Differentially Private De-identification of Dutch Clinical Notes: A Comparative Evaluation

Anonymous ACL submission

Abstract

Protecting patient privacy in clinical narratives is essential to enable research in healthcare. Historically, text de-identification pipelines employed named entity recognition (NER) to identify protected entities for redaction. Although methods based on differential privacy (DP) provide formal privacy guarantees, more recently also large language models (LLMs) are increasingly used for text de-identification in the clinical domain. In this work, we present the first comparative study of DP, NER, and LLMs for *Dutch* clinical text de-identification. We investigate these methods separately as well as hybrid strategies that apply NER or LLM preprocessing prior to DP, and assess performance in terms of privacy leakage and extrinsic evaluation (entity and relation classification). [1]

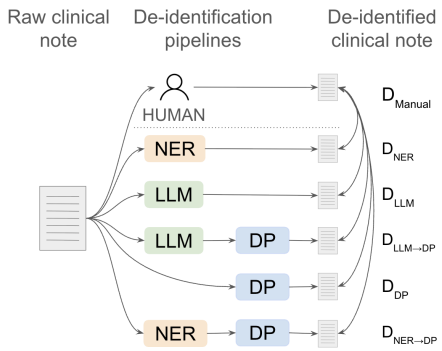

Figure 1: Overview of our comparative analysis. A raw document $\mathcal{D}_{\text{raw}}$ is de-identified using 5 different pipelines, which are evaluated against a manually de-identified version of the same document $\mathcal{D}_{\text{manual}}$. We use a range of open-source and proprietary LLMs that vary in architecture and size in our experiments.

1 Introduction

Ensuring privacy in clinical texts is critical to enable data sharing for healthcare research (Conduah et al., 2025). Privacy regulations like GDPR (European Parliament and Council of the European Union, 2016) and HIPAA (U.S. Department of Health and Human Services) require the redaction or *de-identification* of all personally identifiable information (PII) to protect patient privacy. Methods for PII de-identification based on named entity recognition (NER) have been extensively used for English and other languages (Grouin et al., 2015; Dernoncourt et al., 2017; Bourdois et al., 2021; Tchouka et al., 2022b; Wang et al., 2022). Recently, large language models (LLMs) have been applied to PII de-identification and have shown strong performance (Liu et al., 2023). These methods, despite their performance, do not provide any formal privacy guarantees. Differential privacy (DP; Dwork, 2006; Dwork and Roth, 2014) offers a principled mechanism with formal guarantees against privacy leakage when sharing a *privatized* dataset (Chatzikokolakis et al., 2013; Tong et al., 2025).

The tension in the existing literature lies in the fact that DP-based methods for text de-identification provide formal privacy guarantees but can impact utility (Yu et al., 2022; Yue et al., 2023; Tchouka et al., 2022b,a), whereas LLM-based methods become increasingly strong but do not provide any privacy guarantees (Pissarra et al., 2024; Yang et al., 2025). These works remain limited as they protect only detected entities rather than full text, offer weaker guarantees than DP, overlook instruction-tuned LLMs, and demand resources often unavailable in privacy-critical settings, while all being developed solely for English and thus untested in other languages.

In this study, we compare different PII de-identification methods using state-of-the-art NER (Zaratiana et al., 2024), LLMs (Liu et al., 2023), and synthetic data generation methods with DP guarantees (Chatzikokolakis et al., 2013; Tong

[1] We will release all code to reproduce our experiments upon acceptance.

et al., 2025) for Dutch clinical notes. See Figure 1 for an overview of our comparisions. Unlike prior works on English or, more recently, French clinical text (Tchouka et al., 2022d,c), our work is **the first study to investigate DP-based text anonymization in real-world Dutch clinical notes**, addressing a critical gap in the de-identification research. We evaluate the quality of the de-identified clinical notes *intrinsically* by quantifying the remaining residual PII, and *extrinsically* by measuring the performance in two downstream tasks: entity classification (drugs and disorders) and relation classification (adverse drug events). Finally, we also combine a strong de-identification method (e.g. NER- or LLM-based) as a preprocessing step with DP de-identification. We hypothesise that by doing this we can considerably reduce the DP noise necessary to achieve the same privacy guarantee, possibly leading to a better privacy-utility trade-off.

