Evaluating Privacy Risks in Synthetic Clinical Text Generation in Spanish

Luis Miranda[13] **Jocelyn Dunstan**[13] **Matias Toro**[23]
Federico Olmedo[23] **Félix Melo**[2]
[1]Ponficia Universidad Católica de Chile [2]Universidad de Chile
[3]Instituto Milenio Fundamento de los Datos
{lmirandn,jdunstan}@uc.cl
{mtoro,federico.olmedo,felix.melo}@dcc.uchile.cl

Abstract

Leveraging medical data for Deep Learning models holds great potential, but ensuring the protection of sensitive patient information is paramount in the clinical domain. A widely used approach to balance data utility and privacy is the generation of synthetic text with Large Language Models (LLMs) under the framework of differential privacy (DP). Techniques like Differentially Private Stochastic Gradient Descent (DP-SGD) are typically considered to provide privacy guarantees, but they rely on specific conditions. This research demonstrates how memorization in LLMs can deteriorate when these privacy safeguards are not fully met, increasing the risk of personal and sensitive information being leaked in synthetic clinical reports. Addressing these vulnerabilities could enhance the reliability of DP in protecting clinical text data while maintaining its utility.

1 Introduction

The utilization of Electronic Health Records (EHRs) for Natural Language Processing (NLP) offers numerous benefits, particularly in enhancing healthcare research and outcomes [7]. However, protecting the privacy of the patients in these records is crucial. Privacy is recognized as a core human right in the Universal Declaration of Human Rights, placing the control individuals have over their personal information on par with the authority exercised by corporations and governments [14].

According to the 2021 Annual Report of the United Nations High Commissioner, privacy reflects human dignity and plays a critical role in safeguarding individual autonomy and identity. In today's digital age, privacy concerns are even more pronounced as personal data—often considered a valuable commodity—can be collected, sold, and potentially misused. This is particularly concerning when sensitive health data is involved (e.g., apps that collect reproductive information, or dating apps that ask for HIV status) [6]. The mishandling of such data not only threatens privacy but can also foster discrimination and erode human dignity.

There are several techniques to protect patient privacy in EHRs, such as Named Entity Recognition (NER) for de-identification or pseudo-anonymization [3, 17, 16]. However, synthetic text generation with Differential Privacy (DP) is often preferred due to its formal privacy guarantees and its widespread use [20, 10, 19, 2].

Synthetic text refers to artificially generated text that mimics human language and content. One way to create it is by using Large Language Models (LLMs), which generate text through "next-token prediction." This process involves predicting the next word in a sentence based on the previous ones, allowing the model to generate coherent text. In this context, the goal is to create realistic synthetic

Latinx in AI @ NeurIPS-24 (LXAI 2024).

Electronic Health Records (EHRs) that are similar to original EHRs, making them useful for research and other purposes. To achieve this, an LLM can be trained using real EHR data.

Training an LLM involves exposing the model to a dataset and adjusting its parameters based on the patterns it learns. However, during this process, the model might memorize personal information and reproduce it [4], which is critical when dealing with clinical data. To prevent this, DP can be applied. DP, in essence, ensures that individual data points within a dataset do not significantly influence the outcome of an algorithm, protecting information quantified by a level of privacy ϵ [9]. A common technique used for training an LLM with DP is Differentially Private Stochastic Gradient Descent (DP-SGD), which adds noise during training to prevent memorization, ensuring both privacy and utility [1, 11].

However, the mere use of DP-SGD often leads to an assumption of privacy guarantees, but in practice, is frequently overlooked. DP-SGD provides "sample-level" privacy [18, 11], meaning it protects individual data points as long as the same individual does not appear in multiple samples. In clinical datasets, this assumption is unfeasible, as the same individual may be represented in multiple samples. This raises serious concerns about the true effectiveness of DP in such contexts.

To address potential privacy concerns, it is important to evaluate privacy beyond standard guarantees, such as by assessing the level of memorization. Previous research has primarily focused on measuring model memorization and the leakage of sensitive information in synthetic data, particularly the leakage of isolated pieces of Personally Identifiable Information (PII) [20, 5]. Building on these studies, this work introduces a novel method for analyzing the memorization of LLMs and the risk of information leakage in synthetic EHRs generated in Spanish. This presents unique challenges specific to the language (e.g. the more frequent use of gendered terms throughout sentences).

2 Experimental Setup

In this study we used the MEDDOCAN dataset [12] (available here), which consists of 1,000 manually crafted Spanish clinical reports enriched with personal information and annotated with NER for PII and sensitive data. For computing limitations, the final dataset used consisted of 1000 reports, divided into 750 documents for training and 250 for validation. These documents are used to analyze information leakage at the document level. We conducted the experiments using the LLM `meta-llama/Meta-Llama-3.1-8B-Instruct` [8].

The training was executed with the following parameters: 7 epochs, batch size of 2, gradient accumulation steps of 1, LoRA dimension of 4, LoRA alpha, and a learning rate of 3e-4. Additionally, the training of the LLMs was performed on 2 NVIDIA RTX 6000 Ada Generation GPUs.

