Intelligent Systems with Applications 30 (2026) 200659

Contents lists available at ScienceDirect

Intelligent Systems with Applications

journal homepage: www.journals.elsevier.com/intelligent-systems-with-applications

Assessment of differentially private fine-tuning of large language models for synthetic clinical note generation ☆

Atiquer Rahman Sarkar [a,*], Fatima Jahan Sarmin [a], Djedjiga Mouheb [b], Benjamin C. M. Fung [c], Noman Mohammed [a]

[a] Department of Computer Science, University of Manitoba, Winnipeg, Manitoba, R3T 5V6, Canada
[b] Department of Computer Science and Software Engineering, Laval University, Quebec City, Quebec, G1V 0A6, Canada
[c] School of Information Studies, McGill University, Montreal, Quebec, H3A 1X1, Canada

ARTICLE INFO

Dataset link: https://github.com/Privacy4all/dp_llm_finetuning

Keywords:
Clinical note generation
Differential privacy
HIPAA
Privacy leakage
LLM

ABSTRACT

The sharing of clinical notes is limited by privacy regulations such as HIPAA, PIPEDA, and GDPR. Synthetic data generation (SDG) has emerged as a promising approach, but it lacks quantifiable privacy guarantee. This study investigates three differentially private (DP) fine-tuned large language models for generating synthetic clinical notes, employing a recently proposed context-aware note generation technique. We investigate the impact of differential privacy bound on two key aspects: (i) the reappearance of protected health information (PHI), and (ii) utility preservation, measured via linguistic and semantic similarity metrics in the generated notes. We experimented with various privacy budgets and generation settings. Our results, obtained on the I2B2-2014 de-identification corpus, suggest that differential privacy finetuning does not automatically provide privacy guarantees. Some synthetic notes were observed to leak protected health information. Consequently, the synthetic notes generated from the DP-finetuned LLMs are not compliant with the HIPAA privacy rule for de-identification standards. Rank-based statistical analyses demonstrate that the expected positive monotonic association between the privacy budget ϵ and leakage is not consistently significant when comparing DP-finetuned models (i.e., $\epsilon \in \{1, 4, 8\}$). More importantly, the data indicate that, while leakage increases when moving to a non-private baseline ($\epsilon = \infty$), the leakage difference shows a substantial baseline of PHI reappearance even under the strict privacy setting (e.g., at $\epsilon = 1$, 27 of 1304 notes are affected in the I2B2-2014 corpus). This suggests that a significant portion of leakage is an inherent artifact of the fine-tuning process, which is not fully mitigated by the DP mechanism, positioning ϵ-dependent leakage as only one component of the overall risk. The findings contribute to advancing the discourse on privacy-aware data sharing in healthcare research.

1. Introduction

The sharing of clinical notes is restricted by laws such as the Health Insurance Portability and Accountability Act (HIPAA) in the USA, the Personal Information Protection and Electronic Documents Act (PIPEDA) in Canada, and the General Data Protection Regulation (GDPR) in the EU (Forcier et al., 2019; Ness et al., 2008). These laws protect privacy-sensitive health information, including names, dates of birth, phone numbers, emails, zip codes, etc. HIPAA defines 18 categories of attributes that are protected health information (PHI), including an open-ended attribute that covers any information that may be used to uniquely identify an individual. However, restricting access to such information-rich clinical notes limits the potential for research to advance the healthcare sector. Clinical notes have been shown to be useful in developing machine learning models to predict life expectancy, the duration of a patient's stay under care, and so on Cai et al. (2016) and Ye et al. (2020). This information can be used for resource planning, management, and optimization. One approach to comply with the rules and still share clinical notes is to de-identify the notes by removing all protected health information tokens. Although automated tools for PHI detection are continuously being improved, they are not yet fully accurate enough to be deployed (Urbain et al., 2022; Yang et al., 2019). Additionally, some de-identifier models that worked well in certain datasets did not perform well in others (Ahmed et al.,

☆ NM was supported by the NSERC Discovery (RGPIN-04127-2022) and NSERC Alliance Grants (ALLRP 592951-24). ARS was supported by the UMGF fellowship.

* Corresponding author.
E-mail address: sarkarar@myumanitoba.ca (A.R. Sarkar).

https://doi.org/10.1016/j.iswa.2026.200659
Received 4 July 2025; Received in revised form 29 January 2026; Accepted 24 March 2026
Available online 7 April 2026

2020). Consequently, researchers are also exploring other techniques for sharing data while complying with regulatory laws.

Synthetic data generation (SDG) is increasingly recognized for its potential in both data augmentation and privacy preservation (Cai et al., 2021; Jordon et al., 2022). It has attracted considerable interest from both academic and industrial sectors as a potential remedy for data sharing limitations (El Emam et al., 2020; Hernandez et al., 2022; Liu et al., 2022). A recent report from the National Institute of Standards and Technology (NIST) has mentioned synthetic data as one of the best practices developed over the past several decades for de-identifying government datasets, advising agencies to consider using synthetic data (Garfnkel et al., 2023). However, synthetic data lacks a readily quantifiable measure of privacy. In contrast, Differential Privacy (DP) offers a quantifiable definition of privacy and ensures that if an algorithm or process is differentially private, an adversary's ability to discern whether a particular individual's data was included in the process is bounded by certain thresholds (ϵ, δ) (Abadi et al., 2016). Consequently, a potential solution to provide privacy guarantee to synthetic data may involve training or fine-tuning a generative language model on private clinical notes using differential privacy, then releasing the model or the generated synthetic clinical notes to the researchers. To verify the feasibility of this approach, the synthetic clinical notes generated by such a DP-trained/fine-tuned large language model (LLM) should be investigated to assess the level of privacy protection provided and the retained usefulness.

This work aims to investigate both the privacy aspect and the utility aspect of the synthetic clinical notes generated under differential privacy guarantees. Differential privacy limits the impact that an individual record can have on the training or fine-tuning of a model. We aim to determine whether the limit imposed by differential privacy is related to the reappearance rate of PHI in synthetic notes. If such a relationship exists, it would imply that we could control the reappearance rate of PHI by adjusting the privacy budget (ϵ) in differential privacy. Additionally, we seek to understand whether the privacy budget impacts the quality of the synthetic notes. To this end, this work seeks to answer two specific research questions:

1. What is the relationship between differential privacy bounds and the reappearance of protected health information?
2. What impact, if any, different differential privacy bounds have on the data utility preservation in the synthetic notes?

Contributions. Synthetic clinical note generation is a relatively new area that is receiving attention due to recent breakthroughs in the field of LLMs. In this study, we investigated three large language models, LLAMA-3.1-8B, Llama-2-7B, and GPT-2 Large for the generation of synthetic clinical notes under differentially private fine-tuning and measured the privacy (HIPAA compliance via sensitive PHI token reappearance) and the quality. We found that differentially private fine-tuning does not automatically provide privacy-guarantee. Some synthetic notes leaked protected health information making the synthetic notes generated from the DP-finetuned LLMs noncompliant with the HIPAA privacy act. Our experiments reveal a complex relationship between the privacy budget (ϵ) and PHI reappearance. Contrary to the simple assumption that leakage uniformly increases with ϵ, our rank-based tests found that a statistically significant positive monotonic association is not always present across all model-prompt strata when comparing only the DP-finetuned models ($\epsilon \in \{1, 4, 8\}$). A more critical finding is the presence of a substantial baseline of leakage even under the strict privacy setting ($\epsilon = 1$). This suggests that a significant component of PHI reappearance is an inherent artifact of the finetuning process itself, rather than a simple function of the ϵ value. Although we believe the qualitative insight that DP fine-tuning alone is insufficient to prevent PHI leakage likely generalizes to other settings, the specific leakage rates reported here are based on the I2B2-2014 corpus and should not be directly generalized to other note types, institutions, or languages.

The rest of the article is organized as follows: Section 2 discusses the related works. Section 3 presents the preliminaries, describing the major components of this project: differential privacy and fine-tuning LLMs using Low-Rank Adaptation (LoRA). Section 4 provides the methods and implementation details, including the conditional generation of synthetic clinical notes and the evaluation metrics used. Section 5 discusses the results of our experiments. Finally, Section 6 concludes the article with directions for future research.

2. Related works

The generation of synthetic text using large language models under differential privacy has emerged as a promising yet challenging direction for privacy-preserving data sharing. While LLMs have demonstrated remarkable capacity for producing fluent and contextually relevant text, they are also prone to memorizing and reproducing sensitive information from their training data. As a result, recent studies have explored differential privacy as a formal safeguard against such memorization. However, the literature on DP text generation varies significantly across domains, with clinical applications facing unique ethical and regulatory constraints under frameworks such as HIPAA. The following literature review is organized into two themes: **(1)** studies focusing on differentially private text generation in clinical domains, and **(2)** studies addressing DP-based text generation in non-clinical contexts.

DP Clinical Note Generation. Synthetic clinical note generation using large language models (LLMs) is a relatively new research direction. Early approaches in this space involved sampling from trained or fine-tuned LLMs with minimal or no contextual prompting, such as using a disease name or starting token (Al Aziz et al., 2021; Li et al., 2021). More recent works have explored fine-tuning and zero-shot strategies with richer contextual prompts (Baumel et al., 2024; Boulanger et al., 2024; Chuang et al., 2025; Sarkar et al., 2024).