2 Methods and Experiments

Below we detail the dataset (§2.1), the de-identification modules (§2.2) used in our de-identification pipelines that generate privacy-preserving clinical notes (§2.3; Fig. 1), and the intrinsic (§2.4) and extrinsic evaluation on the entity and relation classification tasks (§2.5).

2.1 Dataset

We use the Dutch ADE dataset (Murphy et al., 2025a), a benchmark corpus containing 102 clinical notes from ICU patients annotated with entity-level labels—*drug* and *disorder* mentions—and relation labels between *drug-disorder* pairs—*adverse drug event* or *ADE* (when the drug caused the disorder as an adverse event) and *prescribing indication* (when the drug was prescribed to treat the disorder). We refer to the raw corpus as $\mathcal{D}_{\text{raw}}$ and to its manually de-identified version as $\mathcal{D}_{\text{manual}}$. $\mathcal{D}_{\text{raw}}$ consists of 107,110 words, 0.83% of which are identified as containing PII. The PII has been redacted and replaced by 17 different placeholders, each indicating one type of PII in the data. Please refer to Appendix B.1 for the full list of PII types in our dataset and other details.

2.2 De-identification Modules

We apply NER-based, LLM-based, and two DP-based text methods to de-ideintify PII: metric differential privacy (Metric-DP; Chatzikokolakis et al., 2013), and RANTEXT (Tong et al., 2025).

NER-based de-identification We use the pre-trained model `gliner-multi-v2.1`.[2] We provide it with our target entity types (see Table 1) and apply it to extract labeled spans from clinical text. Each prediction consists of text span, label, and confidence score. To balance precision and recall, we use a confidence threshold that maximises the F-1 score on the validation set, i.e., for threshold $t \in [0, 1)$ we only retain entities predicted with confidence score $c \geq t$.

LLM-based de-identification Here, we adapt the approach of Liu et al. (2023) and prompt an LLM in a zero-shot setting to redact predefined PII categories from Table 1 (e.g., `<NAAM>`, `<ZIEKENHUIS>`). When applying LLMs directly as a de-identification pipeline (P_{LLM}), we experiment with several LLMs that vary in architecture, domain specialization, and size, including GPT-4o (Achiam et al., 2023), DeepSeekR1 (8B and 70B) (Guo et al., 2025), LLaMA-3.1 8B (Dubey et al., 2024), and MedGEMMA 27B (Sellergren et al., 2025). When using LLMs combined with DP ($\text{P}_{\text{LLM}\rightarrow\text{DP}}$), we use the best-performing open-source LLM in terms of privacy.

Metric-DP Metric-DP introduces geometric noise in the embedding space to privatize tokens (Chatzikokolakis et al., 2013). We use embeddings from BERTje[3] (De Vries et al., 2019) and construct an approximate nearest-neighbor (ANN) index over the vocabulary.[4] For each token, we sample a noise vector from a γ distribution with scale $1/\epsilon$, add it to the token's embedding, and select the closest replacement from the ANN index.

RANTEXT RANTEXT (Tong et al., 2025) employs an LLM to generate embedding vectors for tokens (we use the Dutch GPT-2[5] from de Vries and Nissim 2020). For each token, a Laplace distribution is used to dynamically determine the size of a random adjacency list, and then new tokens are sampled from this list to replace the original token. Candidate replacements within the noise radius are

[2] https://huggingface.co/urchade/gliner_multi-v2.1

[3] https://huggingface.co/GroNLP/bert-base-dutch-cased

[4] An ANN index is a data structure designed to quickly retrieve the nearest vectors to a query vector without computing all pairwise distances, which would be prohibitively expensive in high-dimensional spaces.

[5] https://huggingface.co/GroNLP/gpt2-small-dutch

selected via the exponential mechanism, favoring tokens closer to the original in embedding space.

