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

Michele Miranda[12*], **Xinlan Yan**[34*], **Nishant Mishra**[34], **Rachel Murphy**[34], **Ameen Abu-Hanna**[34], **Sébastien Bratières**[2], **Iacer Calixto**[34]
[1]Sapienza University of Rome, [2]Translated, [3]Amsterdam UMC, [4]University of Amsterdam
miranda@di.uniroma1.it, sebastien@translated.com
{x.yan, n.mishra, r.m.murphy, a.abu-hanna, i.coimbra}@amsterdamumc.nl

Abstract

Protecting patient privacy in clinical narratives is essential for enabling secondary use of healthcare data under regulations such as GDPR and HIPAA. While manual de-identification remains the gold standard, it is costly and slow, motivating the need for automated methods that combine privacy guarantees with high utility. Historically, most automated 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). We show that DP mechanisms alone degrade utility substantially, but combining them with linguistic preprocessing, especially LLM-based redaction, significantly improves the privacy–utility trade-off.

Keywords: Clinical Notes, De-Identification, Differential Privacy, NER

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). More recently, large language models (LLMs) have been applied to PII de-identification and have shown strong performance (Liu et al., 2023). NER- and LLM-based methods, despite their performance, do not provide any formal privacy guarantees. Differential privacy (DP; Dwork, 2006; Dwork and Roth, 2014), on the other hand, 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 severely 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 LLM-based methods remain limited as they redact only detected entities rather than full text—offering weaker privacy preservation than DP—, overlook instruction-tuned LLMs, demand computational resources often unavailable in privacy-critical settings, and are developed solely for English and are thus untested in other languages.

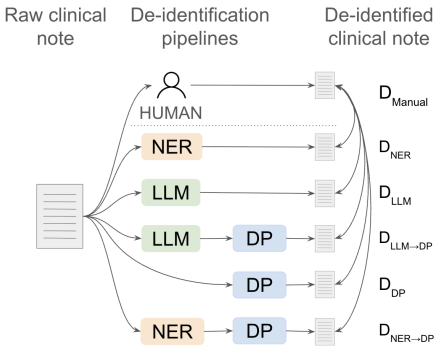

Figure 1: Overview of our comparative analysis. A raw document $\mathcal{D}_{\text{raw}}$ is de-identified using 5 different pipelines, which are evaluated against a manually de-identified version of the same document $\mathcal{D}_{\text{manual}}$. We use a range of open-source and proprietary LLMs that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

In this study, we compare different PII de-

*These authors contributed equally to this work.

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 113–124
16 May 2026.

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 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 Dutch real-world hospital 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 after de-identification, and *extrinsically* by measuring the performance of prediction models trained using the generated de-identified notes for 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. Related Works

De-identification 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 PII 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. In other words, the key idea is that two models ($\mathcal{M}_1$, $\mathcal{M}_2$) trained on two datasets ($\mathcal{D}_1$, $\mathcal{D}_2$) that differ only by one entry, i.e., one patient record, should be indistinguishable in terms of their predicted outcomes.

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 can be applied to text PII de-identification in different ways. DeID-GPT (Liu et al., 2023), built on GPT-4, asks the model to directly 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. This can be considered very good performance, especially given the simplicity of the approach.

We note that it is unlikely that organisations would want to use GPT-4 for this task, since using an online model would mean sending sensitive data off-site. Often, this may be even forbidden under privacy regulations. In this work, we use the OpenAI API via a cloud-based service provider in a way that preserves privacy in agreement with privacy regulations. We understand this is not always possible for every hospital, and for that reason we include experiments with both proprietary and open-weight LLMs.

3. Methods and Experiments

Below we detail the dataset (§3.1), the de-identification modules (§3.2) used in our de-identification pipelines that generate privacy-preserving clinical notes (§3.3; Fig. 1), and the intrinsic (§3.4) and extrinsic evaluation on the entity and relation classification tasks (§3.5).

3.1. Dataset

We use the Dutch ADE dataset (Murphy et al., 2025a), a benchmark corpus containing 102 clinical notes from intensive care unit (ICU) patients annotated with entity-level labels—*drug* and *disorder* mentions—and relation labels between *drug-disorder* pairs. The two relation labels annotated are *adverse drug event* or *ADE* (when the drug

Placeholder	Meaning
<AFDELING>	Department
<APOTHEEK>	Pharmacy
<ARTS>	Doctor
<EHR>	Electronic Health Record System
<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.

