Analyzing the Leakage of Personal Information in Synthetic Clinical Spanish Texts

Anonymous ACL submission

Abstract

Using medical data for Deep Learning models can be highly beneficial, but protecting sensitive and personal patient information in the clinical field is critical. One of the most common ways to use this data while protecting patient privacy is by generating synthetic text with Large Language Models (LLMs) using differential privacy (DP). Although DP techniques, such as the Differentially Private Stochastic Gradient Descent (DP-SGD), are often assumed to guarantee privacy, they require specific conditions to be met. This study shows how memorization in LLMs can occur when these privacy guarantees are compromised, potentially leading to the leakage of personal and sensitive information in generated clinical reports. If these gaps are addressed, DP could offer more reliable safeguards for clinical data, improving privacy without sacrificing 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 (Dalianis, 2018). 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 (Nampewo et al., 2022).

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) (Citron, 2022). 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 (Aracena et al., 2024; Verkijk and Vossen, 2022; Vakili et al., 2023). However, synthetic text generation with Differential Privacy (DP) is often preferred for due to its formal privacy guarantees and its widespread use (Yue et al., 2023; Flemings and Annavaram, 2024; Xin et al., 2022; Abay et al., 2019).

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 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 (Bender et al., 2021), 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 (Dwork, 2006). 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 memoriza-

Injected 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 (Injected Can.). The evaluation metrics include MAUVE, Perplexity (PPL), and the number of leaked canaries (Leaked Can.) in the 500 synthetic generated data.

tion, ensuring both privacy and utility (Abadi et al., 2016; Klymenko et al., 2022).

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 (Wang et al., 2023; Klymenko et al., 2022), 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) (Yue et al., 2023; Carlini et al., 2019). 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 (Marimon et al., 2019), 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 750 reports, divided into 500 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 LLMs `mistralai/Mistral-7B-v0.1` (Jiang et al., 2023) and `meta-llama/Meta-Llama-3.1-8B-Instruct` (Dubey et al., 2024).

3 Methodology

The training used DP-SGD, which adds noise to gradients during the training process to safeguard the original data's privacy (Abadi et al., 2016). 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 (Pillutla et al., 2021) 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 (Miaschi et al., 2021). 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 (Carlini et al., 2019). 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

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 ($\epsilon = \infty$), 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 Future Work

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.

As future work, we propose employing feature extraction and NER algorithms for personal and sensitive information in each synthetically generated text to further analyze memorization in various differentially private algorithms for generating synthetic clinical data.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. 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. Association for Computing Machinery.

Nazmiye Ceren Abay, Yan Zhou, Murat Kantarcioglu, Bhavani Thuraisingham, and Latanya Sweeney. 2019.

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.

Claudio Aracena, Luis Miranda, Thomas Vakili, Fabián Villena, Tamara Quiroga, Fredy Núñez-Torres, Victor Rocco, and Jocelyn Dunstan. 2024. A privacy-preserving corpus for occupational health in Spanish: Evaluation for NER and classification tasks. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 111–121, Mexico City, Mexico. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. 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. Association for Computing Machinery.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. 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. USENIX Association.

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

Hercules Dalianis. 2018. *Clinical Text Mining: Secondary Use of Electronic Patient Records*. Springer International Publishing.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, and (...). 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Cynthia Dwork. 2006. Differential privacy. In *Automata, Languages and Programming*, pages 1–12, Berlin, Heidelberg. Springer Berlin Heidelberg.

James Flemings and Murali Annavaram. 2024. Differentially private knowledge distillation via synthetic text generation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12957–12968, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Oleksandra Klymenko, Stephen Meisenbacher, and Florian Matthes. 2022. Differential privacy in natural language processing the story so far. In *Proceedings of the Fourth Workshop on Privacy in Natural Language Processing*, pages 1–11, Seattle, United States. Association for Computational Linguistics.

Montserrat Marimon, Aitor Gonzalez-Agirre, Ander Intxaurrondo, Heidy Rodriguez, Jose Lopez Martin, Marta Villegas, and Martin Krallinger. 2019. Automatic De-identification of Medical Texts in Spanish: the MEDDOCAN Track, Corpus, Guidelines, Methods and Evaluation of Results. In *IberLEF@ SEPLN*, pages 618–638.

Alessio Miaschi, Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. 2021. What makes my model perplexed? a linguistic investigation on neural language models perplexity. In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 40–47, Online. Association for Computational Linguistics.

Zahara Nampewo, Jennifer Heaven Mike, and Jonathan Wolff. 2022. Respecting, protecting and fulfilling the human right to health. *International Journal for Equity in Health*, 21(1).

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. In *Advances in Neural Information Processing Systems*, volume 34, pages 4816–4828. Curran Associates, Inc.

Thomas Vakili, Aron Henriksson, and Hercules Dalianis. 2023. End-to-End Pseudonymization of Fine-Tuned Clinical BERT Models.

Stella Verkijk and Piek Vossen. 2022. 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. European Language Resources Association.

Yanling Wang, Qian Wang, Lingchen Zhao, and Cong Wang. 2023. Differential privacy in deep learning: Privacy and beyond. *Future Generation Computer Systems*, 148:408–424.

Bangzhou Xin, Yangyang Geng, Teng Hu, Sheng Chen, Wei Yang, Shaowei Wang, and Liusheng Huang. 2022. Federated synthetic data generation with differential privacy. *Neurocomputing*, 468:1–10.

Xiang Yue, Huseyin Inan, Xuechen Li, Girish Kumar, Julia McAnallen, Hoda Shajari, Huan Sun, David Levitan, and Robert Sim. 2023. Synthetic text generation with differential privacy: A simple and practical recipe. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1321–1342, Toronto, Canada. Association for Computational Linguistics.