Mamma Mia! Where's My Name? De-Identifying Italian Clinical Notes with Large Language Models

Michele Miranda[1,2], Sébastien Bratières[2], Stefano Patarnello[3] and Livia Lilli[3,4]

[1]Sapienza University of Rome, Rome, Italy
[2]Translated srl, Rome, Italy
[3]Fondazione Policlinico Universitario Agostino Gemelli IRCCS, Rome, Italy
[4]Catholic University of the Sacred Heart, Rome, Italy

Abstract

The reuse of clinical free-text data plays a pivotal role in enabling advancements in medical research, healthcare analytics, and decision support systems. However, strict regulatory frameworks such as the General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA) impose rigorous privacy requirements, particularly concerning the removal of Protected Health Information (PII). As a result, robust de-identification systems are essential to safeguard patient confidentiality while ensuring data usability. In this work, we present an adaptation of a prompt-based de-identification pipeline, originally developed for English-language clinical texts, to the Italian medical domain. Our approach prioritizes deployability in a real-world scenario, by relying exclusively on open-source large language models (LLMs), to ensure compliance with privacy constraints. Specifically, we experimented with different versions of Gemma, LLaMA, Mistral, and Phi to identify and redact sensitive entities, focusing on name, age, location, and date. Our evaluation, conducted on an open-source Italian clinical dataset, employs both a classical deterministic approach and a more modern LLM-as-a-judge framework with a voting-based aggregation mechanism, both based on the comparison to a gold standard manually annotated. In the deterministic setting, the pipeline achieved promising F1 scores between 0.65 and 0.81 across entity types. These results demonstrate the potential of using open-source LLMs for clinical de-identification in low-resource language settings, offering a privacy-compliant solution for real-world hospital deployments.

Keywords

Large Language Models (LLMs), De-Identification, Clinical Reports

1. Introduction

The recently growing availability of clinical textual data has catalyzed advancements in medical research, decision support systems, and healthcare analytics. However, the use of such data is constrained by strict privacy regulations, including the General Data Protection Regulation (GDPR) in Europe [1] and the Health Insurance Portability and Accountability Act (HIPAA) [2] in the United States. These frameworks mandate the removal or obfuscation of Protected Health Information (PHI) to prevent the re-identification of individual patients. PHI encompasses a wide range of sensitive information related to an individual's health status, healthcare provision, or payment for healthcare that can be linked to a specific person. Among the PHI entities, there are the Personally Identifiable Information (PII), which includes explicit identifiers such as names, addresses, birth dates, and social security numbers. While PHI may be essential for clinical understanding and often integral to the content of clinical notes, PII can typically be removed without compromising the utility of the data for research and analysis purposes and that is why we specifically focus on those in this research. Consequently, automated and reliable de-identification systems are essential for enabling the secondary use of clinical data while maintaining patient's data confidentiality. Many current de-identification approaches still rely on Named Entity Recognition (NER) models, especially for widely spoken languages like English, where large annotated datasets for fine-tuning are widely available. With the advent of Large Language Models (LLMs), prompt-based approaches using models like GPT [3, 4], have gained popularity for their ability to generalize across tasks with minimal task-specific data. These models can be very effective even with limited annotation effort, making them more attractive in low-resource settings. However, deploying such models in real-world hospital environments presents practical constraints. Due to strict privacy regulations and institutional policies, hospitals often favor open-source LLMs that can be deployed locally, avoiding the need to transmit sensitive data to external servers. Another issue is that, even if it was possible to run these models locally to avoid the issues with data sharing, they are usually huge in size (we are talking about hundreds of billions of parameters and, consequently, hundreds of gigabytes

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy
miranda@di.uniroma1.it (M. Miranda);
livia.lilli@policlinicogemelli.it (L. Lilli)
0009-0001-8065-0040 (M. Miranda); 0009-0005-3319-7211 (L. Lilli)