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 PII. The PII has been redacted and replaced by 17 different placeholders, each indicating one type of PII in the data. There are thus 17 types of PII annotated in $\mathcal{D}_{\text{manual}}$. Names of the placeholders and their meaning are listed in Table 1. For further details on the corpus, we refer the reader to Murphy et al. (2025a).

3.2. 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 off-the-shelf pretrained multilingual NER model `gliner-multi-v2.1`[1], without any fine-tuning on our clinical data. We provide it with our target entity types (see Table 1, e.g., `<NAAM>`, `<ZIEKENHUIS>`) 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 a 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). Please see Appendix A.1 for the prompt we use. 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 (Deepseek-70B).

For our experiments with GPT models, we use the OpenAI API in a privacy-compliant setting whereby no private data leaves the premises of our hospital. Moreover, all our experiments are conducted in a privacy-compliant setting and strictly follow the data usage agreement of the Dutch ADE dataset.

Metric-DP Metric-DP introduces geometric noise in the embedding space to privatize tokens (Chatzikokolakis et al., 2013). We use embeddings from BERTje[2] (De Vries et al., 2019) and construct an approximate nearest-neighbor (ANN) index over the vocabulary.[3] 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[4] 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 selected via the exponential mechanism, favoring tokens closer to the original in embedding space.

3.3. 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 §3.1, is very 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-

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

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

[3] 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.

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

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.

In all experiments with pipeline $\mathbf{P}_{\texttt{LLM}\rightarrow\texttt{DP}}$, we use Deepseek-70B as it is the best-performing open-source LLM in terms of privacy.

3.4. Privacy Leakage Evaluation

We quantify the privacy leakage of a de-identification pipeline as *the percentage of personally identifiable information (PII) leaked in the de-identified text compared to* $\mathcal{D}_{manual}$. PIIs are not all the same, and the leakage of different types of PII can, in practice, have very different impact in terms of privacy preservation. For that reason, in this work we differentiate between leakage of *direct* and *indirect PII*.

Direct vs. Indirect PII Direct PII (e.g., a patient's name or telephone number) can directly identify an individual. For that reason, their leakage is the most damaging in terms of privacy preservation. Indirect PII have lower risk but can also be used to identify an individual, usually in combination with other indirect PII or with extra auxiliary information, e.g., Sweeney (2000)'s classical example whereby zip code, date of birth and gender is reported to uniquely identify 87% of the population in the USA.

According to U.S. Department of Health and Human Services, the following identifier types are classified as direct PII: name; address (all geographic subdivisions smaller than state, including street address, city county, and zip code); all date-related elements (except years) related to an individual (including birth date, admission date, discharge date, date of death, and exact age if over 89); telephone number; fax number; email address; social security number; medical record number; health plan beneficiary number; account number; certificate or license number; vehicle identifiers and serial numbers, including license plate numbers; device identifiers and serial numbers; web URL; internet protocol (IP) address; finger or voice print; photographic image; and any other characteristic that could uniquely identify the individual.

In our data, we consider name of doctor, electronic health record ID, date of birth, name of patient, city, and telephone number as direct PIIs, and all the rest in Table 1 as indirect PIIs.

In all our experiments, we apply a strict evaluation setting where the leakage of any subword token part of a PII entity span is treated as leakage.

3.5. 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).[5] We use the original train/dev/test splits in the Dutch ADE corpus for training, model selection, and testing, respectively. For configuration details please see Appendix A.2 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. At low ϵ values under some settings, i.e. $\epsilon\leq$128 for metric-DP and $\epsilon\leq$16 for RANTEXT, we only have less than 10 annotated relations in the generated text, making it impossible to train the relation classification model. Thus, we omit these experiments and report performance using macro F-1 score for both entity and relation classification. We do not report new confidence intervals for the downstream tasks because its protocol and uncertainty were established in prior work (Murphy et al., 2025b), which already provides CIs under the same evaluation procedure.

