User-Centric Evaluation of LLM-Based Pseudonymization in Medical Texts

Stig Hellemans[0000-0001-9441-3882], Pieter Meysman[0000-0001-5903-633X], and Kris Laukens[0000-0002-8217-2564]

University of Antwerp, Adrem Data Lab

Keywords: Natural Language Processing · Privacy preservation · Evaluation Frameworks

Protecting patient privacy while enabling data-driven research requires reliable de-identification. A common approach is pseudonymization, where identifiers are replaced with re-linkable pseudonyms under strict conditions [1]. Training such systems still depends on large manually annotated datasets. LLMs can reduce this burden by generating pre-annotations that humans refine rather than create from scratch [6, 7]. To assess true effort savings, we adopt a user-centric metric inspired by the annotation edit distance [5], which quantifies the number of user actions required to reach gold-standard quality—providing a more realistic measure of annotation workload than traditional precision or recall. While structured data pseudonymization methods like the Shift and Truncate (SANT) approach [2] mitigate temporal leakage through per-patient random offsets and truncation, extending such principles to unstructured text remains challenging. Free-text may reveal sensitive information through indirect means, such as temporal event references or linguistic cues, that demand adversarial testing to verify de-identification robustness [4]. Even seemingly safe transformations can provide false security if latent temporal patterns remain [3]. Effective anonymization must therefore balance privacy protection with clinical usefulness—preserving information valuable for research while preventing reidentification.

We used 100 synthetic medical texts containing realistic clinical narratives, patient identifiers, dates, and other protected health information (PHI). Each document was manually annotated to create gold-standard labels for evaluation.

We evaluated two grammar-based large language model (LLM) approaches for text anonymization, comparing them to a baseline condition without any pre-annotations (Fig. 1). The Finite-State Machine (FSM) method scans the text and wraps sensitive entities in span tags, achieving the highest accuracy but requiring full text rewriting, making it less token-efficient. The dictionary JSON grammar combines Waterman–Smith fuzzy matching with structured JSON outputs, a format commonly used by LLMs. JSON grammars also have the distinct advantage of compatibility with all OpenAI APIs, unlike custom EBNF grammars.

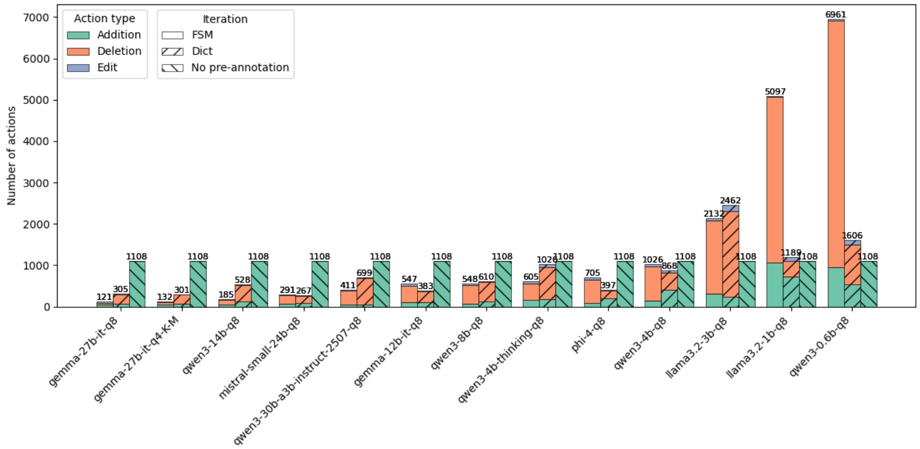

Fig. 1: User-based metric evaluation of LLMs and grammar approaches.

To better capture real annotation effort, we adopted the concept of annotation edit distance [5] into a user-centric metric that quantifies the number of actions such as additions, edits, and deletions required to reach gold-standard quality (Fig. 1). A user study also measured the distribution of durations for each action (Fig. 2). Using this metric, the gemma-27b-it-q8 model with the FSM grammar reduced the required annotation operations by up to tenfold, demonstrating substantial gains in annotation efficiency.

We also evaluated the robustness of different date-shifting strategies using Bayesian inference. Depending on the chosen strategy, adversaries could still exploit temporal cues such as weekdays (“Monday”) or event anchors (“New Year’s Eve”) to infer true dates. By sampling possible offsets and estimating posterior probabilities for each, we quantified how quickly the inference converged on the correct offset. Results show that poorly designed shifting schemes can leak substantial temporal information, underscoring the importance of carefully selecting and testing pseudonymization techniques.

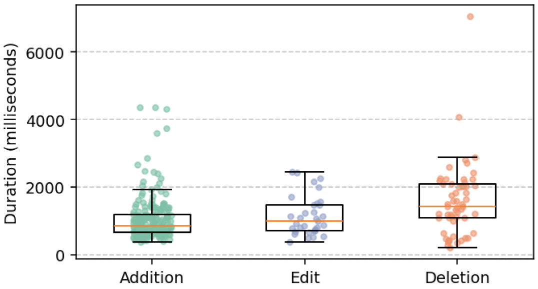

Fig. 2: Measured durations of actions.

This work was supported by the Flemish Government (Flanders AI Research Program) and the Research Fountation Flanders (FWO).

References

1. Abu Attieh, H., Müller, A., Wirth, F., *et al.*: Pseudonymization tools for medical research: a systematic review. *BMC Medical Informatics and Decision Making* **25**, 128 (2025).
2. Hripcsak, G., Mirhaji, P., Low, A.F., Malin, B.A.: Preserving temporal relations in clinical data while maintaining privacy. *Journal of the American Medical Informatics Association* **23**(6), 1040–1045 (2016).
3. Cimino, J.J.: The false security of blind dates: chrononymization's lack of impact on data privacy of laboratory data. *Applied Clinical Informatics* **3**(4), 392–403 (2012).
4. Morris, J.X., Campion, T.R., Nutheti, S.L., Peng, Y., Raj, A., Zabih, R., Cole, C.L.: DIRI: Adversarial Patient Reidentification with Large Language Models for Evaluating Clinical Text Anonymization. *AMIA Joint Summits on Translational Science Proceedings* **2025**, 355–364 (2025).
5. Eilbeck, K., Moore, B., Holt, C., Yandell, M.: Quantitative measures for the management and comparison of annotated genomes. *BMC Bioinformatics* **10**, 67 (2009).
6. Kuo, T.-T., Huh, J., Kim, J., El-Kareh, R.E., Singh, S., Feudjio Feupe, S., Kuri, V., Lin, G., Day, M.E., Ohno-Machado, L., Hsu, C.-N.: The impact of automatic pre-annotation in clinical note data element extraction – the CLEAN tool. *arXiv preprint* arXiv:1808.03806 (2018).
7. Xu, H.A., Loftsson, V., Kulynych, B., Kaabachi, B., Raisaro, J.L.: Accelerating clinical text annotation in underrepresented languages: A case study on text de-identification. *Studies in Health Technology and Informatics* **316**, 853–857 (2024).