Only a few studies have examined differentially private (DP) text generation in clinical domains, and even fewer have focused on clinical notes. Al Aziz et al. (2021) generated clinical notes by fine-tuning GPT-2 with differential privacy, experimenting with a single privacy budget ($\epsilon = 23$) on the I2B2 dataset. However, their study did not evaluate the reappearance of protected health information (PHI) tokens. Baumel et al. (2024) created an instruct dataset using a clinical entity annotator and an LLM, and then fine-tuned the LLM in a DP manner. They evaluated privacy via membership inference attacks and found no clear relationship between different DP budgets ($\epsilon = \{2, 4, 6\}$ in PHI-2 LLM) and the non-DP baseline. Importantly, they also did not report PHI reappearance statistics.

The study by Lukas et al. (2023) is conceptually close to the present work but differs in dataset choice, threat model, and privacy evaluation. While this project fine-tunes on real clinical notes, Lukas et al. used Yelp reviews related to healthcare facilities—thus not directly reflecting the structure or semantics of clinical documentation. Their attack setting also differs: they assume the attacker has access to the LLM's internal probability vectors, whereas our work assumes the attacker only sees the synthetic outputs, aligning with realistic data-sharing scenarios. Furthermore, their generation length was limited to 256 tokens, and they explored only a single privacy level ($\epsilon = 8$), precluding an analysis of privacy–utility trade-offs across multiple DP budgets.

Overall, prior DP text-generation studies in clinical settings suffer from several gaps: (1) lack of PHI reappearance analysis as a privacy metric, (2) use of short synthetic notes ($\leq$256 tokens), (3) restrictive attacker threat models assuming model access, and (4) reliance on a single DP level. Hence, the relationship between DP strength, linguistic utility, and PHI leakage in synthetic clinical notes remains poorly understood.

DP Text Generation on Non-Clinical Domains. A larger body of research has examined differentially private text generation outside

clinical domains, typically using benchmark datasets such as Yelp, Reddit, or general corpora.

Bo et al. (2021) showed that text generated under DP constraints can provide effective authorship anonymization, with consistent privacy protection across budgets ($\epsilon = \{2, 4, 6\}$). Although not focused on healthcare data, their findings highlight DP's potential for textual privacy protection. Yu et al. (2021) fine-tuned GPT-2 under DP to generate synthetic texts across several non-clinical datasets. Their emphasis was on downstream utility and linguistic coherence, not privacy leakage, and none of their corpora were healthcare-related. Yue et al. (2022) similarly used a DP-fine-tuned GPT-2 on Yelp reviews, conditioning synthetic generation on business category and star rating (e.g., "Business Type: Bar | Review Stars: 5.0"). They experimented with $\epsilon = 4$ and capped synthetic review length at 128 tokens. To assess privacy, they inserted canary sequences—phrases containing private information—into training data. With $\epsilon = 4$, none of the canaries reappeared despite 100 repetitions, while the non-DP model reproduced four of five canaries at proportional rates. They also evaluated topic coverage by comparing the top-10 topics between real and synthetic reviews.

These non-clinical studies demonstrate that DP-fine-tuned language models can preserve textual utility while mitigating exact memorization of sensitive sequences. However, they typically focus on general-purpose privacy (e.g., authorship or canary protection) rather than regulatory privacy (e.g., HIPAA compliance). Consequently, their methodologies and findings cannot be directly extrapolated to the generation of synthetic clinical text, where PHI leakage carries distinct ethical and legal implications.

Summary. The privacy threat model considered in our study assumes that the attacker has access only to the generated notes and not to the underlying generator model. This reflects a realistic data-sharing scenario, where external parties receive synthetic notes but are not granted access to the generator itself. Such a setting mirrors practical use cases, for example when institutions release synthetic datasets for research collaboration or public use. While this assumption bounds the attacker's capabilities, it does not completely prevent privacy attacks. For instance, suppose a generated synthetic note inadvertently reproduces a patient name and date (e.g., "Mr. X admitted to the Cardiac ICU on March 12"). At first glance, this snippet may appear innocuous. However, consider an insurance broker who already knows that Mr. X was hospitalized on that date. By cross-referencing this auxiliary knowledge with the synthetic notes released by the hospital, the broker could infer sensitive medical information revealed in the text (e.g., that Mr. X is suffering from Alzheimer's disease). Such an inference could be misused, for example, to justify higher insurance premiums or to enable discriminatory practices.

Differentially private text generation remains under-investigated, particularly for long, structured narratives like clinical notes. Clinical studies rarely assess PHI reappearance or regulatory compliance, while non-clinical works, though methodologically insightful, lack domain-specific privacy considerations. This gap highlights the need to systematically evaluate how differential privacy budgets affect PHI reappearance and utility in post-release settings where only synthetic notes, rather than the generator model, are accessible.

3. Preliminaries

Here, we provide a brief description of (a) differential privacy and (b) the process of fine-tuning a large language model with differential privacy. We utilized the low-rank adaptation technique from the literature for fine-tuning the LLM with differential privacy. In the next section (Section 4), we present the methodology and implementation details of the fine-tuning process.

3.1. Differential privacy

Differential privacy safeguards individual privacy by minimizing the risk of identifying specific records in the underlying dataset. It ensures that the output from an algorithm remain indistinguishable for two underlying datasets which differ by one record, achieved through the addition of noise. This approach to privacy offers a definition of privacy preservation and standardized evaluation methods. An algorithm A is ϵ-differentially private if, for two datasets D, D' differing by 1 record, all of their corresponding output (S) are bounded by ϵ: (Abadi et al., 2016)

$$\Pr[A(D) \in S] \leq e^{\epsilon} \times \Pr[A(D') \in S]. \tag{1}$$

Two important properties that make ϵ - Differential Privacy an attractive approach are its (i) robustness to post-processing and (ii) composition property. The robustness to post-processing property ensures that if an algorithm A is ϵ -differentially private, then any other algorithm that operates on the output of A is guaranteed to be at least ϵ -differentially private. The composition property states that if we make 't' queries to an ϵ -differential privacy mechanism, each query being randomized independently, the overall result will be ϵt-differentially private. This property provides a way to compute the differential privacy bound when a deep learning model is trained over multiple iterations. On the other hand, once a model is trained or fine-tuned with ϵ -differential privacy, it will remain ϵ -differentially private no matter how many responses to queries are taken from the model.

In 2016, Abadi et al. (2016) introduced a stochastic gradient descent (SGD) algorithm with differential privacy, offering a more precise accounting of privacy budget and showing that privacy in deep neural networks can be achieved with minimal impact on model performance. Their algorithm aims to achieve (ϵ, δ) differential privacy which includes an additional additive term δ to accommodate a violation of ϵ -differential privacy with a probability of δ. This violation probability should be very small, ideally less than $1/|dataset|$. They proved that, to provide a differential privacy guarantee, a deep learning model needs to control the impact of each training sample on the gradient. Their algorithm controls the privacy loss during gradient update by clipping the gradient and adding noise to the gradient updates. The clipping ensures that the gradient gets scaled down if it gets larger than a clipping threshold.

3.2. Fine-tuning LLMs with differential privacy

In this article, we experimented with three LLMs- GPT-2 (Solaiman et al., 2019), LLAMA-2-7B, and LLAMA-3.1-8B. They are autoregressive text generation models developed by OpenAI and Meta AI, to make the next token prediction given the earlier sequence. The model tries to learn a conditional probability distribution of tokens (Abadi et al., 2016) in the form of:

$$\Pr(w_1, \ldots, w_n; \theta) = \prod_{i=1}^{n} \Pr(w_i \mid w_1, \ldots, w_{i-1}; \theta). \tag{2}$$

Through training and fine-tuning, the aim is to optimize the language model θ's performance.

Low-Rank Adaptation (LoRA) is an effective technique for fine-tuning large LLMs to adapt the models for domain-specific tasks in a memory- and compute-efficient manner. LoRA of a pretrained large language model introduces new trainable parameters in the form of two low-rank matrices for a dense weight matrix, resulting in a significantly reduced number of trainable parameters compared to the original count (Yu et al., 2021). The pretrained weights (θ) are frozen, and only the newly introduced low-rank matrices are fine-tuned. By limiting the number of parameters that are fine-tuned, LoRA helps maintain the generalization abilities of the pretrained model while still allowing it to adapt to new tasks. This targeted adaptation reduces the risk of overfitting to the fine-tuning dataset. This technique is particularly

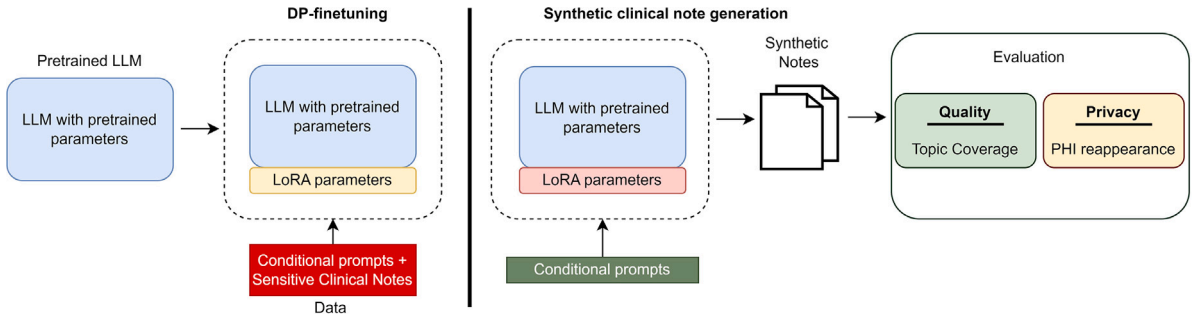

Fig. 1. Synthetic clinical notes generation and evaluation using DP-finetuned LLM. (On the left, the LLM is fine-tuned on sensitive data. On the right, the fine-tuned LLM generates synthetic notes based on prompts. The fine-tuning process utilizes the low-rank adaptation (LoRA) technique, keeping the pretrained parameters frozen while updating only the LoRA parameters.)