4. Results

We now report on our experiments regarding the empirical privacy leakage of PII in the generated clinical notes (4.1), and the downstream evaluations according to drug and disorder (entity) classification (4.2) and adverse drug event (relation) classification (4.3).

4.1. Privacy Leakage Results

Figure 2 presents privacy leakage percentage for various pipeline configurations across DP budgets (ϵ), spanning from 8 to 1024. We acknowledge that with ϵ higher than 10, the theoretical guarantees of DP fade away, but it is also usual in practical setting to work with high epsilons in order to get meaningful empirical results, as seen in works like DP-BART (Igamberdiev and Habernal, 2023). 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, $\mathrm{P}_{\texttt{DP}}$ consistently exhibits the highest privacy leakage across all ϵ values. In contrast, pipelines that combine DP with NER preserve privacy for slightly higher privacy budgets, and LLM-based de-identification show significantly lower empirical leakage. Notably,

[5] 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.

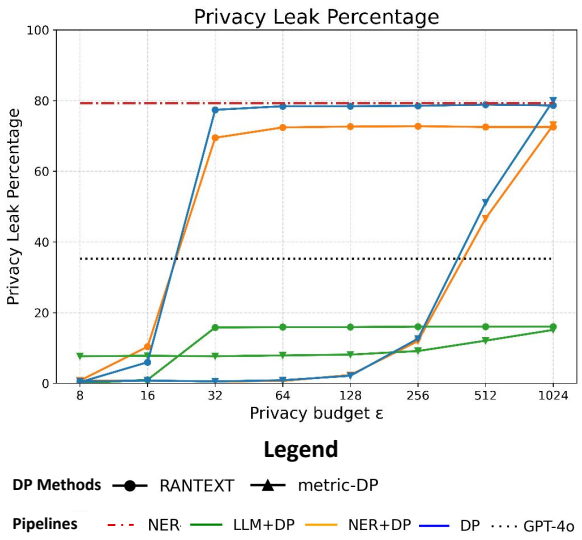

Figure 2: Comparison of privacy leakage across different de-identification pipelines and DP budgets (ϵ). This figure includes two DP mechanisms: RANTEXT and Metric-DP, each applied to three pipelines: P_{DP}, $P_{NER \to DP}$, and $P_{LLM \to DP}$. For $P_{LLM \to DP}$, we use Deepseek-70B as the de-identification module as it performs the best in terms of privacy. Horizontal lines indicate non-DP baselines, including one NER-based pipeline (GLiNER) and one best performing LLM (GPT-4o). As ϵ increases, privacy leakage of applying DP directly to raw text also significantly increases. Meanwhile, combining DP with LLM-based redaction largely enhances privacy even with high ϵ values.

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.

In our next experiments, we use the Health Insurance Portability and Accountability Act (HIPAA; Garfinkel et al., 2015) to differentiate between leakage of *direct* (i.e., exact identifiers such as names, emails, phone numbers) versus *indirect PII* (i.e., *quasi-identifiers* and other attributes that enable re-identification when linked). Figure 3 illustrates the leakage rates of direct and indirect PIIs for each privacy budget ϵ across pipelines. Overall, leakage increases with ϵ, but the extent and composition differ largely according to the pipeline. Metric-DP consistently leaks far less PII than RANTEXT at comparable budgets: while RANTEXT and NER+RANTEXT reach 38–40% direct (and 76–71% total) leakage at high ϵ, metric-DP (MDP) better preserves privacy more at low-to–mid ϵ, and only approaches RANTEX when ϵ is large. Moreover, leakage rises with ϵ for all methods, but the slope differs. RANTEXT variants escalate quickly, with the direct component dominating the increase, whereas MDP grows more gradually until the largest budgets, where its bars steepen.

Moreover, the baseline results of LLMs and NER confirms that stand-alone methods without DP are markedly leakier: the pure NER system shows high overall leakage with a large direct component, and LLM-only variants also leak substantially. However, inserting an LLM de-identification before the DP step shows a clear advantage for both families. Finally, pipelines combining LLMs with DP (LLM+MDP and LLM+RANTEXT) were the only pipelines we investigated that managed to keep empirical privacy leakage below 10%, the majority of which consisted of indirect PII.