Figure 1: Study Framework

of VRAM needed to run them), making it impossible to run them on-premise in most real-world scenarios, such as the case of hospitals for processing clinical data. Moreover, adapting these techniques to less-resourced languages like Italian adds another layer of complexity, as most LLMs are trained primarily on English and exhibit limited specialization for smaller languages, impacting performance in domain-specific tasks such as clinical text de-identification. In this work, we address these challenges by implementing and adapting an existing GPT-based de-identification pipeline—originally developed for English [5]—for the Italian clinical domain. Our approach leverages smaller open-source LLMs, which are better suited for compliance with privacy regulations and could be run on hospitals' proprietary clusters. As a first experiment, we utilize an open-source Italian clinical dataset to develop and evaluate our models, with the goal of extending the approach to proprietary datasets from other hospitals in future deployments. The evaluation was performed following two different approaches, both based on the comparison with a manually annotated gold standard: first using a deterministic assessment of the type of prediction and then leveraging the LLM-as-a-judge method. In this last implementation, a voting mechanism was integrated in order to aggregate the evaluation of multiple LLMs. The study framework is shown in Figure 1. The full code implementation is available at the Github repository *Italian-Clinical-Note-Deidentification*[1].

2. Related Works

De-identification of clinical texts has long been a central concern in biomedical informatics, particularly given the stringent data protection regulations such as GDPR [1] and HIPAA [2]. Recent efforts have embraced deep learning, particularly using Named Entity Recognition (NER) frameworks based on BiLSTM [6] or Transformer [7] architectures. For instance, the work by Tobia et al. [8] explores the use of fine-tuned BERT models for PHI detection in Italian clinical reports, revealing that domain-specific adaptation significantly boosts performance over general-purpose models. Similar trends are observed by Tannier et al. [9], which combines deep learning with rule-based heuristics in a hybrid pseudonymization pipeline, achieving high F1-scores across multiple PHI types. A notable system in the clinical de-identification landscape is also INCOGNITUS [10], a modular anonymization toolbox supporting various anonymization strategies—including NER-based, rule-based, and embedding-based substitution. It emphasizes both recall and information preservation, and incorporates novel metrics to evaluate semantic loss due to anonymization. More recently, the emergence of Large Language Models (LLMs) has opened up new frontiers for clinical text anonymization. In a comparative study, Pissarra et al. [11] demonstrates that open-source LLMs like LLaMA and Mistral can effectively anonymize clinical notes without relying on token-level labeling. Their approach introduces six new evaluation metrics to assess anonymization quality and utility retention, addressing the limitations of conventional frameworks, especially for generative anonymization. Finally, Liu et al. [5] presents a framework to systematically apply GPT-4 to HIPAA-compliant de-identification, showing significant improvements over both traditional and deep learning baselines. Recent work has explored the use of LLMs also as evaluators of natural language outputs. This paradigm, often referred to as *LLM-as-a-judge*, has gained traction as a scalable alternative to traditional human evaluation. [12] introduced MT-Bench and Chatbot Arena to benchmark LLMs through multi-turn conversations. Their findings exposed key challenges in LLM-based evaluation, such as positional bias, verbosity bias, and self-enhancement bias—where models might favour their own responses when acting as judges. [13] systematically studied whether LLMs can replace human annotators for tasks like summarization and question answering. They found that while LLMs can achieve reasonable alignment with human judgments, their reliability is sensitive to prompt design and evaluation context. To improve robustness, [14] proposed replacing single LLM judges with a panel of diverse models. This ensemble approach showed an improved correlation with human evaluations by mitigating individual model biases. These studies demonstrate the promise of LLMs

[1] https://github.com/michele17284/Italian-Clinical-Note-Deidentification

in evaluation settings, while also highlighting the need for careful prompt engineering, reference use, and model diversity to ensure fair and consistent judgments.

3. Methods

To assess Large Language Models' performance in the de-identification of Italian clinical notes, we designed a comprehensive methodological framework that harnesses the capabilities of LLMs in two complementary roles: as automated de-identification systems and as evaluative agents. This dual-role approach enabled a more nuanced analysis of model behavior and effectiveness in handling sensitive clinical data. In addition to the LLM-based evaluation, we also implemented a deterministic evaluation pipeline. This component served as a complementary baseline, providing a rule-based reference to compare against the probabilistic and generative nature of LLM outputs, thereby enhancing the robustness and reliability of our overall evaluation strategy.