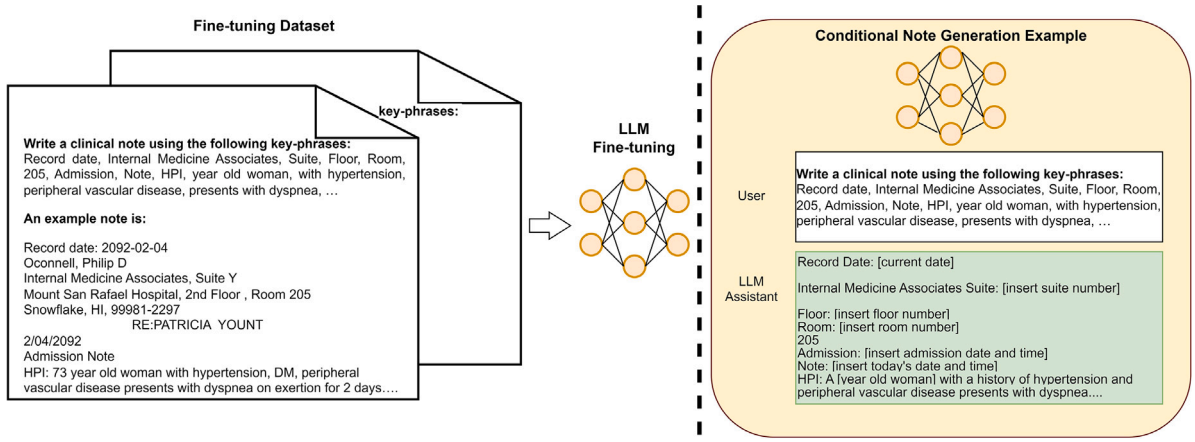

Fig. 2. To generate notes, we first fine-tune the LLM using differential privacy (left) and then send the prompts to the fine-tuned model (right).

attractive when the dataset is small or when computational resources are limited. For example, for a dense weight matrix of size 768×2304 in a certain layer, the two LoRA matrices could be of size 768×4 and 4×2304, resulting in only 12,288 trainable LoRA parameters, much less than the frozen 768×2304 matrix. Across 12 layers of decoder block, the total number of tunable parameters is only 1,47,456.

4. Methods and implementation

In this section, we present the methodology of our experimentation with implementation details. The experimentation can be divided into four major components:

1. Dataset preparation for fine-tuning.
2. Fine-tuning of the model using Low-Rank Adaptation and differential privacy.
3. Conditional generation of synthetic notes.
4. Evaluation of synthetic clinical notes.

The overall process is summarized in Fig. 1.

Dataset. We used the I2B2-2014 de-identification data set (Stubbs et al., 2015) for our experiment. It contains 1304 notes with annotated PHI tokens. It includes 23 annotation tags encompassing a wide range of personally identifiable attributes, such as names, demographic details, and contact information. The complete set of tags is: age, bioID, city, date, device, email, fax, healthplan, idnum, medicalrecord, organization, patient, phone, street, url, zip, state, country, location-other, hospital, doctor, username, and profession.

Among these, 17 categories directly correspond to identifiers specified under the U.S. HIPAA Privacy Rule's Safe Harbor provision [45 CFR 164.514(b)(2)(i)] (U.S. Department of Health and Human Services, 2025). These include: age, bioid, city, date, device, email, fax, healthplan, idnum, medicalrecord, organization, patient, phone, street, url, zip. The HIPAA Safe Harbor rule enumerates 18 identifiers whose removal is necessary for a dataset to be considered de-identified. These encompass personal names; all geographic subdivisions smaller than a state (e.g., street address, city, county, and ZIP codes with fewer than 20,000 individuals); all elements of dates (except year) related to an individual; and direct numerical, biometric, or image-based identifiers such as telephone and fax numbers, email addresses, social security and medical record numbers, device serial numbers, URLs, IP addresses, and full-face photographic images. Any other unique identifying number, characteristic, or code that could re-identify an individual is also considered PHI under this rule.

In our experiments, we restricted PHI reappearance evaluation to the 17 HIPAA-targeted categories, as these directly correspond to identifiers regulated under Safe Harbor, and therefore constitute the primary privacy risk in synthetic clinical text generation.

4.1. Dataset preparation for fine-tuning

To fine-tune the model for conditional generation, we need to provide the model with examples of the prompts and the corresponding notes. To do this, we used the recently developed conditional clinical note-generation approach (Chuang et al., 2025; Sarkar et al., 2024) which utilized two key-phrase extraction algorithms on the clinical note and subsequently used the keywords/key-phrases as conditions in the generation of the synthetic notes. Fig. 2 depicts the process. The generation prompt to the LLMs is:

"Write a clinical note using the following key-phrases: <LIST_OF_KEY_PHRASES>".

Table 1
Hyperparameters for fine-tuning the LLMs.

Hyperparameters			
ϵ in DP	1, 4, 8	δ in DP	$\frac{1}{\lvert dataset \rvert}$
adam_beta1	0.9	num_epochs	4
adam_beta2	0.999	learning_rate	1e−4
per_sample_max_grad_norm	1.0	lora_dim	4
weight_decay	0.01	lora_alpha	32

To adapt the LLM for this task, each record in the fine-tuning dataset is formatted as follows:

"Write a clinical note using the following key-phrases: <LIST_OF_KEY_PHRASES>. An example note is: <CORRESPONDING_REAL_NOTE>"

Implementation Details. Considering the context size limitation of GPT-2 (1024 tokens including prompt and response), we restricted the search-space for key-phrases to the first 700 tokens of the real notes. For the KP-miner algorithm, the least allowable seen frequency (lasf) parameter was set to 1 because, in clinical notes, a term that appeared only once could be of major significance. For YAKE, we kept the maximum n-gram size to its default value 3 and employed the default 'sequencematcher' de-duplication function with a threshold of 0.70. Subsequently, we sorted the key-phrases extracted by the two algorithms based on their location of appearance within the note. Key-phrases are sorted by their appearance location to preserve the natural clinical information flow, which may correspond to the chronological sequence of health events and is therefore important. Additionally, to ensure that the prompts do not contain any protected health information, the extracted key-phrases were compared against a list of PHI tokens, and any occurrences of PHIs in the extracted key-phrases were removed. The key-phrases were then merged in case of overlap, and this refined list of phrases serves as the context for the conditional generation prompt.

4.2. Differentially private finetuning using LoRA

For finetuning the model, we used the dp-transformers repository (Wutschitz et al., 2022) from Microsoft and used the configuration from Lukas et al. (2023) with a change in batch size (due to GPU size limitation (Titan X GTX GPU), batch size of 8 was used). We briefly mention the key parameters in Table 1. We selected $\epsilon \in \{1, 4, 8\}$ to align with privacy budgets commonly used in prior work (e.g., Lukas et al. (2023) evaluate $\epsilon = 8$, and both Baumel et al. (2024) and Bo et al. (2021) used $\epsilon = 4$), while also including a stricter setting ($\epsilon = 1$). The per-sample maximum gradient norm functions as a gradient clipping threshold, strictly bounding the L2 norm of the gradient contribution from any single training example. This limitation is essential for satisfying the sensitivity constraints required to effectively mask individual contributions with calibrated noise. Concurrently, the privacy formalism is defined under the (ϵ, δ)-differential privacy framework, where δ introduces a necessary relaxation to the strict guarantee. The delta was set to $\delta = \frac{1}{\lvert dataset \rvert}$. This term represents a negligible 'failure probability', permitting a slight deviation from pure differential privacy to achieve practical model utility while maintaining robust statistical protection for individual records.

LLAMA models were fine-tuned on an NVIDIA A100-SXM4-40GB GPU. We finetuned all models for 4 epochs. Each LLAMA finetuning run took about 90 min on average, and each note-generation run (1304 notes) took about 7 h on average. Generation halts automatically at the first EOS token (End-Of-Sequence); if none is produced, it hard-stops after 700 new tokens (i.e., $max\_new\_tokens = 700$).