4.2. Utility Preservation Results - Entity Classification

Figure 4 illustrates the relationship between privacy budgets (ϵ) and entity classification F1 scores (utility) 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 $\epsilon = 512$. P_{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 $P_{NER \to DP}$ achieving higher F1 scores than P_{DP}, especially at mid-to-high privacy budgets ($\epsilon \geq 128$). $P_{LLM \to DP}$ further outperforms $P_{NER \to DP}$ in most cases, showcasing the advantage of leveraging advanced language models for preprocessing. Both DP methods (RANTEXT and metric-DP) exhibit similar trends within each pipeline, suggesting that preprocessing, rather than the DP mechanism itself, plays a more critical role in improving utility. The LLM baseline maintains consistently high F1 scores near 1.0, thereby serving as an upper bound for utility.

4.3. Utility Preservation Results - Relation Classification

Figure 5 shows relation classification F1 scores (utility) across various privacy budgets (ϵ) and for different 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 $P_{NER \to DP}$ and $P_{LLM \to DP}$ achieve higher F1 scores, with $P_{LLM \to DP}$ consistently outperforming $P_{NER \to DP}$, particularly at mid-to-high privacy budgets ($\epsilon \geq 128$). Non-DP baselines still maintain

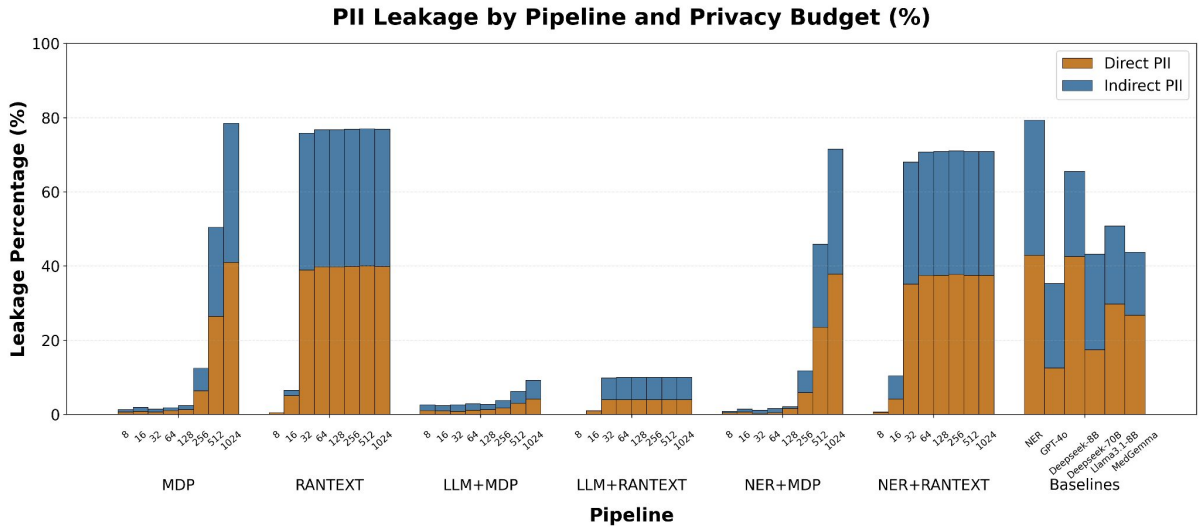

Figure 3: PII leakage by pipeline and privacy budget ϵ. Bars are stacked (dark: direct PII; light: indirect PII). Across budgets, metric-DP (MDP) leaks substantially less than RANTEXT, and all methods show increasing leakage as ϵ grows, with the steepest rise for RANTEXT variants. Prepending an LLM rewrite reduces leakage for both families, with LLM+MDP achieving the lowest overall rates, followed by LLM+RANTEXT, while plain RANTEXT and NER+RANTEXT exhibit high direct leakage. NER and LLMs alone without DP also remain markedly leakier.