3 Methodology

The training used DP-SGD, which adds noise to gradients during the training process to safeguard the original data's privacy [1]. We trained the models using identical parameters across different dataset versions, each with varying levels of differential privacy. ϵ, a key parameter in differential privacy, measures privacy loss, with lower values providing stronger protection. The used values are $\epsilon = 8$, 16, and ∞ (no privacy).

After training, 500 synthetic documents were generated with each model. These documents were analyzed to assess memorization and evaluate the quality and utility of the generated text. The generation process was standardized putting the same training parameters to ensure comparable results across models. Finally, we applied various metrics to examine the privacy-utility trade-off and the extent of memorization.

3.1 Utility Metrics

The utility of the synthetic documents generated by each model was evaluated using key metrics such as MAUVE and perplexity (PPL). MAUVE [15] measures the quality and diversity of generated text using divergence frontiers, reflecting how closely the synthetic data aligns with the distribution of real text. PPL assesses how well a model predicts a sample, with lower values indicating better performance [13]. These metrics were used to evaluate the impact of differential privacy on the quality and coherence of the generated EHRs.

3.2 Leakage of Sensitive Information

To evaluate the impact of synthetic text generation with DP-SGD when private patient information is repeated across documents, we adapted the "canary" experiment [5]. This involved injecting a "canary" sentence containing a single piece of PII repeated across documents, allowing us to track how often it appeared in generated samples. In our version, two pieces of information—a reference to positive HIV as sensitive data and the name "Lopez Perez" to link it to personal information—were embedded into 0, 50, and 200 documents. We then counted how often this information appeared in the generated samples. In this way, we assess the memorization of links between sensitive data and individuals rather than the memorization of individual data points, which is crucial in the context of sensitive clinical data, as the ability to link sensitive information (such as an illness or medical history) to an individual must be protected.

4 Results and Discussion

Inj. Can.	ϵ	**MAUVE**		**PPL**		**Leaked Can.**	
		Model 1	**Model 2**	**Model 1**	**Model 2**	**Model 1**	**Model 2**
0	8	0.48	0.84	7.84±0.42	8.27±0.43	0	0
0	16	0.55	0.88	7.56± 0.21	8.44±0.39	0	0
0	∞	0.83	0.89	6.02±0.29	4.73±0.24	0	0
50	8	0.47	0.76	7.76±0.44	8.34±0.37	0	1
50	16	0.59	0.80	7.57± 0.23	8.76±0.32	2	2
50	∞	0.82	0.87	6.06±0.27	**4.39±0.07**	76	120
200	8	0.41	0.81	8.05±0.35	8.32±0.39	1	1
200	16	0.55	0.85	7.75±0.30	8.45±0.19	3	8
200	∞	0.84	**0.95**	5.72±0.32	4.91±0.64	103	331

Table 1: Privacy-utility evaluation results for Model 1 : `mistralai/Mistral-7B-v0.1` and Model 2 : `meta-llama/Meta-Llama-3.1-8B-Instruct`. The models were evaluated across varying privacy levels ($\epsilon = 8, 16, \infty$) and different quantities of injected canaries (Inj. Can.). The evaluation metrics include MAUVE, Perplexity (PPL), and the number of leaked canaries (Leaked Can.) in the 500 synthetic generated data.

Table 1 shows the results of synthetically generated texts evaluated by models trained with different privacy levels (ϵ = 8, 16, ∞) and varying numbers of injected canaries (0, 50, 200). The utility metrics, MAUVE and PPL, reveal that as privacy increases (lower ϵ), MAUVE decreases and PPL rises, indicating lower text quality and diversity due to the added noise from DP-SGD. Additionally, Model 1 displays lower PPL but also a lower MAUVE than Model 2, suggesting that while the text generated by Model 1 is more predictable, it is less natural and diverse—consistent with the definitions of MAUVE and PPL. Except in the case where there is no privacy (ϵ = ∞), where Model 1 shows both lower MAUVE and higher PPL than Model 2.

Regarding canary leakage, the more frequently a canary (e.g., name and disease) is injected into the training data, the more it appears in the generated texts, with over 15% of the text containing personal information in some cases. However, when differential privacy is applied, this percentage drops to less than 2%. Despite this reduction, conditions for privacy guarantees are still violated, as differential privacy requires that no individual appear in more than one sample. Consequently, the generated text would be leaking that the individual with the surname "Lopez Perez" is HIV positive.

5 Conclusions and Limitations

While DP-SGD is widely believed to provide strong privacy guarantees, our findings reveal that memorization in LLMs occurs when those privacy guarantees are compromised, particularly in cases where the same individual appears across multiple samples—an aspect rarely considered when applying these methods. This was done by injecting the same linked personal and sensitive information multiple times in the training data of an LLM and then quantifying the leakage of this information in synthetic generated data by the model, offering a more comprehensive view of information leakage across entire documents, rather than focusing on individual PII entities. This raises concerns about the effectiveness of DP in clinical datasets, where privacy protection is paramount. Despite these challenges, DP can still serve as a valuable tool for safeguarding individuals if its conditions are properly fulfilled.