4.3. Conditional generation of clinical notes

Conditional generation of clinical notes allows for providing context through a prompt, guiding the model to produce output that aligns with specific requirements. To prepare the training or fine-tuning dataset for a large language model for conditional clinical note generation, each clinical record should be preceded by corresponding contextual information. Since involving physicians to label the dataset with context is costly, an alternative approach could be to automate this process using existing key-phrase extraction techniques. Since key-phrases are widely used for document summarization, indexing, and classification (Piskorski et al., 2021), it is reasonable to assume that these automatically extracted key-phrases may provide the expected context information in a befitting manner. To enhance the robustness of the key-phrase extraction process, we combined two key-phrase extraction algorithms, KP-miner (Boudin, 2016) and YAKE (Campos et al., 2020) which have been shown to perform well in various benchmarks (Piskorski et al., 2021).

KP-miner (Boudin, 2016) is an unsupervised key-phrase extraction algorithm that improves upon the TF-IDF (Qaiser & Ali, 2018) technique, which was biased towards single words, resulting in worse performance. KP-miner overcomes this bias issue by introducing a boosting factor that influences the weight calculation of candidate key-phrases. The weight calculation formula for competing candidates, which lies at the heart of this technique, is as follows:

$$\text{Weight}_{\text{candidate}} = \text{BF} \times \text{TF}_{\text{candidate}} \times \text{IDF}_{\text{candidate}} \times \text{PF}_{\text{candidate}}. \tag{3}$$

where,

- Boosting Factor (BF):

$$\text{BF} = \min\left(\frac{\text{number of candidate terms}}{\alpha \times \text{number of candidates with length} > 1}, \sigma\right);$$

$$[\alpha = 3, \sigma = 2.3]$$

- Term Frequency (TF):

$$\text{TF}_{\text{candidate}} = \frac{\text{number of candidate's occurrences}}{\text{number of total candidates}}$$

- Inverse Document Frequency (IDF):

$$\text{IDF}_{\text{candidate}} = \log\left(\frac{\text{number of documents}}{\text{no. of candidate containing document}}\right)$$

- Position Factor (PF):

$$\text{PF}_{\text{phrase}} = \begin{cases} 1, & \text{if position rule is not defined} \end{cases}$$

Initially, the algorithm generates a list of candidates by eliminating stop-words, disregarding candidates exceeding a specified length cutoff, and excluding those that do not meet a minimum occurrence threshold (referred to as the 'least allowable seen frequency'). Subsequently, it computes the weights of these candidates, followed by a refinement process to address overlapping cases within the list. YAKE (Campos et al., 2020) is another unsupervised key-phrase extraction algorithm. In the initial step, the text is segmented into individual terms using various predefined delimiters. Following this, a set of five features is devised to characterize each term: casing of the word (lowercase/uppercase), word's position in the text (higher significance near the beginning of a document), word's frequency in the text (more frequent is more significant), relatedness to context measured by the number of unique words on the left and right side (more unique words in the surrounding area make the term less significant), and occurrence frequency in sentences (more frequent occurrences have higher significance). These features are then consolidated into a single score

Table 2
PHI Re-appearance statistics.

Prompt context	ϵ	LLAMA-3.1-8B			LLAMA-2-7b			GPT-2 Large		
		Affected record	No. of unique PHI	Avg. Occur.	Affected record	No. of unique PHI	Avg. Occur.	Affected record	No. of unique PHI	Avg. Occur.
100	1	27	1.09	1.61	13	1.06	1.26	15	1.04	1.07
	4	32	1.03	1.45	16	1.15	1.55	20	1.09	1.28
	8	30	1.05	1.37	19	1.08	1.30	28	1.03	1.08
	∞	29	1.15	1.54	25	1.07	1.48	49	1.06	1.28
200	1	25	1.07	1.58	20	1.03	1.75	11	1.00	1.12
	4	27	1.12	1.50	17	1.04	1.32	17	1.02	1.12
	8	36	1.08	1.78	16	1.04	1.41	14	1.03	1.13
	∞	39	1.09	1.48	14	1.02	1.40	38	1.03	1.25

for each term. In the subsequent step, keywords are generated using a sliding window approach, producing candidate keywords ranging from 1 to 3-grams (n = 3 by default). Each candidate keyword is assigned a final score, with lower scores indicating higher relevance and meaning. Additionally, closely related candidates are eliminated using Levenshtein distance.

4.4. Evaluation metrics

Privacy. For each prompt, we generated 3 synthetic notes so that the evaluation statistics of the notes could be more representative of the language model. For measuring privacy (and HIPAA compliance) through PHI token reappearance, we reported the following metrics:

1. Number of affected records: the total number of real notes that are directly compromised (i.e., synthetic notes contain PHIs from the corresponding real note)
2. Average number of unique PHI leakages on the directly affected records.
3. Average number of occurrences of those unique PHIs reported in Step 2.

Monotonic Association Testing. We assessed the monotonic association between the privacy budget (ϵ) and PHI reappearance using Spearman's ρ and Kendall's τ_b. Analyses were stratified by model (LLAMA-3.1-8B, LLAMA-2-7B, GPT-2 Large) and prompt granularity level. For each stratum, we evaluated the ϵ grid: $\{1, 4, 8\}$ ($n = 9$; three runs per ϵ). To address the small sample size and the presence of ties induced by the fixed ϵ levels, we avoided standard asymptotic approximations. Instead, statistical significance was assessed using exact permutation tests (20,000 iterations), testing a one-sided hypothesis ("greater") based on the expectation that weaker privacy (larger ϵ) increases PHI reappearance. Furthermore, to quantify the precision of our estimates, we computed 95% confidence intervals using a stratified bootstrap approach (5000 resamples). This method resamples the data while preserving the fixed experimental design (three observations per ϵ level), ensuring that the confidence intervals accurately reflect the variability within our specific experimental constraints. All analyses were conducted in Python, utilizing `scipy.stats` for statistic calculation and custom routines for permutation and bootstrapping procedures.

Note Quality. Throughout this study, *utility* refers exclusively to linguistic and semantic similarity between synthetic and reference notes, as measured by the metrics described below. We do not assess task-level clinical utility (e.g., performance on downstream clinical NLP tasks or clinician judgments). To evaluate the linguistic quality and content fidelity of the generated synthetic clinical notes, we utilize a combination of established natural language generation metrics: BLEU (Papineni et al., 2002), ROUGE (Chin-Yew, 2004), and cosine similarity with sentence embedding. BLEU (Bilingual Evaluation Understudy) is employed to assess the fluency and surface-level overlap between generated and reference texts. In line with prior generative studies (Oh et al., 2024; Yim & Yetisgen, 2021; Zhou et al., 2023), we report the BLEU score (BLEU-1 to BLEU-4, which measure the precision of 1- to 4-gram matches, respectively) with higher scores indicating greater local n-gram consistency and syntactic fluency. Similar to prior studies (Abacha et al., 2023; Chuang et al., 2025; Oh et al., 2024), we also incorporate the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metrics to evaluate content preservation and recall of words or phrases. Specifically, we use ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-Lsum, which quantify unigram, bigram, and longest common subsequence overlaps, as well as sentence-level summaries. These metrics are particularly useful for assessing the extent to which information from the reference texts is retained in the generated outputs.

To complement the above metrics, we check whether each synthetic note incorporates the topics (keywords/key phrases) mentioned in the corresponding prompt. We believe this statistic is an effective metric for assessing content retention. The topic coverage score is calculated as the average percentage of specified topics present in the generated outputs, and we report the overall average across all generations. However, since we employed exact matching criteria, this statistic will miss the topics that are expressed with any variation. To complement this result, we also used the cosine similarity of text embeddings as a semantic quality metric to evaluate the alignment between synthetic clinical notes and their corresponding reference notes (example of usage in prior studies: Chuang et al. (2025) and Oh et al. (2024)). Unlike surface-level n-gram-based metrics such as ROUGE or BLEU, which primarily capture lexical overlap, cosine similarity measures the angular distance between high-dimensional vector representations of texts, enabling the assessment of deeper semantic similarity at the note level, offering a more nuanced measure of factual and topical alignment beyond exact word overlap. This is particularly valuable in the clinical domain, where synonymous or paraphrased content may convey equivalent meaning despite low lexical similarity. By leveraging text embeddings generated using pre-trained language models, cosine similarity provides a measure of content relevance. Together, these metrics provide a comprehensive evaluation of both the linguistic quality and informational consistency of the synthetic clinical text.

5. Results

We experimented with 3 privacy budgets (ϵ = 1, 4, and 8) and two levels of verbosity of the prompts (number of keyphrases = 100 and 200 for KP-miner and YAKE during prompt generation). In total, we have $3 \times 2 = 6$ DP-finetuned models for each of the three LLMs. Given the limited sample size, the correlation analyses are exploratory and may have insufficient power to detect modest associations; accordingly, a lack of statistical significance should not be interpreted as evidence of no effect. We also experimented without differential privacy ($\epsilon = \infty$). Sections 5.1 and 5.2 present the PHI reappearance and note quality results, respectively. Appendix B provides illustrative examples of leakage, including the corresponding fine-tuning note excerpts and the model input context used to generate each output, with leakage instances highlighted.