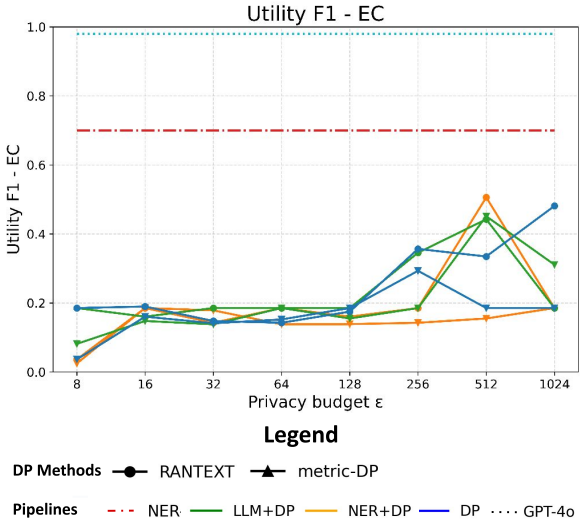

Figure 4: Comparison of utility F1-score for Entity Classification (EC) task across different de-identification pipelines and DP budgets (ϵ) (see Figure 2 for more details). It shows that the utility scores improve with higher ϵ, but still cannot recover baseline performances.

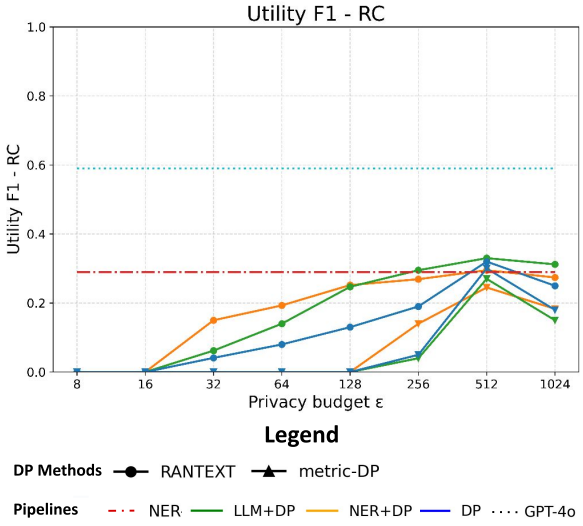

Figure 5: Comparison of utility F1-score for Relation Classification (RC) task across different de-identification pipelines and DP budgets (ϵ) (see Figure 2 for more details). It shows that the utility scores slightly improve with higher ϵ, but are significantly harmed even with high ϵ, and cannot recover baseline performances.

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.

4.4. LLM Baseline Results

Figure 6 shows the performances of privacy leakage and utility F1-scores for both EC and RC tasks among P_{LLM} using 5 various LLMs. Amongst the LLM baselines, GPT-4o achieves the best, i.e. lowest privacy leakage percentage while still offering

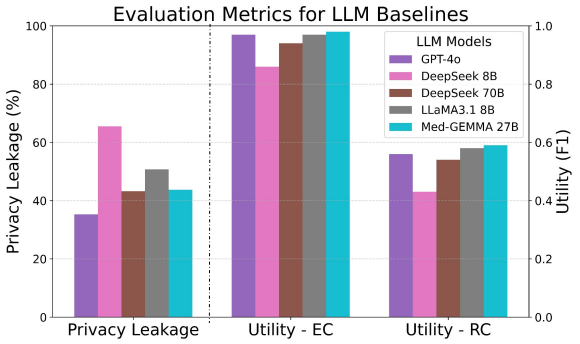

Figure 6: Comparison of evaluation metrics including Privacy Leakage, Utility - EC, and Utility - RC for P_{LLM} using 5 different LLMs. Privacy Leakage, where lower is better, shows that GPT-4o achieves the lowest leakage. Utility metrics for EC and RC, where higher is better, reveal that Med-GEMMA 27B consistently outperforms others while obtaining relatively low privacy leakage. GPT-4o achieves similar utility even with the lowest privacy leakage, indicating its high overall performance.

near-optimal utility for both EC and RC tasks. Med-GEMMA 27B exhibits the highest utility for both tasks while keeping the privacy leakage at a relatively low level. This aligns with the assumption that LLMs' advanced linguistic capabilities allow them to effectively redact sensitive information.

4.5. Operational Efficiency

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 RANTEXT 26 minutes and 40 seconds per run. While manual de-identification remains the gold standard for accuracy and consistency, these results highlight a substantial operational advantage for the automated methods, enabling rapid iteration over various configurations and large-scale or repeated de-identification in practice.