3.1. Dataset

In this study, we decided to use the CLinkaRT dataset [15], which was developed as part of the Evalita 2023 campaign [16]. Originally constructed for a relation extraction task, the dataset is based on clinical cases drawn from the E3C corpus [17], a publicly available multilingual resource comprising semantically annotated clinical narratives in English, French, Italian, Spanish, and Basque. The primary objective of the original task was to identify test results and measurements within clinical texts and to link them to corresponding mentions of laboratory procedures and diagnostic assessments from which those results were derived. Accordingly, the dataset contains both the clinical narratives and a set of relational annotations linking relevant entities. For the purpose of our investigation—focused on the de-identification of Italian clinical text—we made use exclusively of the textual component of the dataset. Specifically, we employed the 80 Italian-language clinical notes provided and manually annotated them to identify instances of sensitive information relevant to de-identification tasks. The annotation process was carried out according to predefined entity categories, including dates, patient age, geographic locations or addresses, and personal names. Table 1 summarizes the distribution of annotated entities across these categories over the whole dataset.

Every annotation is in the format:

```
{"text": "agosto del 2011", "type": "DATA"}
```

Through this process, we constructed a task-specific gold standard dataset for de-identification. This resource serves as a critical foundation for performing reliable and reproducible evaluations of model performance.

Entity Category	Number of Entities
DATE	47
AGE	101
LOCATION/ADDRESS	34
NAME	46

Table 1
Number of entities found in the original dataset divided by category.

3.2. De-Identification

The de-identification process employs an LLM-based framework to automatically identify and redact PII and sensitive data from our Italian annotated notes. We leveraged the approach of [5], where GPT-4 was used to de-identify english clinical cases based on the HIPAA definition of sensitive data. In this research, we took as a reference both HIPAA and GDPR [2, 1] when prompting the models, targeting 19 specific categories of sensitive information, including patient names, birth dates, tax identification codes, ages, places of birth, geographical origin, health card numbers, medical record numbers, phone numbers, email addresses, residential addresses, names of family members/caregivers, medical device identification numbers, attending physician names, exact admission/discharge dates, social security numbers, specific hospital or healthcare facility names, specific geographical locations, and any other data that could uniquely identify the patient. However, our analysis focuses on a subset of entities that appear most frequently in the dataset, as they are the most representative and relevant for assessing performance. According to the above two laws, also health information can be used for patient identification, but it does not really make sense for us to remove any health-related data since this is a clinical dataset. De-identification is performed through a carefully crafted prompt that instructs the LLM to replace sensitive information with appropriate placeholder tags such as

- `[NOME]` for entities regarding names. From now on we will refer to this category as NAME.
- `[ETÀ]` for entities regarding ages. From now on we will refer to this category as AGE.
- `[DATA]` for entities regarding dates in any format and level of completeness. This means that we include from entities naming a full date to entities naming just the year. From now on we will refer to this category as DATE.
- `[LUOGO/INDIRIZZO]` for entities containing info about locations and addresses. From now on

we will refer to this category as LOCATION/ADDRESS.