2.3 De-identification pipelines

We investigate five different *pipelines*, i.e., automatic methods to de-identify the text in the raw dataset ($\mathcal{D}_{\text{raw}}$). Our dataset, as detailed in §2.1, is extremely small; thus, all pipelines we propose next require models without fine-tuning. The pipelines are: $\mathbf{P}_{\texttt{NER}}$: Apply NER de-identification to $\mathcal{D}_{\text{raw}}$. $\mathbf{P}_{\texttt{LLM}}$: Apply LLM de-identification to $\mathcal{D}_{\text{raw}}$. $\mathbf{P}_{\texttt{DP}}$: Apply DP-based de-identification to $\mathcal{D}_{\text{raw}}$. $\mathbf{P}_{\texttt{NER}\rightarrow\texttt{DP}}$: Apply a NER model to $\mathcal{D}_{\text{raw}}$ and apply DP-based de-identification to its output. $\mathbf{P}_{\texttt{LLM}\rightarrow\texttt{DP}}$: Apply LLM de-identification to $\mathcal{D}_{\text{raw}}$ and apply DP-based de-identification to its output.

2.4 Privacy Leakage Evaluation

We quantify the privacy leakage of a de-identification pipeline as *the percentage of sensitive information leaked in the de-identified text compared to* $\mathcal{D}_{manual}$. We apply a strict evaluation setting where the leakage of any subword token part of a sensitive entity span is treated as leakage.

2.5 Utility Evaluation

We assess the preservation of relevant clinical information by using the generated notes in two downstream tasks: classifying entity spans in a clinical text as drug or disorder mentions, and predicting whether a drug–disorder entity pair denotes an adverse drug event (ADE).[6] We use the original train/dev/test splits in the Dutch ADE corpus for training, model selection, and testing, respectively. We use the annotated entities and relations that remain in the generated texts for training and development, and evaluate the model against the original test set with gold annotations. We omit some experiments at low ϵ values where we only have less than 10 training data points, and report performance using macro F-1 score for both entity and relation classification.

3 Results and Discussion

In Figure 2, we show the empirical privacy leakage of PII in the generated clinical notes (plot a), drug and disorder (entity) classification F-1 scores (plot b), and adverse drug event (relation) classification F-1 scores (plot c) achieved by different de-identification pipelines.

3.1 Privacy Leakage Results

DP-based Pipelines Figure 2 (subplot a) presents privacy leakage percentage for various pipeline configurations across DP budgets (ϵ). As ϵ increases, privacy leakage also rises, demonstrating the trade-off between reduced noise (higher ϵ) and weaker privacy preservation. At high privacy budgets, the protection provided by DP diminishes, leading to increased privacy leakage across all pipelines. Among the DP-based pipelines, $\text{P}_{\texttt{DP}}$ consistently exhibits the highest privacy leakage across all ϵ values. In contrast, pipelines that combine DP with NER or LLM-based de-identification show significantly lower leakage. Notably, pipelines with LLM preprocessing significantly outperform other approaches with a high budget. This improvement can be attributed to the strong ability of LLMs to identify and redact sensitive information more effectively, reducing the amount of private data exposed to DP noise.

Non-DP Baselines Amongst non-DP baselines, NER-based ones exhibit the worst performance. In contrast, LLM pipelines, with models such as GPT-4o obtain noticeably lower privacy leakage. This aligns with the assumption that LLMs' advanced linguistic capabilities allow them to redact sensitive information more effectively than NER systems.