Table 3
Association between privacy budget (ϵ) and leakage. We report Spearman's ρ and Kendall's τ_b rank correlation coefficients with 95% confidence intervals derived via stratified bootstrapping (resampling within each ϵ level). Permutation p-values are computed using $N = 20{,}000$ iterations; one-sided tests correspond to the *a priori* hypothesis that leakage increases with ϵ, while two-sided p-values are reported as a sensitivity analysis.

Prompt	Model	ρ [95% CI]	$p_{\text{perm,1s}}$	$p_{\text{perm,2s}}$	τ_b [95% CI]	$p_{\text{perm,1s}}$	$p_{\text{perm,2s}}$
100	LLAMA-3.1-8B	0.397 [−0.215, 0.866]	0.147	0.293	0.325 [−0.173, 0.795]	0.156	0.313
	LLAMA-2-7B	0.664 [0.487, 0.973]	0.034	0.067	0.528 [0.311, 0.933]	0.042	0.083
	GPT-2 Large	0.957 [0.957, 0.986]	<0.001	0.002	0.891 [0.891, 0.965]	<0.001	0.002
200	LLAMA-3.1-8B	0.858 [0.696, 0.986]	0.004	0.008	0.771 [0.577, 0.965]	0.004	0.008
	LLAMA-2-7B	−0.585 [−0.961, −0.055]	0.951	0.113	−0.495 [−0.905, −0.035]	0.961	0.110
	GPT-2 Large	0.458 [0.174, 0.767]	0.113	0.227	0.340 [0.115, 0.679]	0.148	0.294

5.1. PHI re-appearance

Table 2 shows the average PHI reappearance statistics across three runs for 1304 generated notes. The major revelation is that the synthetic notes generated by DP-finetuned LLMs are not HIPAA-compliant, as some of the generated notes leaked protected health information (PHI). For instance, LLAMA-3.1 under $\epsilon = 8$ leaked PHI in an average of 30 synthetic notes across three runs, with a mean of 1.37 occurrences across 1.05 unique PHI terms per affected record (prompt context level of 100). Similarly, LLAMA-2 leaked PHI in an average of 19 notes, and GPT-2 leaked PHI in 28 notes. Even at the higher privacy level ($\epsilon = 1$), PHI leakage persisted: LLAMA-3.1 leaked PHI in 27 records, LLAMA-2 in 13 records, and GPT-2 in 15 records.

This constitutes a severe breach, as the leakage occurred in the specific context of a patient whose clinical background was provided in the prompt. Although the threat model is different, this observation supports the claim made by Lukas et al. (2023) that differential privacy does not prevent the reappearance of sensitive tokens (rather, it protects against inferring the origin of a particular record).

Statistical Tests. It is generally expected that the privacy level increases with decreasing values of ϵ. To assess whether the privacy budget (ϵ) had a systematic effect on the PHI reappearance rate, we performed nonparametric tests separately for each model. Spearman's and Kendall's rank correlations were used to assess monotonic associations over $\epsilon \in \{1, 4, 8\}$. Table 3 presents the results. We report the results across two prompt-granularity levels.

Table 3 shows that the monotonic association between ϵ and PHI reappearance is statistically supported in three of the six model-prompt strata: GPT-2 Large (100-context, $\rho = 0.957$, $p_{\text{perm,1s}} < 0.001$), LLAMA-2-7B (100-context, $\rho = 0.664$, $p_{\text{perm,1s}} = 0.034$), and LLAMA-3.1-8B (200-context, $\rho = 0.858$, $p_{\text{perm,1s}} = 0.004$). In the remaining three strata, the evidence is insufficient to draw a firm conclusion about a nonzero monotonic association: LLAMA-3.1-8B at 100-context ($\rho = 0.397$, $p_{\text{perm,1s}} = 0.147$) and GPT-2 Large at 200-context ($\rho = 0.458$, $p_{\text{perm,1s}} = 0.113$) show weak positive trends that do not reach significance, while LLAMA-2-7B at 200-context exhibits a reversed association ($\rho = -0.585$, $p_{\text{perm,1s}} = 0.951$) that also fails to reach significance under the two-sided permutation analysis ($p_{\text{perm,2s}} = 0.113$) and should therefore be interpreted as exploratory rather than conclusive. The full statistics, including confidence intervals derived via stratified bootstrapping, are provided in Table 3.

Overall, these results indicate that the relationship between ϵ and PHI reappearance is significantly monotonic in some instances, but not uniformly so across all model and prompt settings. We note that the small experimental sample (n = 9 observations per stratum) limits statistical power, and variability across runs may contribute to inconclusive findings in some configurations. Consequently, these results should be interpreted as evidence of complex model-dependent behavior rather than a categorical absence of ϵ effects. This is consistent with prior observations that stronger privacy budgets do not always translate into monotonic changes in attack- or leakage-related metrics (e.g., Baumel et al., 2024).

The lack of a consistent significant monotonic relationship between ϵ and PHI reappearance points to a fundamental mismatch between the granularity of the privacy guarantee and the leakage metric. The DP implementation studied here is enforced at the sequence granularity, meaning neighboring datasets considered under DP differ by the inclusion or exclusion of one entire record. This approach, also noted in recent sequence-level DP deployments (e.g., VaultGemma Sinha & McKenna, 2025), provides a strong formal guarantee against the holistic memorization of any single record (i.e., sequence-level indistinguishability). VaultGemma applied sequence-level DP at the sequence-length of 1024-tokens. Memorization was evaluated by prompting the model with a 50-token prefix from a training document and testing whether it can reproduce the corresponding 50-token suffix. Under this evaluation, VaultGemma exhibited no detectable memorization of its training data. However, our privacy metric tracks the reappearance of individual PHI tokens, which do not constitute an entire record (sequence). To illustrate: consider a training note containing the PHI token *Alaska Air* (as in Appendix B, Example 1). The DP noise added during fine-tuning is calibrated to mask *record membership*, not the individual token's signal. Consequently, the model may still reproduce *Alaska Air* or *Alaska Airlines* when provided a prompt containing *Alaska* as a key-phrase, as observed in our ablation results. The DP guarantee is intact at the sequence level, yet the token-level leakage persists.

We posit that the leakage we observe primarily reflects an inherent artifact of the fine-tuning process itself. This inherent leakage likely stems from PHI tokens that were already present in the pre-training corpus and are simply reinforced during finetuning. Therefore, adjusting ϵ within the private range (e.g., 1 vs. 8) primarily tunes the sequence(record)-level guarantee and has a minimal impact on this inherent token-level leakage. While secondary factors like LoRA's parameter efficiency may contribute to this inconsistency, we believe that this mismatch in the privacy mechanism's goal (sequence/record) versus the privacy metric's target (PHI token) is the dominant cause.

Table 4
Topic coverage (%) and cosine similarity across models with varying DP levels and prompts.

Prompt context	ϵ	One-to-one coverage (%)			Cosine similarity		
		LLAMA-3	LLAMA-2	GPT-2	LLAMA-3	LLAMA-2	GPT-2
100	1	21.9	17.5	5.3	0.58	0.59	0.41
	4	20.1	14.0	5.9	0.61	0.59	0.43
	8	21.3	15.5	6.0	0.61	0.58	0.43
	∞	19.2	12.9	24.1	0.62	0.55	0.60
200	1	22.0	16.0	5.2	0.40	0.38	0.32
	4	15.9	15.9	5.6	0.39	0.39	0.32
	8	19.6	15.8	5.6	0.40	0.38	0.32
	∞	20.8	9.4	24.7	0.40	0.35	0.40

5.2. Quantitative quality metrics

Under differential privacy ($\epsilon \in \{1, 4, 8\}$), the LLAMA models consistently exhibit higher topic coverage and greater semantic alignment than GPT-2 at both context sizes (Table 4). At the 100-context setting, LLAMA-3.1-8B attains ∼20%–22% coverage with cosine 0.58–0.61, and LLAMA-2-7B ∼14%–18% with cosine 0.58–0.59, while GPT-2 remains low (≈5%–6% coverage; cosine 0.41–0.43). At the 200-context setting,

Table 5
BLEU and ROUGE scores across models with varying levels of differential privacy.

