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

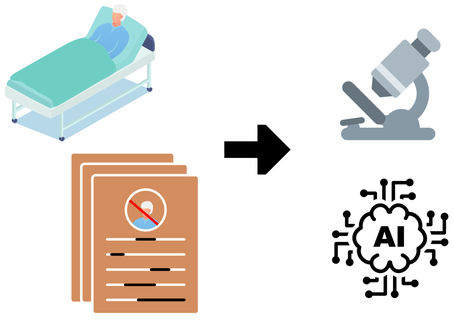
Stig Hellemans[0000-0001-9441-3882], Pieter Meysman[0000-0001-5903-633X], and Kris Laukens[0000-0002-8217-2564]

University of Antwerp | Adrem | Adrem Data Lab

Flanders AI Research Program

fwo

Abstract



Protecting patient privacy in clinical research demands effective pseudonymization of medical texts. This work evaluates LLM-based de-identification from a user-centric perspective, emphasizing actual annotation effort rather than traditional precision–recall metrics. Using 100 synthetic medical documents with gold-standard labels, we compared several grammar-driven structured output methods. By extending the idea of annotation edit distance, we quantified the concrete user actions required to reach gold-standard quality. The FSM method combined with the Gemma-27B-IT-Q8 model reduced annotation workload by up to tenfold, translating into a twofold increase in annotator efficiency in a real-world experiment. We also analyzed the robustness of temporal pseudonymization using Bayesian inference, showing that date shifting per patient or per encounter is generally necessary. Using a single fixed offset for an entire dataset increases re-identification risk as the number of samples grows. Overall, these results underscore the need for user-centric and adversarial approaches to develop practical, privacy-preserving anonymization tools.

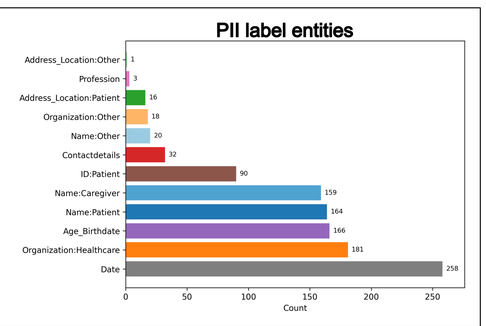
Methods

Semi-structured synthetic dataset generation

Generate a medical text given:
- PII labels: <PII-labels>
- Topic: <topic>
- Discipline: <discipline>

Custom PII generator

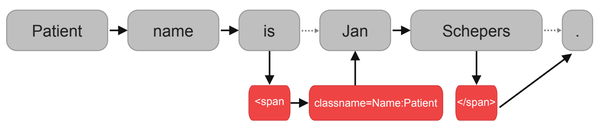
Diverse synthetic dataset



Pre-annotations using LLMs with structured outputs

(1) Finite State Machine (FSM)

Patient name is Jan Schepers.



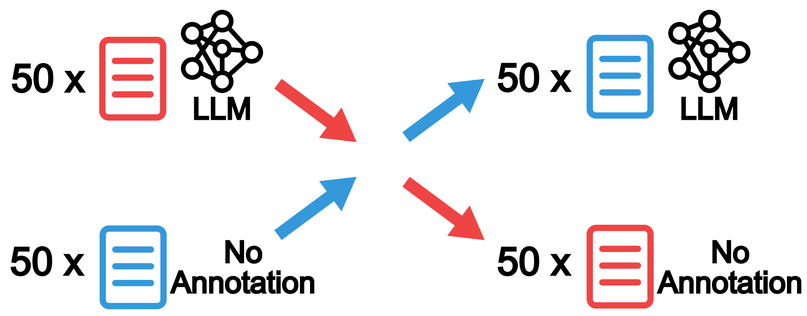
(2) Dictionary + Waterman Smith lookup

{"annotated_text": "Jan Schepers", "label": "Name:Patient"}

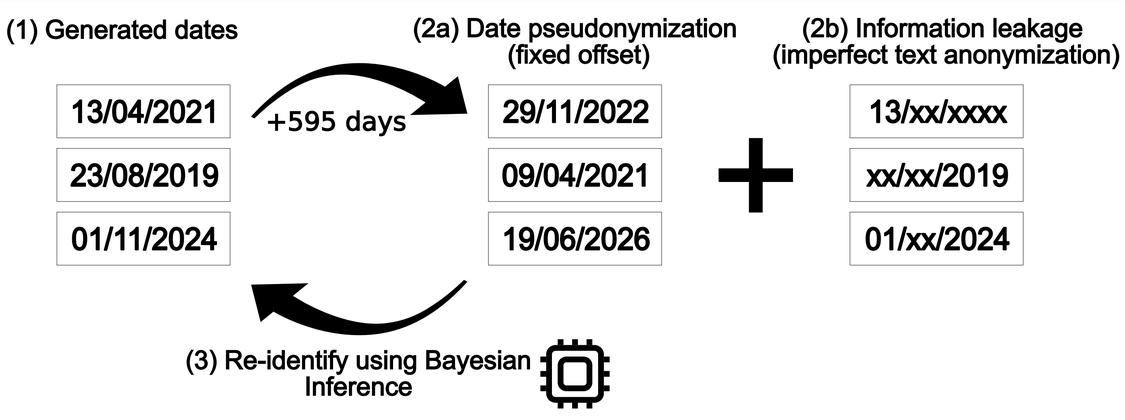
(3) Tuple + Waterman Smith lookup

["Jan Schepers", "Name:Patient"]