3.2 Utility Preservation Results - Entity Classification

Figure (b) illustrates the relationship between privacy budgets (ϵ) and utility (F1 scores) for entity classification across different pipelines. At very low privacy budgets ($\epsilon \leq 32$), all DP pipelines demonstrate poor utility due to the high levels of noise, but utility improves as ϵ increases, peaking around $\epsilon = 512$. $\text{P}_{\texttt{DP}}$ consistently exhibits lower F1 scores across all privacy budgets, demonstrating utility degradation caused by noise addition, particularly at low ϵ values. Introducing linguistic preprocessing improves utility, with $\text{P}_{\texttt{NER}\rightarrow\texttt{DP}}$ achieving higher F1 scores than $\text{P}_{\texttt{DP}}$, especially at mid-to-high privacy budgets ($\epsilon \geq 128$). $\text{P}_{\texttt{LLM}\rightarrow\texttt{DP}}$ further outperforms $\text{P}_{\texttt{NER}\rightarrow\texttt{DP}}$ in most cases, showcasing the advantage of leveraging advanced language models for preprocessing. Both DP methods (RANTEXT and metric-DP) exhibit similar trends

[6] We note that originally the Dutch ADE dataset was collected to support these two tasks (Murphy et al., 2025b), making these tasks ideal for utility evaluation.

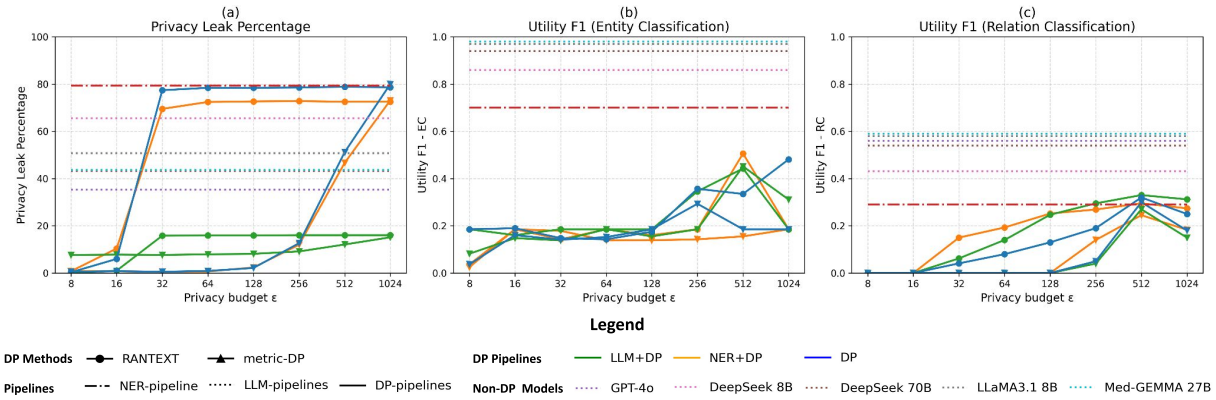

Figure 2: Comparison of privacy leakage (a) and F1 scores for EC (b) and RC (c) tasks across different de-identification pipelines and DP budgets (ϵ). The figure includes two DP mechanisms: RANTEXT and Metric-DP, each applied to three pipelines: P_{DP}, $P_{NER \rightarrow DP}$, and $P_{LLM \rightarrow DP}$. Horizontal lines indicate non-DP baselines, including one NER-based pipeline (GLiNER) and five LLMs. Markers identify the specific DP mechanism used. As ϵ increases, privacy leakage of applying DP directly to raw text also significantly increases. Meanwhile, combining DP with NER or LLM-based redaction largely enhances privacy even with high ϵ values. Conversely, utility scores improve with higher ϵ, but still cannot recover baseline performances.

within each pipeline, suggesting that preprocessing, rather than the DP mechanism itself, plays a more critical role in improving utility. Non-DP baselines maintain consistently high F1 scores near 1.0, as they do not introduce noise, thereby serving as an upper bound for utility.