Metric	ϵ	LLAMA-3.1-8B				LLAMA-2-7B				GPT-2 Large			
		B-1	B-2	B-3	B-4	B-1	B-2	B-3	B-4	B-1	B-2	B-3	B-4
BLEU	1	0.216	0.044	0.016	0.007	0.311	0.063	0.021	0.009	0.237	0.020	0.003	0.001
	4	0.214	0.042	0.015	0.006	0.308	0.057	0.017	0.006	0.238	0.021	0.003	0.001
	8	0.215	0.042	0.014	0.006	0.353	0.073	0.024	0.009	0.239	0.021	0.003	0.001
	∞	0.247	0.045	0.014	0.005	0.255	0.046	0.014	0.005	0.334	0.072	0.026	0.010
		R-1	R-2	R-L	L-sum	R-1	R-2	R-L	L-sum	R-1	R-2	R-L	L-sum
ROUGE	1	0.23	0.05	0.11	0.21	0.27	0.06	0.13	0.23	0.20	0.02	0.08	0.08
	4	0.23	0.05	0.10	0.21	0.24	0.05	0.11	0.19	0.21	0.02	0.08	0.08
	8	0.23	0.05	0.10	0.21	0.26	0.06	0.12	0.21	0.21	0.02	0.09	0.09
	∞	0.26	0.05	0.11	0.23	0.22	0.04	0.11	0.18	0.29	0.06	0.14	0.14

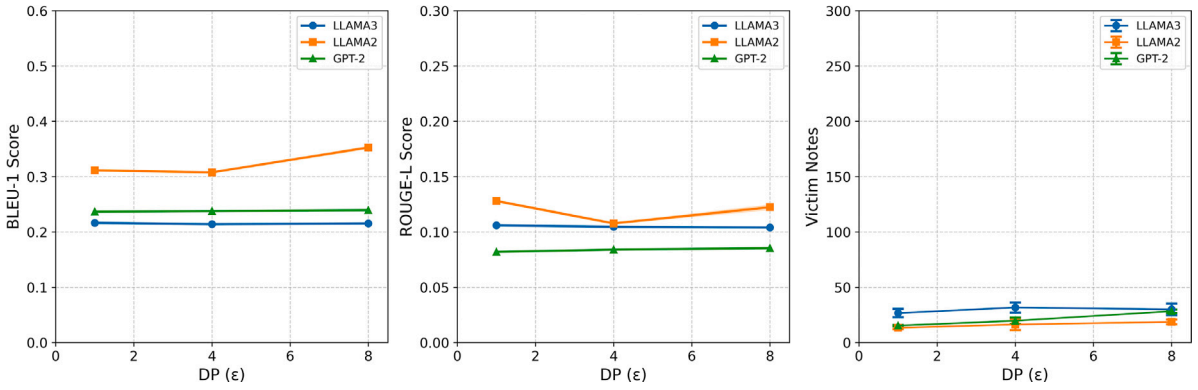

Fig. 3. BLEU-1 scores (left), ROUGE-L scores (middle), and the number of notes affected by PHI reappearance (right) for three models: LLAMA-3.1-8B, LLAMA-2-7B, and GPT-2 Large, across three differential privacy (DP) budgets ($\epsilon \in \{1, 4, 8\}$).

the pattern holds: LLAMA-3.1-8B yields ~16%–22% coverage (cosine $\approx$ 0.39–0.40) and LLAMA-2-7B ~16% (cosine $\approx$ 0.38–0.39), versus GPT-2 near 5%–6% (cosine $\approx$ 0.32). The fluctuations are modest and contained within each model–prompt setting. Increasing the context-length was sometimes associated with degraded qualitative score. For example, cosine similarity decreased from approximately 0.6 with the 100-context setting to approximately 0.4 with the 200-context setting, suggesting that the model may not reliably integrate an increasingly large set of lexical constraints into its output. This trend is consistent with prior findings that LLMs can become less reliable as the number of simultaneously specified constraints increases in multi-constraint instruction-following settings (Jiang et al., 2024; Wan et al., 2025).

When DP is removed ($\epsilon = \infty$), GPT-2 shows a marked jump in topic coverage—to 24.1% (100) and 24.7% (200)—and improves in cosine similarity. LLAMA-3.1-8B's cosine also peaks at 0.62 (100, $\epsilon = \infty$), but its coverage changes are comparatively mild. Overall, the table indicates that LLAMA models are more faithful to reference topics and semantics under DP, whereas GPT-2's fidelity is strongly contingent on the no-DP regime, underscoring a clear model-by-privacy interaction in topic preservation and semantic coherence.

Table 5 and Fig. 3 summarizes BLEU (B-1/2/3/4) and ROUGE (R-1/R-2/R-L/L_{sum}) scores across privacy budgets for synthetic notes generated in the 100-context setting. LLAMA-3.1-8B is stable under DP (B-1 $\approx$ 0.214–0.216) and shows a modest unigram gain without DP (B-1 = 0.247 at $\epsilon = \infty$), with minimal changes to B-2/3/4. LLAMA-2-7B attained better BLEU scores than LLAMA-3.1-8B, but fluctuated across DP budgets. In contrast, GPT-2 Large is stable under DP (B-1 $\approx$ 0.237–0.239) but exhibits the largest no-DP jump (0.334/0.072/0.026/0.010), indicating that GPT-2 Large benefits most from the no-DP regime.

We observed similar pattern in the ROUGE metrics (R-1, R-2, R-L, and L-sum) as well. Removing DP ($\epsilon = \infty$) yields the largest quality gains for GPT-2 Large (e.g., ROUGE-1/2/L = 0.29/0.06/0.14). The ROUGE scores remain relatively stable across privacy settings, with only minor fluctuations. Overall, we observe that surface-level and semantic similarity metrics (BLEU, ROUGE, cosine similarity) remain relatively stable across ϵ for some of the LLAMA models; however, we emphasize that these proxy measures do not constitute validation of clinical utility.

5.3. Ablation (PHI filtering disabled in generation)

Although key-phrase extraction from clinical notes might seem unlikely to capture PHIs, our ablation demonstrates that extracted key-phrases do, in fact, contain PHIs. This leads to leakage through the prompt context itself. Empirically, we observe hundreds of affected records across $\epsilon \in \{1, 4, 8\}$ with no consistent monotonic trend, indicating that leakage is not simply a function of the privacy budget in this scenario. When contrasted with earlier results of PHI-removed context, the evidence suggests that prompt-borne leakage can dominate model-borne leakage. Detailed results are provided in Appendix A. While prompts clearly serve as a vector for explicit PHI, the complex interaction between context-driven and model-intrinsic leakage mechanisms remains to be characterized. Specifically, the degree to which sanitized contexts act as retrieval cues for memorized data remains undetermined. This limitation highlights a critical avenue for future research: systematically evaluating how the nature and length of sanitized contexts influence the risk of PHI leakage. These findings also underscore the need for layered safeguards, particularly preprocessing and generation-time filters for synthetic notes, in addition to differential privacy.

6. Conclusion

Differential privacy (DP) is widely regarded as a gold standard for privacy-preserving learning, offering principled guarantees across diverse settings. In this work, using the I2B2-2014 de-identification corpus, we examined the impact of DP on PHI reappearance and the quality of synthetic clinical notes after fine-tuning a large language model. We evaluated multiple privacy budgets. Despite DP,

we observed PHI leaks, indicating that DP fine-tuning alone is insufficient and should be complemented with additional safeguards. While our qualitative conclusion, namely that DP fine-tuning alone is insufficient, is likely broadly applicable, the quantitative leakage rates should not be directly generalized to other datasets or clinical settings. Recent advances in agentic AI (e.g., AgentClinic and AgentHospital) (Karunanayake, 2025) open the door to privacy attacks when models are fine-tuned on raw data containing PHI. Such systems risk exposing sensitive details in conversation-like interactions, highlighting the need for strict privacy controls. These findings also have implications beyond healthcare; analogous risks may arise in legal, financial, and proprietary textual domains.

Limitations and future work

Our findings are based on a single de-identification corpus (I2B2-2014), limiting external validity. This corpus was constructed specifically for de-identification research and shared-task evaluation, and contains longitudinal clinical narratives with manually annotated PHI; PHI mentions were replaced with realistic surrogates, and annotation was performed via double-annotation and adjudication procedures (Stubbs et al., 2015). Because the corpus is purpose-built for de-identification, its PHI density, note composition, and reporting style may differ from routine contemporary EHR text. Moreover, the dataset's provenance and release timeframe (2014) introduce temporal limitations: clinical language, templates, coding practices, and institutional workflows have evolved since then. Consequently, the quantitative leakage rates and utility estimates reported here should not be assumed to transfer directly to other note types (e.g., radiology vs. discharge summaries), institutions, languages, or more recent clinical corpora. While permutation tests ensure the validity of our significance estimates despite the small sample size ($n = 9$), the statistical power of these tests remains constrained. Moreover, for one model configuration (e.g., GPT-2 Large, 100-context), we observe a strong, statistically significant association ($\rho = 0.957$, $p_{\text{perm,1s}} < 0.001$). In contrast, for another model configuration (e.g., LLAMA-3.1-8B, 100-context), the correlation is inconclusive ($\rho = 0.397$, $p_{\text{perm,1s}} = 0.147$). Consequently, our analysis is conservative: it is likely to detect only large, robust effects (such as those observed in GPT-2) while potentially failing to detect moderate associations. Non-significant findings in this study should be interpreted as a lack of strong evidence for a monotonic trend, rather than definitive evidence of independence. We therefore acknowledge this limitation and its implications for generalizability of our findings.

Another significant limitation of this study is that we did not empirically validate clinical utility through downstream tasks or clinician-facing evaluation. Our utility assessment relies primarily on surface-form similarity and distributional proxies, which may not perfectly correlate with task-specific performance (e.g., phenotype extraction, medical coding, or decision-support relevance). Therefore, higher or lower surface-level scores alone should not be taken as definitive evidence that the synthetic notes reliably retain (or fail to retain) clinically actionable information. We leave the comprehensive validation of these models via domain-expert review and task-specific clinical applications as an essential direction for future research.