It is important to highlight a limitation of this work: while the goal is to analyze the level of memorization in an entire clinical report, this study only focuses on whether a name paired with a disease appears in any generated text. Although this approach is closer to identifying more than just a name, it still falls short of capturing a complete clinical record. Clinical records (for the dataset used in this study) typically include more detailed information such as medical history, addresses, phone numbers, prescriptions, medications, diagnoses, and more. Therefore, a more robust method for analyzing memorization in clinical reports that takes into account all this additional information is needed.

As future work, we propose employing feature extraction and Named Entity Recognition (NER) algorithms to detect personal and sensitive information in each synthetically generated document. Once extracted, this information can be used to analyze memorization across different differentially private algorithms for generating synthetic clinical data, comparing how these techniques perform in both the original and synthetic texts.

References

[1] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, page 308–318, New York, NY, USA, 2016. Association for Computing Machinery.

[2] N. C. Abay, Y. Zhou, M. Kantarcioglu, B. Thuraisingham, and L. Sweeney. Privacy preserving synthetic data release using deep learning. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18*, pages 510–526. Springer, 2019.

[3] C. Aracena, L. Miranda, T. Vakili, F. Villena, T. Quiroga, F. Núñez-Torres, V. Rocco, and J. Dunstan. A privacy-preserving corpus for occupational health in Spanish: Evaluation for NER and classification tasks. In T. Naumann, A. Ben Abacha, S. Bethard, K. Roberts, and D. Bitterman, editors, *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 111–121, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[4] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery.

[5] N. Carlini, C. Liu, Ú. Erlingsson, J. Kos, and D. Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 267–284, Santa Clara, CA, Aug. 2019. USENIX Association.

[6] D. Citron. *The Fight for Privacy: Protecting Dignity, Identity and Love in the Digital Age*. Random House, 2022.

[7] H. Dalianis. *Clinical Text Mining: Secondary Use of Electronic Patient Records*. Springer International Publishing, 2018.

[8] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, and (...). The llama 3 herd of models, 2024.

[9] C. Dwork. Differential privacy. In M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, editors, *Automata, Languages and Programming*, pages 1–12, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[10] J. Flemings and M. Annavaram. Differentially private knowledge distillation via synthetic text generation. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Findings of the Association for Computational Linguistics ACL 2024*, pages 12957–12968, Bangkok, Thailand and virtual meeting, Aug. 2024. Association for Computational Linguistics.

[11] O. Klymenko, S. Meisenbacher, and F. Matthes. Differential privacy in natural language processing the story so far. In O. Feyisetan, S. Ghanavati, P. Thaine, I. Habernal, and F. Mireshghallah, editors, *Proceedings of the Fourth Workshop on Privacy in Natural Language Processing*, pages 1–11, Seattle, United States, July 2022. Association for Computational Linguistics.

[12] M. Marimon, A. Gonzalez-Agirre, A. Intxaurrondo, H. Rodriguez, J. L. Martin, M. Villegas, and M. Krallinger. Automatic De-identification of Medical Texts in Spanish: the MEDDOCAN Track, Corpus, Guidelines, Methods and Evaluation of Results. In *IberLEF@ SEPLN*, pages 618–638, 2019.

[13] A. Miaschi, D. Brunato, F. Dell'Orletta, and G. Venturi. What makes my model perplexed? a linguistic investigation on neural language models perplexity. In E. Agirre, M. Apidianaki, and I. Vulić, editors, *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 40–47, Online, June 2021. Association for Computational Linguistics.

[14] Z. Nampewo, J. H. Mike, and J. Wolff. Respecting, protecting and fulfilling the human right to health. *International Journal for Equity in Health*, 21(1), Mar. 2022.

[15] K. Pillutla, S. Swayamdipta, R. Zellers, J. Thickstun, S. Welleck, Y. Choi, and Z. Harchaoui. Mauve: Measuring the gap between neural text and human text using divergence frontiers. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 4816–4828. Curran Associates, Inc., 2021.

[16] T. Vakili, A. Henriksson, and H. Dalianis. End-to-End Pseudonymization of Fine-Tuned Clinical BERT Models, Sept. 2023.

[17] S. Verkijk and P. Vossen. Efficiently and Thoroughly Anonymizing a Transformer Language Model for Dutch Electronic Health Records: a Two-Step Method. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1098–1103, Marseille, France, June 2022. European Language Resources Association.

[18] Y. Wang, Q. Wang, L. Zhao, and C. Wang. Differential privacy in deep learning: Privacy and beyond. *Future Generation Computer Systems*, 148:408–424, 2023.

[19] B. Xin, Y. Geng, T. Hu, S. Chen, W. Yang, S. Wang, and L. Huang. Federated synthetic data generation with differential privacy. *Neurocomputing*, 468:1–10, 2022.

[20] X. Yue, H. Inan, X. Li, G. Kumar, J. McAnallen, H. Shajari, H. Sun, D. Levitan, and R. Sim. Synthetic text generation with differential privacy: A simple and practical recipe. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1321–1342, Toronto, Canada, July 2023. Association for Computational Linguistics.