5. Discussion and Conclusions

General results. This work provides a comparative analysis of privacy-preserving de-identification for Dutch clinical notes by contrasting text-level differential privacy (DP) mechanisms, NER-based masking, and LLM-based masking, and by evaluating both privacy leakage and downstream utility on entity and relation classification tasks. Our results show a consistent tension for DP perturbation: increasing the privacy budget ϵ improves utility but increases residual leakage, and applying DP directly to raw notes yields the weakest privacy outcomes among DP pipelines. By contrast, preprocessing that removes explicit identifiers before DP (via NER or LLM masking) improves the overall trade-off, indicating that, in our no-training setting, most gains come from reducing sensitive content *before* perturbation rather than relying on perturbation alone. In utility terms, token-level perturbation is especially harmful for relation classification (more than for entity classification), plausibly because relations depend on longer-range coherence and on the availability of sufficient high-quality labeled instances: under tighter privacy budgets, the perturbation can effectively corrupt the signal beyond recovery, whereas higher ϵ only partially closes the gap with non-private baselines. Finally, our operational analysis highlights the substantial gap between the cost of producing benchmark-quality manually de-identified reference data and the speed of automated pipelines, which enables rapid iteration over configurations in practice.

Advice for practitioners. When the goal is to share Dutch clinical narratives while explicitly controlling privacy risk, our results suggest prioritizing hybrid strategies: use a strong high-recall masking stage (LLM-based when feasible, otherwise NER-based) and then apply DP as a second-stage safeguard on the residual text, rather than applying DP directly to raw notes. This design reduces the sensitive surface exposed to DP and consistently improves both leakage and downstream utility relative to DP-on-raw-text, while retaining an explicit privacy knob through ϵ. Practitioners should also anticipate that preserving utility for relation-centric tasks will be harder than for entity-centric tasks under text perturbation, and should validate performance on the intended downstream use case (not only on intrinsic leakage). Finally, even when automated pipelines are adopted, a manually de-identified reference (or equivalent audit protocol) remains important for trustworthy evaluation and monitoring. Although the best-performing LLM in our experiments is proprietary (GPT-4o), recent work on Italian clinical notes has shown that smaller, open-source LLMs can achieve competitive de-identification performance in a similar zero-shot setting (Miranda et al., 2025), suggesting that our hybrid pipeline design is not inherently tied to proprietary models.

Future work. Three extensions are particularly important: (i) expanding beyond Dutch to assess whether the same trade-offs hold across languages with different resource availability; (ii) enabling broader experimentation and external validation via public benchmarks or carefully designed synthetic alternatives; and (iii) widening utility evaluation be-

yond entity/relation classification to additional clinical NLP use cases (e.g., document-level prediction), where sensitivity to perturbation may differ. More generally, future work should also strengthen adversarial evaluation of re-identification risk and explore hybrid designs that better target quasi-identifiers without unnecessarily corrupting clinical signals.

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. Third, our privacy leakage metric - residual PII percentage against a manually de-identified reference - is an empirical proxy, not a formal privacy measure. While DP provides theoretical guarantees on output indistinguishability, empirical leakage captures a complementary quantity: how much identifiable information survives in practice. These two notions should not be conflated, yet empirical evaluation against a gold-standard reference remains one of the few viable approaches for assessing de-identification effectiveness on real data. 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.

6. Bibliographical References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 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.

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.

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, et al. 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.

Simson Garfinkel et al. 2015. De-identification of personal information:.

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, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Timour Igamberdiev and Ivan Habernal. 2023. DP-BART for privatized text rewriting under local differential privacy. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13914–13934, Toronto, Canada. Association for Computational Linguistics.

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.

Michele Miranda, Sébastien Bratières, Stefano Patarnello, and Livia Lilli. 2025. Mamma mia! where's my name? de-identifying Italian clinical notes with large language models. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 735–746, Cagliari, Italy. CEUR Workshop Proceedings.

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, et al. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

Latanya Sweeney. 2000. Simple demographics often identify people uniquely. *Health (San Francisco)*, 671(2000):1–34.

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.

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].