3.3 Utility Preservation Results - Relation Classification

Figure (c) presents the utility (F1 scores) for relation classification also across various privacy budgets (ϵ) and pipelines. At low privacy budgets, such as $\epsilon = 8$ or in some cases 16, no results are obtained due to the severe obfuscation which reduces the amount of usable training data to negligible levels. As ϵ increases, utility scores gradually improve across all pipelines. P_{DP} exhibit the lowest F1 scores throughout the range of privacy budgets. In contrast, both $P_{NER \rightarrow DP}$ and $P_{LLM \rightarrow DP}$ achieve higher F1 scores, with $P_{LLM \rightarrow DP}$ consistently outperforming $P_{NER \rightarrow DP}$, particularly at mid-to-high privacy budgets ($\epsilon \geq 128$). Non-DP baselines still maintain consistently high F1 scores and establish the upper bounds for utility. These results underscore the crucial role of linguistic preprocessing, particularly with LLMs, in mitigating the utility degradation inherent to DP-based methods, especially in tasks requiring complex relational reasoning.

3.4 Discussion

Building the manually de-identified corpus ($\mathcal{D}_{\text{manual}}$) was a labor-intensive process, requiring over 47 hours of annotation by two annotators (Murphy et al., 2025a). In contrast, our automated pipelines are way more time- and cost-efficient, with metric-DP taking 3 minutes and 16 seconds per run and InferDPT 26 minutes and 40 seconds per run. Our study also addresses a critical gap in the research on non-English data by focusing on Dutch text, as most prior research assumes English by default. This might partially explain the relatively lower performance of our results, as the models and pipelines used may not be as well-optimized for Dutch compared to English. The limited availability of high-quality Dutch-language datasets and pretrained models likely impacts the ability of these methods to effectively handle linguistic nuances, resulting in suboptimal outcomes.

4 Conclusions and Future work

In this paper, we presented the first comparative study of DP for Dutch clinical text de-identification, contrasting DP approaches with state-of-the-art NER and LLM-based redaction methods. Our results show that while DP mechanisms offer formal guarantees, they suffer utility degradation at low privacy budgets. Combining linguistic preprocessing with DP improves the privacy–utility trade-off, with LLM-based preprocessing performing best. We also highlight the cost-efficiency of automated DP pipelines. Future work should explore advanced hybrid strategies, scalable DP training for LLMs, and broader evaluations to enhance the practicality of privacy-preserving clinical NLP.

Limitations

While our study provides valuable insights into the application of differential privacy (DP) for clinical text de-identification, it has certain limitations. First, our work focuses exclusively on Dutch clinical narratives, which may limit the generalizability of our findings to other languages or multilingual clinical datasets. Expanding this research to additional languages would provide a broader understanding of the effectiveness and scalability of the proposed techniques. Second, the dataset used in this study is private, as it contains sensitive medical information, which restricts reproducibility and external validation by other researchers. Although this is a common constraint in clinical research, the use of publicly available or synthetic benchmark datasets in future studies could help foster more extensive experimentation and comparison. Lastly, while we evaluated our pipelines on two downstream tasks (entity and relation classification), further exploration of their performance across a wider variety of use cases, such as document classification or clinical outcome prediction, would provide a more comprehensive assessment of their utility.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Loick Bourdois, Marta Avalos, Gabrielle Chenais, Frantz Thiessard, Philippe Revel, Cedric Gil-Jardine, and Emmanuel Lagarde. 2021. De-identification of emergency medical records in french: Survey and comparison of state-of-the-art automated systems. *The International FLAIRS Conference Proceedings*, 34.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Konstantinos Chatzikokolakis, Miguel E. Andrés, Nicolás Emilio Bordenabe, and Catuscia Palamidessi. 2013. Broadening the scope of differential privacy using metrics. In *International Symposium on Privacy Enhancing Technologies*.

Andrew Kweku Conduah, Sebastian Ofoe, and Dorothy Siaw-Marfo. 2025. Data privacy in healthcare: Global challenges and solutions. *Digital Health*, 11:20552076251343959.

Wietse de Vries and Malvina Nissim. 2020. As good as new. how to successfully recycle english gpt-2 to make models for other languages. *Preprint*, arXiv:2012.05628.

Wietse De Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. Bertje: A dutch bert model. *arXiv preprint arXiv:1912.09582*.

Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. 2017. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association (JAMIA)*, 24(3):596–606.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Cynthia Dwork. 2006. Differential privacy. In *Proceedings of the 33rd International Conference on Automata, Languages and Programming - Volume Part II*, ICALP'06, page 1–12, Berlin, Heidelberg. Springer-Verlag.