Future work will broaden the scope of datasets, models, and conditional generation techniques. The potential of open-weight models such as gpt-oss-120B and gpt-oss-20B to improve privacy and utility under differential privacy should be investigated. Furthermore, given that releasing models trained on sensitive text entails residual risk even under DP, future studies should investigate alternative DP fine-tuning strategies alongside complementary deployment-time defenses. Key areas for investigation include: (i) Training Data Sanitization: While automated de-identification does not guarantee 100% recall, it is critical to quantify the reduction in residual leakage when the dataset undergoes rigorous de-identification prior to training or finetuning. (ii) Instruction-Based Defense: Future studies should examine the impact of incorporating explicit de-identification constraints (e.g., 'Do not output any name or {PII}') directly into the generative system prompt to leverage the model's instruction-following capabilities as a soft privacy safeguard. (iii) Output Filtering (Re-deidentification): Finally, the efficacy of post-hoc defenses, applying de-identification techniques or dictionary checks to the generated output, remains a vital area for evaluation to mitigate residual leakage persisting despite DP.

While this study (and similar efforts such as VaultGemma (Sinha & McKenna, 2025)) applies sequence-level differential privacy, future work should explore token-level DP training for large language models, which would offer stronger protection against the reappearance of individual PHI tokens. However, token-level DP currently faces substantial limitations: it is computationally prohibitive at LLM scale, as each token must be treated as a separate record for per-sample gradient computation and clipping. This results in extreme privacy budget consumption and makes training impractical with current algorithms and hardware. Overcoming these efficiency and scalability barriers is therefore an important challenge for future research. Researchers should explore whether feature-level differential privacy from the structured data domain can be incorporated into synthetic note generation.

CRediT authorship contribution statement

Atiquer Rahman Sarkar: Conceptualization, Methodology, Software, Visualization, Writing – original draft. **Fatima Jahan Sarmin:** Methodology, Writing – reviewing & editing. **Djedjiga Mouheb:** Methodology, Writing – reviewing & editing. **Benjamin C. M. Fung:** Methodology, Writing – reviewing & editing. **Noman Mohammed:** Conceptualization, Methodology, Funding acquisition, Resources, Supervision.

Declaration of Generative AI and AI-assisted technologies in the writing process

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process: During the preparation of this work the author(s) used ChatGPT 5 in order to improve sentence clarity. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

NM was supported by the NSERC Discovery Grants (RGPIN-04127-2022) and NSERC Alliance Grants (ALLRP 592951-24). ARS was supported by the University of Manitoba Graduate Fellowship (UMGF). We sincerely thank the anonymous reviewers for their constructive insights, which improved the clarity of presentation and the analysis.

Appendix A. Ablation study

This appendix contains the ablation observation where no steps were taken to ensure that PHIs were removed from the extracted context. The context prompt for conditional generation can themselves carry PHI and thereby induce leakage independent of model parameters or privacy budget.

Table A.1 reports the leakage statistics for each model and privacy budget $\epsilon \in \{1, 4, 8\}$ and for two context sizes (100, 200). It is observed that context-borne leakage dominates the privacy budget effects. The counts of affected records are in the hundreds for both LLAMA

Table A.1
PHI Re-appearance statistics.

Prompt		LLAMA-3.1-8B			LLAMA-2-7b			GPT-2 Large		
context	ϵ	Affected record	No. of unique PHI	Avg. Occur.	Affected record	No. of unique PHI	Avg. Occur.	Affected record	No. of unique PHI	Avg. Occur.
100	1	293	1.22	1.81	273	1.21	1.87	39	1.06	1.56
	4	273	1.22	1.78	304	1.27	1.82	53	1.12	1.73
	8	278	1.23	1.87	261	1.19	1.71	52	1.05	1.57
200	1	323	1.22	1.95	280	1.24	1.92	15	1.07	1.18
	4	257	1.19	1.72	261	1.21	1.84	16	1.07	1.24
	8	298	1.23	1.88	277	1.21	1.87	15	1.02	1.14

Table A.2
ROUGE and BLEU scores across models with varying levels of differential privacy.

Metric	ϵ	LLAMA-3.1-8B				LLAMA-2-7b				GPT-2 Large			
		R-1	R-2	R-L	L-sum	R-1	R-2	R-L	L-sum	R-1	R-2	R-L	L-sum
ROUGE	1	0.24	0.06	0.11	0.21	0.29	0.07	0.14	0.24	0.15	0.01	0.05	0.11
	4	0.23	0.05	0.10	0.20	0.31	0.09	0.16	0.26	0.15	0.01	0.05	0.11
	8	0.23	0.05	0.10	0.21	0.25	0.06	0.12	0.21	0.15	0.01	0.05	0.11
		B-1	B-2	B-3	B-4	B-1	B-2	B-3	B-4	B-1	B-2	B-3	B-4
BLEU	1	0.223	0.047	0.017	0.008	0.336	0.075	0.026	0.011	0.229	0.019	0.003	0.002
	4	0.216	0.043	0.015	0.006	0.258	0.069	0.027	0.012	0.232	0.020	0.003	0.001
	8	0.217	0.043	0.015	0.006	0.335	0.062	0.021	0.009	0.232	0.020	0.003	0.002

Table A.3
Topic coverage (%) and cosine similarity across models with varying DP levels and prompts.

Prompt		One-to-one coverage (%)			Cosine similarity		
context	ϵ	LLAMA-3	LLAMA-2	GPT-2	LLAMA-3	LLAMA-2	GPT-2
100	1	22.8	21.9	1.9	0.62	0.64	0.40
	4	19.0	29.3	1.9	0.63	0.68	0.42
	8	20.6	16.8	1.9	0.63	0.59	0.43
200	1	21.2	17.9	2.0	0.41	0.41	0.31
	4	15.5	18.2	2.0	0.41	0.41	0.31
	8	18.8	16.9	2.0	0.41	0.41	0.31

models regardless of ϵ. Within each model, affected-record counts and per-record occurrence rates fluctuate across $\epsilon = \{1, 4, 8\}$ without a consistent ordering, showing evidence that leakage here is primarily a function of the provided context rather than the DP budget. GPT-2 shows much smaller absolute counts, consistent with lower capacity and weaker content fidelity of the model (see Tables A.2 and A.3).

Appendix B. Leakage examples

Example 1: Snippet from the note used in fine-tuning

> ...with a past medical history notable for coronary artery disease, status post myocardial infarction with known ventricular tachycardia with AICD in place. He continues to work at **Alaska Air** and has multiple family obligations which are stressful to him...
>
> **Given context for the generation/attack:**
>
> Record date, HISTORY OF PRESENT ILLNESS, lovely 57-year-old gentleman, past medical history notable for coronary artery disease, status post myocardial infarction, known ventricular tachycardia with AICD in place, continues to work at Alaska, multiple family obligations, stressful, exertional neck, chest, arm, jaw pain, felt his AICD fire, lightheadedness or weakness, atenolol, simvastatin and Viagra, PHYSICAL EXAMINATION, exam today, ...
>
> **Snippet from the generated note:**
> The patient presents as follows on February 27.
> HISTORY OF PRESENT ILLNESS
>
> Lovely 57-year-old gentleman came into clinic today complaining of exertional neck, chest, arm, jaw pain. He has been feeling stressed lately due to multiple family obligations. The day before he experienced palpitations which culminated when his automatic internal cardioverter–defibrillator (AICD) fired twice while walking home from work at **Alaska Airlines**, resulting in lightheadedness and weakness lasting several minutes after each discharge. His medication regimen includes atenolol, simvastatin and sildenafil citrate (Viagra).
> PHYSICAL EXAMINATION
> Exam was performed today with results as follows: - Weight: ...

Example 2: Snippet from the note used in fine-tuning

> Record date: 2078-10-28 CARDIOLOGY PREAKNESS HEALTHCARE CENTER Interval History: Roger Easterling, M.D. 27 Caroll St, #734 Bluefield, ME 12111
>
> Dear Dr. Easterling: I had the pleasure of seeing Mr. **Brian Martin** in the Cardiology Service on 10/28/2078. He has a history of chest pain and...
>
> Assessment and plan: Mr. **Martin** has episodic palpitations associated with some chest discomfort...
>
> **Given context for the generation/attack:**
>
> Record date, CARDIOLOGY, Interval History, Dear Dr. Easterling, pleasure, Cardiology Service, history, chest pain, evaluated at Klickitat, treadmill test, unconcerning, episodes of chest, interesting, sharp pain, chest followed, rapid palpitations associated, jaw aching, sudden onset, sudden offset, day, ...
>
> **Snippet from the generated note:**
> ... Dear Dr. Easterling,
> ... When exercising on the tread mill, the patients developed ST elevation on both lateral limbs of the ECG tracings. There have been several spells of chest pains without any related symptoms such as dyspnea shortness of breath, cough, wheezing or nausea. These occurred intermittently throughout the past few days before coming to the emergency room this morning, these attacks began about six hours after his last episode of angina. One episode of severe right sided chest discomfort with radiation into the left arm along with slight tightening sensation of the neck lasted approximately five minutes...
> ...We appreciate your prompt evaluation regarding these complaints.
> Sincerely yours, **Martin** Easterling, M.D.

