxiv.org/abs/2011.05537.*
21. ^*Long Y, et al. 2021. "G-PATE: Scalable Differentially Private Data Generator via Private Aggregation of Teacher Discriminators". In: Proceedings of the 35th Conference on Neural Information Processing Syste*

ms (NeurIPS 2021). Neural Information Processing Systems Foundation; 2021. p. 2965-2977.

22. a, b*Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (201 4). "Generative adversarial nets". Advances in neural information processing systems. **27**.*

23. ^*Cai K, Lei X, Wei J, Xiao X (2021). "Data synthesis via differentially private markov random fields". Pro c. VLDB Endow.. **14** (11): 2190–2202. doi:10.14778/3476249.3476272.*

24. ^*Hardt M, Ligett K, Mcsherry F (2012). "A Simple and Practical Algorithm for Differentially Private Data Release". In: Pereira F, Burges CJ, Bottou L, Weinberger KQ, editors. Advances in Neural Information Pro cessing Systems, vol. 25. Curran Associates, Inc.*

25. ^*Rosenblatt L, Liu X, Pouyanfar S, de Leon E, Desai A, Allen J (2020). "Differentially Private Synthetic D ata: Applied Evaluations and Enhancements". Preprint, arXiv:2011.05537. Available from: https://arxiv. org/abs/2011.05537.*

26. ^*Torkzadehmahani R, Kairouz P, Paten B. (2019). "DP-CGAN: Differentially Private Synthetic Data and Label Generation." In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Worksho ps (CVPRW). IEEE Computer Society. pp. 98–104.*

27. ^*Vietri G, Tian G, Bun M, Steinke T, Wu S (2020). "New Oracle-Efficient Algorithms for Private Syntheti c Data Release". Proceedings of the 37th International Conference on Machine Learning, volume 119 of Proceedings of Machine Learning Research, pages 9765–9774. Available from: https://proceedings.mlr. press/v119/vietri20b.html.*

28. ^*Abadi M, Chu A, Goodfellow I, McMahan HB, Mironov I, Talwar K, Zhang L (2016). "Deep learning wit h differential privacy". Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communicati ons Security, CCS '16, page 308–318, New York, NY, USA. Association for Computing Machinery. doi:10.1 145/2976749.2978318.*

29. ^*Rosenblatt L, Herman B, Holovenko A, Lee W, Loftus J, McKinnie E, Rumezhak T, Stadnik A, Howe B, St oyanovich J (2023). "Epistemic parity: Reproducibility as an evaluation metric for differential privacy". Proc. VLDB Endow.. **16** (11): 3178–3191. doi:10.14778/3611479.3611517.*

30. ^*Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, Almeida D, Altenschmidt J, Altman S, An adkat S, et al. (2023). "Gpt-4 technical report". arXiv preprint arXiv:2303.08774.*

31. ^*Chua L, Ghazi B, Huang Y, Kamath P, Kumar R, Liu D, Manurangsi P, Sinha A, Zhang C (2024). "Mind the Privacy Unit! User-Level Differential Privacy for Language Model Fine-Tuning". First Conference o n Language Modeling.*

32. ^Hurst A, Lerer A, Goucher AP, Perelman A, Ramesh A, Clark A, Ostrow AJ, Welihinda A, Hayes A, Radford A, et al. (2024). "Gpt-4o system card". arXiv preprint arXiv:2410.21276.

33. ^Gemini Team, Georgiev P, Lei VI, Burnell R, Bai L, Gulati A, Tanzer G, Vincent D, Pan Z, Wang S, et al. (2024). "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context". arXiv preprint arXiv:2403.05530.

34. a, bKibriya AM, Frank E, Pfahringer B, Holmes G (2004). "Multinomial naive bayes for text categorization revisited". In: Proceedings of the 17th Australian Joint Conference on Advances in Artificial Intelligence, AI'04, Berlin, Heidelberg: Springer-Verlag. p. 488–499. doi:10.1007/978-3-540-30549-1_43.

35. a, bHearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B (1998). "Support vector machines". IEEE Intelligent Systems and their Applications. **13** (4): 18–28. doi:10.1109/5254.708428.

36. a, bChung J, Gulcehre C, Cho K, Bengio Y (2014). "Empirical evaluation of gated recurrent neural networks on sequence modeling". arXiv preprint arXiv:1412.3555.

37. a, bHochreiter S, Schmidhuber J (1997). "Long short-term memory". Neural Comput.. **9** (8): 1735–1780. doi:10.1162/neco.1997.9.8.1735.

38. a, bZhang X, Zhao J, LeCun Y (2015). "Character-level convolutional networks for text classification". Advances in neural information processing systems. **28**.

39. ^Pennington J, Socher R, Manning C (2014). "GloVe: Global Vectors for Word Representation". In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar: Association for Computational Linguistics. pp. 1532–1543. doi:10.3115/v1/D14-1162.

40. ^Kingma DP, Welling M, et al. (2019). "An introduction to variational autoencoders". Foundations and Trends® in Machine Learning. **12** (4): 307--392.

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.