Cynthia Dwork and Aaron Roth. 2014. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407.

European Parliament and Council of the European Union. 2016. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (general data protection regulation). *Official Journal of the European Union*, L119:1–88.

Oluwaseyi Feyisetan, Tom Diethe, and Thomas Drake. 2019. Leveraging hierarchical representations for preserving privacy and utility in text. *2019 IEEE International Conference on Data Mining (ICDM)*, pages 210–219.

Cyril Grouin, Nicolas Griffon, and Aurélie Névéol. 2015. Is it possible to recover personal health information from an automatically de-identified corpus of French EHRs? In *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis*, pages 31–39, Lisbon, Portugal. Association for Computational Linguistics.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Zhengliang Liu, Yue Huang, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Yiwei Li, Peng Shu, Fang Zeng, Lichao Sun, Wei Liu, Dinggang Shen, Quanzheng Li, Tianming Liu, Dajiang Zhu, and Xiang Li. 2023. Deid-gpt: Zero-shot medical text de-identification by gpt-4. *Preprint*, arXiv:2303.11032.

Rachel M Murphy, Dave A Dongelmans, Nicolette F de Keizer, Rosa J Jongeneel, Christiaan H Koster, Kitty J Jager, Ameen Abu-Hanna, Iacer Calixto, and Joanna E Klopotowska. 2025a. Creation of a gold standard dutch corpus of clinical notes for adverse drug event detection: the dutch ade corpus. *Language Resources and Evaluation*, pages 1–17.

Rachel M Murphy, Nishant Mishra, Nicolette F de Keizer, Dave A Dongelmans, Kitty J Jager, Ameen Abu-Hanna, Joanna E Klopotowska, and Iacer Calixto. 2025b. Detection of adverse drug events in dutch clinical free text documents using transformer models: benchmark study. *arXiv preprint arXiv:2507.19396*.

David Pissarra, Isabel Curioso, João Alveira, Duarte Pereira, Bruno Ribeiro, Tomás Souper, Vasco Gomes, André Carreiro, and Vitor Rolla. 2024. Unlocking the potential of large language models for clinical text anonymization: A comparative study. In *Proceedings of the Fifth Workshop on Privacy in Natural Language Processing*, pages 74–84, Bangkok, Thailand. Association for Computational Linguistics.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

Yakini Tchouka, Jean-François Couchot, Maxime Coulmeau, David Laiymani, Philippe Selles, and Azzedine Rahmani. 2022a. De-identification of french unstructured clinical notes for machine learning tasks. *arXiv preprint arXiv:2209.09631*.

Yakini Tchouka, Jean-François Couchot, and David Laiymani. 2022b. An easy-to-use and robust approach for the differentially private de-identification of clinical textual documents. *arXiv preprint arXiv:2211.01147*.

Yakini Tchouka, Jean-François Couchot, Maxime Coulmeau, David Laiymani, Philippe Selles, and Azzedine Rahmani. 2022c. De-identification of french unstructured clinical notes for machine learning tasks. *arXiv preprint arXiv:2209.09631*.

Yakini Tchouka, Jean-François Couchot, and David Laiymani. 2022d. An easy-to-use and robust approach for the differentially private de-identification of clinical textual documents. *arXiv preprint arXiv:2211.01147*.

Meng Tong, Kejiang Chen, Jie Zhang, Yuang Qi, Weiming Zhang, Nenghai Yu, Tianwei Zhang, and Zhikun Zhang. 2025. Inferdpt: Privacy-preserving inference for black-box large language model. *Preprint*, arXiv:2310.12214.

U.S. Department of Health and Human Services. 45 cfr § 164.514 – de-identification of health information. Health Information Privacy. [Online]. Available: https://www.law.cornell.edu/cfr/text/45/164.514. [Accessed: Dec. 2, 2024].