Cross-over annotation experiment

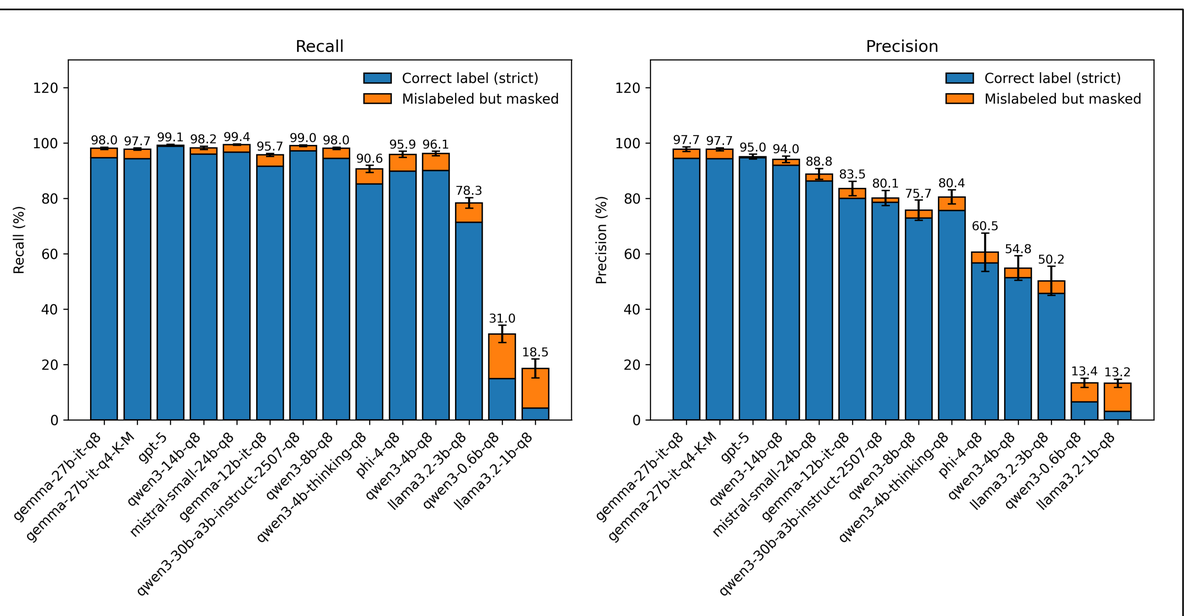


Re-identification risk simulation of dates



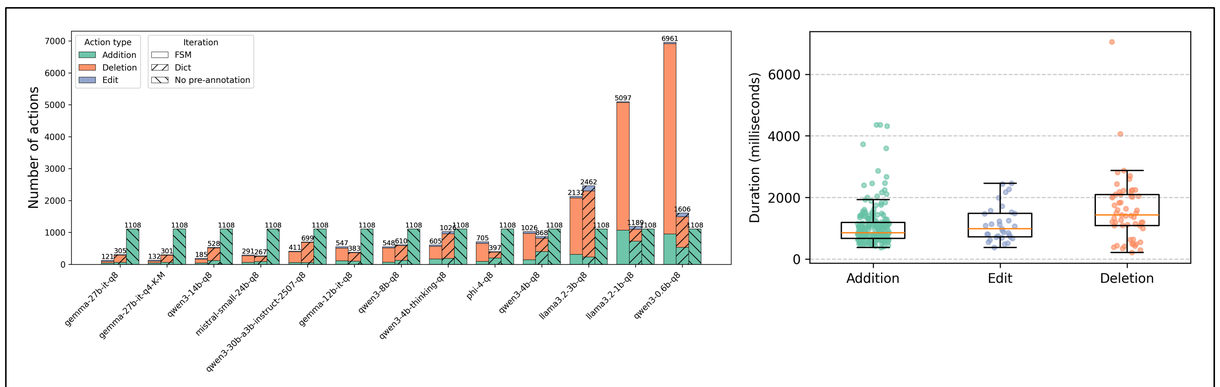
Results

Classic evaluation of LLM pre-annotations



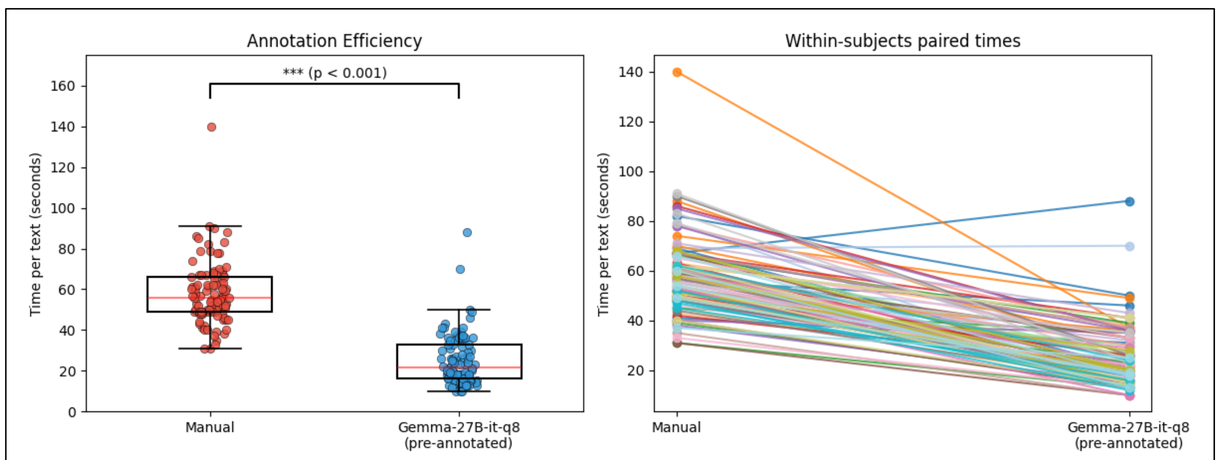
Automatic de-identification models are usually assessed with precision and recall. Where recall matters most because missed identifiers directly increase re-identification risk. These metrics, however, say little about the real manual workload. They don't reflect how much correction a pre-annotated dataset still needs.

Annotator-centric evaluation of LLM pre-annotations



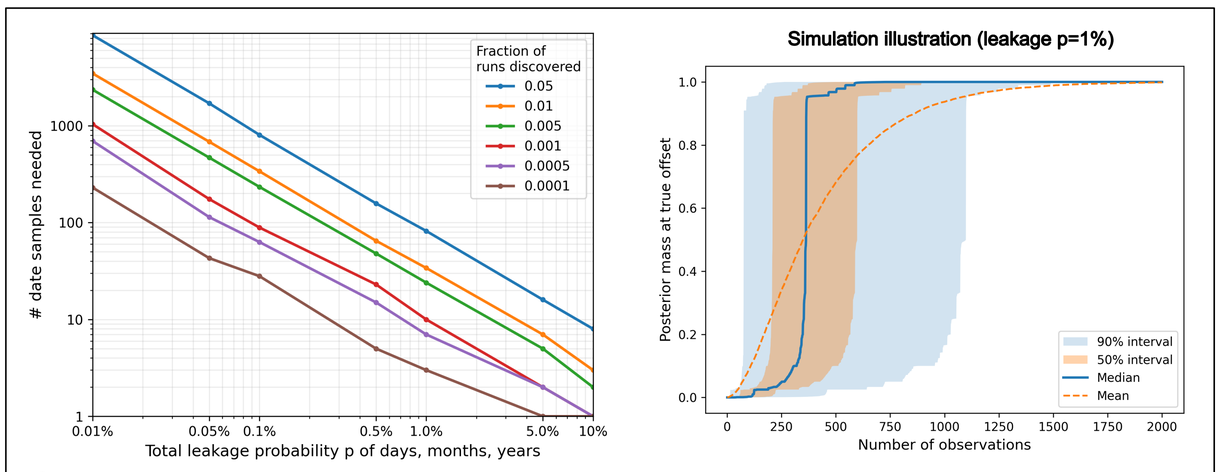
Using structured outputs, LLMs can substantially reduce annotation workload. Annotator workload can be estimated by the annotation edit distance, a user-centric metric, that quantifies the number of actions such as additions, edits, and deletions required to reach gold-standard quality. Among the smaller models, Gemma-27B-IT-Q8 delivers the strongest results, cutting the required actions by roughly a factor of ten. The FSM-based structured output method provides the best overall performance, though it is the least token-efficient. A user study indicates that each action typically costs 1–2 seconds of annotator time.

Measuring annotator efficiency



An annotation experiment demonstrates that LLM-based pre-annotations can double annotation efficiency. A cross-over design with randomly assigned texts was used to minimize bias. The annotator was medically trained and already familiar with the task, supported by an established annotation guide detailing all PII labels and edge cases. Annotations were performed using the open-source INCEpTION platform.

Re-identification risk of dates



Clinical texts only keep their value when relative timing is preserved, typically by shifting all dates by a fixed offset. Because this preserves temporal structure, leaking a single shifted date can expose the rest. The simulation estimates an upper bound on the information content, and thus re-identification risk, given a leakage probability p for day, month, or year. Running 100,000 trials per p gives enough statistical power for even very small risks. Since all dates share one offset, more date samples make it increasingly likely to recover the true offset, as shown in the right-hand figure.

Conclusions

A user study shows that LLMs can roughly double annotation efficiency. The benefit plateaus because human review remains essential, but strong pre-annotations shift the work from manual labeling to verification.

The re-identification simulation clarifies residual privacy risks when sharing clinical texts. It demonstrates how risk grows with more samples under a fixed date offset and how it can be reduced by limiting samples per offset or enforcing strict organizational controls.

Stig.hellemans@uantwerpen.be