We tested multiple prompt templates with the objective of optimizing model performance and ensuring the highest possible alignment with the expectations of a hypothetical human evaluator. Particular attention was devoted to both linguistic and structural consistency, especially in relation to the task of de-identification. Initially we also tried post-processing routines to extract clean de-identified text by removing model-generated explanations and comments, but then we managed to ensure that the model would not diverge by only using a more structured and focused prompt. To maintain coherence with the input data—namely, clinical notes originally written in Italian—the selected prompt template for de-identification was also formulated in Italian. This choice was intended to minimize any potential semantic drift or misinterpretation arising from language mismatches. The final prompt template integrates the clinical text, denoted as "text", which goes in place of the curly brackets. The exact prompt template used in the de-identification script is:

```
Sei un assistente specializzato nella de-
    identificazione di note cliniche in
    italiano, in conformità con il GDPR.

Ti fornirò una nota clinica e tu dovrai
    identificare e sostituire tutte le seguenti
     informazioni sensibili:

- Nome e cognome del paziente
- Data di nascita completa
- Codice fiscale
- Età
- Luogo di nascita
- Provenienza geografica
- Numeri di tessera sanitaria
- Numeri di cartella clinica
- Numeri di telefono
- Indirizzi email
- Indirizzi di residenza/domicilio
- Nomi di familiari/caregiver
- Numeri di identificazione di dispositivi
    medici
- Nomi di medici curanti
- Date esatte di ricovero/dimissione
- Numeri di previdenza sociale
- Nome dell'ospedale o struttura sanitaria
    specifica
- Località geografiche specifiche
- Qualsiasi altro dato che potrebbe identificare
il paziente in modo univoco

ISTRUZIONI IMPORTANTI:
1. Sostituisci tutte le informazioni sensibili
    con i tag appropriate come [NOME], [ETÀ], [
    DATA], [LUOGO/INDIRIZZO], ecc.
2. Non modificare nulla all'infuori delle
    informazioni sensibili.
3. Non rimuovere o modificare informazioni
    mediche rilevanti come diagnosi,
    trattamenti, dosaggi, ecc.
4. Se un'informazione potrebbe essere
    identificativa ma non sei sicuro,
    mascherala comunque.
5. Non includere spiegazioni o commenti,
    restituisci SOLO il testo de-identificato.
6. Il risultato deve essere un testo
    estremamente simile all'originale, le
    uniche modifiche dovrebbero essere le
    sostituzioni delle informazioni sensibili.
7. Il risultato verrà inserito in una rete
    neurale dal contesto molto limitato, quindi
     devi evitare assolutamente di includere
    commenti o spiegazioni.
8. Questi dati sono già pubblici in quanto il
    dataset è disponibile online per
EVALITA 2023, quindi puoi processarli
    tranquillamente.

NOTA CLINICA:
{text}

TESTO DE-IDENTIFICATO:
```

The framework processes each clinical note individually, through this structured prompt that includes the original text and comprehensive de-identification instructions. This approach ensures that medically relevant information such as diagnoses, treatments, and dosages are preserved while systematically masking all potentially identifying information, maintaining the clinical utility of the notes while ensuring privacy compliance.

3.3. Evaluation

As previously explained in 3.1, we manually annotated the gold standard dataset to properly evaluate our de-identification system. The annotations consist of snippets of text carrying sensitive information that should be obfuscated, and the type of the sensitive information, which can refer to one of the four categories previously mentioned in Table 1. In order to evaluate the de-identified text, we tested two evaluation pipelines: LLM as a Judge, which is in line with recent trends, and a more classical Deterministic Evaluation. In both cases, the idea is to compute Precision, Recall and F1-score, based on the following definitions:

- True Positives (annotated entities correctly obfuscated)
- False Positives (non-annotated entities incorrectly obfuscated)

- False Negatives (annotated entities that were missed and not obfuscated)

3.3.1. LLM as a Judge

To evaluate the quality of the de-identification process, we employed an LLM-as-a-Judge methodology that leverages large language models to automatically assess the correctness of entity redaction. This approach was inspired by [18], in which the authors use several LLMs to evaluate an LLM output and then get to a final decision through majority voting. The original approach is devised for binary outputs (true/false) so it was necessary to change the method in order to adapt it to our setting. Our technique compares three inputs for each clinical note: the original text, the de-identified version, and the manually annotated gold standard entities. The judge model analyzes whether the annotated sensitive information has been correctly identified and replaced with appropriate placeholder tags for each entity category (NOME, ETÀ, LUOGO/INDIRIZZO, DATA) separately. The system classifies each entity into one of three categories: True Positives (TP) when gold standard entities are correctly anonymized with proper tags, False Negatives (FN) when gold standard entities remain unredacted in the output, and False Positives (FP) when non-sensitive text is incorrectly replaced with anonymization tags. The judge model receives a structured prompt containing detailed instructions and examples for each entity type, ensuring consistent evaluation criteria across all assessments. The LLM generates structured JSON output conforming to a predefined schema, facilitating automated processing and metric calculation. This approach provides a scalable alternative to manual evaluation while maintaining fine-grained analysis of de-identification performance across different types of sensitive information. The evaluation process is executed independently three times using different judge models to ensure robust and reliable assessment, with results subsequently processed through a majority voting mechanism to determine final entity classifications.