Peng Wang, Yong Li, Liang Yang, Simin Li, Linfeng Li, Zehan Zhao, Shaopei Long, Fei Wang, Hongqian Wang, Ying Li, and Chengliang Wang. 2022. An efficient method for deidentifying protected health information in chinese electronic health records: Algorithm development and validation. *JMIR Med Inform*, 10(8):e38154.

Tianyu Yang, Xiaodan Zhu, and Iryna Gurevych. 2025. Robust utility-preserving text anonymization based on large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28922–28941, Vienna, Austria. Association for Computational Linguistics.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. 2022. Differentially private fine-tuning of language models. In *International Conference on Learning Representations*.

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.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. GLiNER: Generalist model for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376, Mexico City, Mexico. Association for Computational Linguistics.

A Related Work

Anonymization of clinical text is essential to allow secondary use of healthcare data while protecting patient privacy under privacy regulations such as HIPAA (U.S. Department of Health and Human Services) and GDPR (European Parliament and Council of the European Union, 2016). Traditional manual de-identification methods are labour-intensive, costly, and prone to inconsistencies. This

section synthesizes key methodological advancements in automated de-identification.

Differential Privacy Approaches Differential Privacy (Dwork, 2006) offers formal privacy guarantees by ensuring that the outputs of algorithms are statistically indistinguishable across neighboring datasets. Metric DP extends the traditional DP framework by introducing a metric space that quantifies the similarity between data points (Chatzikokolakis et al., 2013). In the context of text data, Metric DP leverages semantic similarity measures to guide the privacy mechanism. (Feyisetan et al., 2019) applied Metric DP to word embeddings, enabling controlled perturbations that account for the semantic relationships between words. This results in text transformations that better preserve the utility of the original data while providing privacy guarantees.

Another promising approach is RANTEXT (Tong et al., 2025), a token-level DP framework designed for privacy-preserving text generation. RANTEXT dynamically constructs randomized adjacency lists for context-aware token substitution and has demonstrated strong resistance to membership inference attacks, outperforming earlier models like SANTEXT+ and CUSTEXT+. In empirical evaluations, it achieved up to 98% F1 scores for semantic preservation while maintaining privacy guarantees with ϵ values below 2. Nonetheless, DP-based methods often struggle with high-dimensional clinical narratives, where stricter privacy constraints can significantly degrade data utility (Tong et al., 2025).

LLM-Based De-identification Large language models have taken every field by storm since the release of GPT-3 (Brown et al., 2020). DeID-GPT (Liu et al., 2023), built on GPT-4, asks the model directly to identify and redact sensitive information and integrates HIPAA identifier categories directly into its prompts, achieving 99.25% precision and 89.73% F1 in zero-shot redaction tasks. For an approach this simple, the performance is extremely good. It should be noted that in a realistic setting, it is unlikely that anyone would want to use GPT-4 for this task, and it is not clear whether this capability is going to be observed also for other models. Using an off-site model would mean sending sensitive data off-site, which is often something not allowed with sensitive data, unless the entity owning the data has some agreement in place that keeps the data safe.

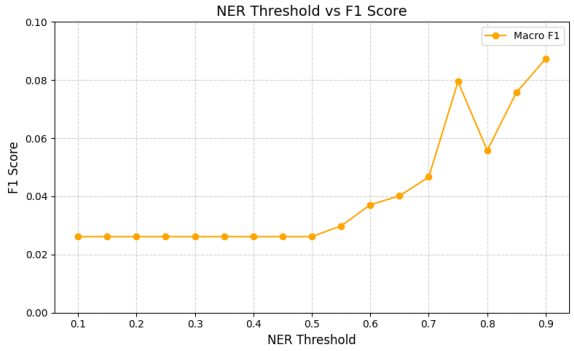

Figure 3: NER Threshold vs F1 Score. The highest score is achieved at 0.90 due to low false positives, and a relatively high score at 0.75 because of high true positive rate.

B Appendix

B.1 PII

There are 17 types of PII annotated in $\mathcal{D}_{\text{manual}}$. Names of the placeholders and their meaning are listed in Table 1.