Data availability

The I2B2-2014 dataset requires authorized access. The dataset can be obtained through the DBMI data portal. The source code is available at https://github.com/Privacy4all/dp_llm_finetuning.

References

Abacha, A. B., Yim, W.-w., Fan, Y., & Lin, T. (2023). An empirical study of clinical note generation from doctor-patient encounters In *Proceedings of the 17th conference of the European chapter of the association for computational linguistics* (pp. 2291–2302).

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security* (pp. 308–318).

Ahmed, T., Aziz, M. M. A., & Mohammed, N. (2020). De-identification of electronic health record using neural network. *Scientific Reports*, *10*(1), 18600.

Al Aziz, M. M., Ahmed, T., Faequa, T., Jiang, X., Yao, Y., & Mohammed, N. (2021). Differentially private medical texts generation using generative neural networks. *ACM Transactions on Computing for Healthcare (HEALTH)*, *3*(1), 1–27.

Baumel, T., Manoel, A., Jones, D., Su, S., Inan, H., Sim, R., et al. (2024). Controllable synthetic clinical note generation with privacy guarantees. arXiv preprint arXiv:2409.07809.

Bo, H., Ding, S. H. H., Fung, B. C. M., & Iqbal, F. (2021). ER-AE: Differentially private text generation for authorship anonymization. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies* (pp. 3997–4007). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.naacl-main.314, URL https://aclanthology.org/2021.naacl-main.314.

Boudin, F. (2016). PKE: An open source python-based keyphrase extraction toolkit. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: system demonstrations* (pp. 69–73).

Boulanger, H., Hiebel, N., Ferret, O., Fort, K., & Névéol, A. (2024). Using structured health information for controlled generation of clinical cases in french. In *The 6th clinical natural language processing workshop at NAACL 2024*. clinicalNLP 2024.

Cai, X., Perez-Concha, O., Coiera, E., Martin-Sanchez, F., Day, R., Roffe, D., & Gallego, B. (2016). Real-time prediction of mortality, readmission, and length of stay using electronic health record data. *Journal of the American Medical Informatics Association*, *23*(3), 553–561.

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*, *54*(6), 1–38.

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., & Jatowt, A. (2020). YAKE! keyword extraction from single documents using multiple local features. *Information Sciences*, *509*, 257–289.

Chin-Yew, L. (2004). Rouge: A package for automatic evaluation of summaries. In *Proceedings of the workshop on text summarization branches out, 2004*.

Chuang, Y.-S., Sarkar, A. R., Hsu, Y.-C., Mohammed, N., & Jiang, X. (2025). Robust privacy amidst innovation with large language models through a critical assessment of the risks. *Journal of the American Medical Informatics Association*, *32*(5), 885–892.

El Emam, K., Mosquera, L., & Hoptroff, R. (2020). *Practical synthetic data generation: Balancing privacy and the broad availability of data*. O'Reilly Media.

Forcier, M. B., Gallois, H., Mullan, S., & Joly, Y. (2019). Integrating artificial intelligence into health care through data access: Can the GDPR act as a beacon for policymakers? *Journal of Law and the Biosciences*, *6*(1), 317–335.

Garfinkel, S., Garfinkel, S., Near, J., Dajani, A., Singer, P., & Guttman, B. (2023). *De-identifying government datasets: Techniques and governance*. US Department of Commerce, National Institute of Standards and Technology.

Hernandez, M., Epelde, G., Alberdi, A., Cilla, R., & Rankin, D. (2022). Synthetic data generation for tabular health records: A systematic review. *Neurocomputing*, *493*, 28–45.

Jiang, Y., Wang, Y., Zeng, X., Zhong, W., Li, L., Mi, F., Shang, L., Jiang, X., Liu, Q., & Wang, W. (2024). Followbench: A multi-level fine-grained constraints following benchmark for large language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 4667–4688).

Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S. N., & Weller, A. (2022). Synthetic data–what, why and how?. arXiv preprint arXiv:2205.03257.

Karunanayake, N. (2025). Next-generation agentic AI for transforming healthcare. *Informatics and Health*, *2*(2), 73–83.

Li, J., Zhou, Y., Jiang, X., Natarajan, K., Pakhomov, S. V., Liu, H., & Xu, H. (2021). Are synthetic clinical notes useful for real natural language processing tasks: A case study on clinical entity recognition. *Journal of the American Medical Informatics Association*, *28*(10), 2193–2201.

Liu, F., Cheng, Z., Chen, H., Wei, Y., Nie, L., & Kankanhalli, M. (2022). Privacy-preserving synthetic data generation for recommendation systems. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval* (pp. 1379–1389).

Lukas, N., Salem, A., Sim, R., Tople, S., Wutschitz, L., & Zanella-Béguelin, S. (2023). Analyzing leakage of personally identifiable information in language models. In *2023 IEEE symposium on security and privacy* (pp. 346–363). IEEE.

Ness, R. B., et al. (2008). Influence of the HIPAA privacy rule on health research. *Obstetrical & Gynecological Survey*, *63*(4), 236–237.

Oh, B.-D., Kim, G.-Y., Kim, C., & Kim, Y.-S. (2024). How to use language models for synthetic text generation in cerebrovascular disease-specific medical reports. In *Proceedings of the 1st workshop on personalization of generative AI systems* pERSONALIZE 2024, (pp. 10–17).

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the association for computational linguistics* (pp. 311–318). Association for Computational Linguistics.

Piskorski, J., Stefanovitch, N., Jacquet, G., & Podavini, A. (2021). Exploring linguistically-lightweight keyword extraction techniques for indexing news articles in a multilingual set-up. In *Proceedings of the EACL hackashop on news media content analysis and automated report generation* (pp. 35–44).

Qaiser, S., & Ali, R. (2018). Text mining: Use of TF-IDF to examine the relevance of words to documents. *International Journal of Computer Applications*, *181*(1), 25–29.

Sarkar, A. R., Chuang, Y.-S., Mohammed, N., & Jiang, X. (2024). De-identification is not enough: a comparison between de-identified and synthetic clinical notes. *Scientific Reports*, *14*(1).

Sinha, A., & McKenna, R. (2025). VaultGemma: The world's most capable differentially private LLM. Google Research. URL https://research.google/blog/vaultgemma-the-worlds-most-capable-differentially-private-llm.

Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., Krueger, G., Kim, J. W., Kreps, S., et al. (2019). Release strategies and the social impacts of language models. arXiv preprint arXiv:1908.09203.

Stubbs, A., Kotfila, C., & Uzuner, Ö. (2015). Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task track 1. *Journal of Biomedical Informatics*, *58*, S11–S19.

Urbain, J., Kowalski, G., Osinski, K., Spaniol, R., Liu, M., Taylor, B., Waitman, L. R., et al. (2022). Natural language processing for enterprise-scale de-identification of protected health information in clinical notes. *AMIA Summits on Translational Science Proceedings*, *2022*, 92.

U. S. Department of Health and Human Services (2025). Guidance regarding methods for de-identification of protected health information in accordance with the health insurance portability and accountability act (HIPAA) privacy rule. URL https://www.hhs.gov/hipaa/for-professionals/special-topics/de-identification/index.html. (Accessed 10 October 2025).

Wan, K., Mu, H., Hao, R., Luo, H., Gu, T., & Chen, X. (2025). A cognitive writing perspective for constrained long-form text generation. arXiv preprint arXiv:2502.12568.

Wutschitz, L., Inan, H. A., & Manoel, A. (2022). Dp-transformers: Training transformer models with differential privacy.

Yang, X., Lyu, T., Li, Q., Lee, C.-Y., Bian, J., Hogan, W. R., & Wu, Y. (2019). A study of deep learning methods for de-identification of clinical notes in cross-institute settings. *BMC Medical Informatics and Decision Making*, *19*, 1–9.

Ye, J., Yao, L., Shen, J., Janarthanam, R., & Luo, Y. (2020). Predicting mortality in critically ill patients with diabetes using machine learning and clinical notes. *BMC Medical Informatics and Decision Making*, *20*, 1–7.

Yim, W.-w., & Yetisgen, M. (2021). Towards Automating Medical Scribing : Clinic Visit Dialogue2Note Sentence Alignment and Snippet Summarization. In *Proceedings of the second workshop on natural language processing for medical conversations* (pp. 10–20). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.nlpmc-1.2.

Yu, D., Naik, S., Backurs, A., Gopi, S., Inan, H. A., Kamath, G., Kulkarni, J., Lee, Y. T., Manoel, A., Wutschitz, L., et al. (2021). Differentially private fine-tuning of language models. arXiv preprint arXiv:2110.06500.

Yue, X., Inan, H. A., Li, X., Kumar, G., McAnallen, J., Shajari, H., Sun, H., Levitan, D., & Sim, R. (2022). Synthetic text generation with differential privacy: A simple and practical recipe. arXiv preprint arXiv:2210.14348.

Zhou, Y., Ringeval, F., & Portet, F. (2023). A survey of evaluation methods of generated medical textual reports. In *Proceedings of the 5th clinical natural language processing workshop* (pp. 447–459). Toronto, Canada: Association for Computational Linguistics.