Stella Verkijk and Piek Vossen. 2021. Medroberta. nl: a language model for dutch electronic health records. In *Computational Linguistics in the*

Netherlands, volume 11, pages 141–159. Computational Linguistics in the Netherlands.

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. Appendix

A.1. LLM De-identification Prompts and Performances

We use the following prompt for using LLMs as de-identification tools: "*Please anonymize the following clinical note. Replace all the following information with corresponding masks: Replace hospital names with <ZIEKENHUIS>; Replace abbreviations for hospitals (e.g. ZKH) with <ZKH>; Replace doctor names with <ARTS>; Replace city names with <STAD>; Replace signs or signals (SEIN) with <SEIN>; Replace patient names with <NAAM>; Replace entries from electronic health records with <EHR>; Replace rare diseases with <RARE_DISEASE>; Replace clinical trial identifiers with <TRIAL-ID>; Replace holidays with <FEESTDAG>; Replace rare disease treatments with <RARE_DISEASE_TREATMENT>; Replace department names with <AFDELING>; Replace rehabilitation centers with <REVALIDATIECENTRUM>; Replace sickbay references with <ZIEKENBOEG>; Replace telephone numbers with <TELNR>; Replace pharmacies or drug stores with <APOTHEEK>; Replace dates of birth with <GEBOORTEDATUM>.*"

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

Table 2 provides a comparison of macro F1-scores, true positives (TP), false positives (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.

A.2. Downstream Evaluation Configuration

For the downstream Entity Classification (EC) and Relation Classification (RC) tasks, we use MedRoberta.nl (Verkijk and Vossen, 2021) as the base model due to its best performance according to Murphy et al. (2025b) and train with the remaining entity/relation in our generated data. We use a batch size of 8, a learning rate of 3e-5, a warmup ratio of 0.2 for EC and a batch size 128, hidden sizes of 512, 128, 32, a dropout rate of 0.5, a learning

rate 1e-6, and a patience of 30 for RC. We report the micro F1-score of EC and macro F1-score of RC as annotated relations suffer from high imbalance. Table 3 and Table 4 show the full results of the EC and RC evaluation.

Pipeline/Epsilon	8	16	32	64	128	256	512	1024	∞
Baseline	–	–	–	–	–	–	–	–	0.91
NER	–	–	–	–	–	–	–	–	0.70
LLM	–	–	–	–	–	–	–	–	0.97
metricDP	0.0368	0.1605	0.1412	0.1524	0.1854	0.2933	0.1854	0.1854	–
RANTEXT	0.1854	0.1898	0.1479	0.1427	0.1751	0.3567	0.3348	0.4812	–
NER+metricDP	0.0249	0.1855	0.1787	0.1383	0.1388	0.1427	0.1552	0.1854	–
NER+RANTEXT	0.0368	0.1854	0.1427	0.1854	0.1605	0.1854	0.5061	0.1854	–
LLM+metricDP	0.0817	0.1479	0.1383	0.1854	0.1552	0.1854	0.4514	0.3110	–
LLM+RANTEXT	0.1854	0.1605	0.1854	0.1854	0.1854	0.3457	0.4426	0.1854	–

Table 3: Performance of Entity Classification task across privacy budgets (ϵ). Dashes indicate settings where ϵ is not applicable.

Pipeline/Epsilon	8	16	32	64	128	256	512	1024	∞
Basline	–	–	–	–	–	–	–	–	0.62
NER	–	–	–	–	–	–	–	–	0.29
LLM	–	–	–	–	–	–	–	–	0.59
metricDP	–	–	–	–	–	0.050	0.300	0.180	–
RANTEXT	–	–	0.041	0.080	0.130	0.190	0.320	0.250	–
NER+metricDP	–	–	–	–	–	0.140	0.245	0.183	–
NER+RANTEXT	–	–	0.150	0.193	0.252	0.269	0.295	0.274	–
LLM+metricDP	–	–	–	–	–	0.040	0.270	0.150	–
LLM+RANTEXT	–	–	0.062	0.140	0.247	0.295	0.330	0.312	–

Table 4: Performance of Relation Classification task across privacy budgets (ϵ). Dashes indicate settings where ϵ is not applicable.