B.2 NER Threshold Selection

Figure 3 shows the F1 scores of NER under different confidence thresholds. A confidence threshold of 0.90 yields the best overall performance for GLiNER, achieving an F1-score of 0.87. Table 2 provides the results for various thresholds applied to the NER process. The first few thresholds (0.10 to 0.45) yield identical results, with a macro F1-score of 0.026149, due to the unchanged true positives (TP = 294), false positives (FP = 19961), and false negatives (FN = 586). This consistency reflects that the model does not recognise any entities below the confidence of 0.45. Notably, at a threshold of 0.75, the F1-score jumps to 0.079465, driven primarily by a significant increase in true positives (TP = 377), highlighting the model's improved ability to correctly identify relevant entities. The highest F1-score of 0.087330 is observed at a threshold of 0.90, primarily due to a substantial reduction in false positives (FP = 2062), which improves the model's precision. Given the balance between precision and recall at this threshold, we selected 0.90 as the final threshold for all subsequent experiments involving the $\mathrm{P}_{\mathrm{NER}\rightarrow\mathrm{DP}}$ pipeline, as it optimizes the trade-off between minimizing false positives and maintaining sufficient true positives.

Similarly, Table 3 provides a comparison of macro F1-scores, true positives (TP), false positives

Name	Meaning
<AFDELING>	Department
<APOTHEEK>	Pharmacy
<ARTS>	Doctor
<EHR>	Electronic Health Record ID
<FEESTDAG>	Holiday
<GEBOORTEDATUM>	Date of Birth
<NAAM>	Name
<RARE_DISEASE>	Rare Disease
<RARE_DISEASE_TREATMENT>	Rare Disease Treatment
<REVALIDATIECENTRUM>	Rehabilitation Center
<SEIN>	Signal
<STAD>	City
<TELNR>	Telephone Number
<TRIAL-ID>	ID of Clinical Trial
<ZIEKENBOEG>	Sickbay
<ZIEKENHUIS>	Hospital
<ZKH>	Abbreviation for Hospital

Table 1: List of placeholders and their meanings.

(FP), and false negatives (FN) across various LLM models applied to the redaction task. DeepSeek-8b and DeepSeek-70b exhibit relatively low F1-scores of 0.337265 and 0.337807, respectively, with DeepSeek-70b achieving slightly higher true positives (TP = 499) but at the cost of significantly more false positives (FP = 3614). LLaMA3-8b demonstrates a notable improvement with a good balance between true positives and false positives. MedGemma emerges as the best-performing model, with the highest macro F1-score of 0.591672. This is supported by its significant reduction in false positives and the highest true positives.

Threshold	F1-score	TP	FP	FN
0.10	0.026149	294	19961	586
0.15	0.026149	294	19961	586
0.20	0.026149	294	19961	586
0.25	0.026149	294	19961	586
0.30	0.026149	294	19961	586
0.35	0.026149	294	19961	586
0.40	0.026149	294	19961	586
0.45	0.026149	294	19961	586
0.50	0.026150	294	19959	586
0.55	0.029845	303	17828	577
0.60	0.037103	320	14989	560
0.65	0.040175	306	12862	574
0.70	0.046647	298	10368	582
0.75	0.079465	377	7951	503
0.80	0.055811	226	5845	654
0.85	0.075774	219	3520	661
0.90	0.087330	177	2062	703

Table 2: Thresholds for the replacement of text by NER type names when applying NER to $\mathcal{D}_{\text{raw}}$, F1-scores of each threshold, True Positives (TP), False Positives (FP), and False Negatives (FN).

Model	F1-score	TP	FP	FN
GPT-4o	0.205839	569	4848	310
DeepSeek-8b	0.337265	303	718	576
DeepSeek-70b	0.337807	499	3614	380
LLaMA3-8b	0.492872	433	652	446
MedGemma	0.591672	495	399	384

Table 3: Comparison of LLMs by F1-scores, True Positives (TP), False Positives (FP), and False Negatives (FN).