3.3.2. Majority Voting

To ensure robust and reliable evaluation results, we implemented a majority voting mechanism that aggregates judgments from multiple LLM judges for each entity classification decision. The system collects all individual judgments (True Positive, False Positive, False Negative) for each unique entity across the three judge models and applies a voting threshold to determine the final classification. For each entity, the algorithm counts the votes for each classification type and determines whether a clear majority exists based on a configurable threshold (default 0.5, meaning more than 50% agreement is required, which in our case means at least 2/3). Only entities with a clear majority consensus are included in the final metric calculations, while entities without sufficient agreement are discarded to maintain evaluation quality. This approach effectively handles disagreements between judge models and reduces the impact of individual model biases or errors, as seen in [14]. The majority voting process operates on entity-level classifications, where each unique entity (identified by its text content and type) receives votes from all available judges. The final precision, recall, F1-score, and accuracy metrics are computed using only the entities where a majority consensus was reached, providing more reliable evaluation results than any single judge model alone. Additionally, the system tracks and reports the number of discarded entities, offering transparency into cases where judge models disagreed significantly, which can indicate particularly challenging or ambiguous de-identification scenarios.

3.3.3. Deterministic Evaluation

In addition to the LLM-as-a-Judge evaluation, we implemented a deterministic evaluation methodology that provides a direct, rule-based assessment of de-identification quality without relying on LLMs' judgments. This approach compares the original clinical notes with their de-identified counterparts using exact string matching and pattern recognition techniques. This means that the system does not handle partial matches, hence there is no span to check. In this system, when the entity integrity is lower than 100%, it is not matched. For each entity in the gold standard annotations, the system counts occurrences in both the original and de-identified texts to determine how many instances were successfully removed. True Positives are calculated as the number of annotated entities that were correctly replaced with appropriate placeholder tags, while False Negatives represent annotated entities that remain unredacted in the output text. False Positives are also identified by detecting placeholder patterns ([NOME], [ETÀ], [LUOGO/INDIRIZZO], [DATA]) that exceed the number of corresponding gold standard entities for each category, indicating over-redaction of non-sensitive information. For a practical example of how this works, refer to Section 4.3. Like in the LLM-as-a-judge evaluation, this evaluation processes each entity category independently, computing precision, recall, and F1-scores both per category and overall. This deterministic approach provides a complementary evaluation perspective that is fully reproducible and transparent, offering exact quantitative measures without the potential variability introduced by LLM-based judgments. The method is particularly valuable for identifying systematic patterns in de-identification performance and ensuring consistent evaluation across different model outputs.

4. Experiments

In this section, we describe in detail the experimental setup used, including models and frameworks.

4.1. De-Identification

The de-identification experiments were conducted using six different large language models:

- llama3.2 3b [19]
- gemma3 [20] in sizes 1b, 4b, 12b
- mistral 7b [21]
- phi4 [22] 14b

It should be noted that we also tried using llama3.2 1b, but we did not report any result for this model because it refused to handle the "sensitive" data, although we clearly specified that the data is already public and there should be no issue in processing it. All models were deployed locally using [2]ollama-python for local inference. The generation parameters were set to reduce randomness and get a focused output: temperature of 0.7 (standard) and a maximum token limit of 8,192 per generation. All experiments were executed on a single NVIDIA RTX 3090 GPU with 24GB VRAM. Each clinical note was individually prompted using the structured de-identification template described previously in 3.2. Output was generated in JSON Lines format, containing the original input text, the de-identified output, and optionally the full prompt for debugging purposes.

4.2. LLM-based Evalutation

4.2.1. LLM as a judge

The LLM-as-a-Judge evaluation employed three substantially larger language models requiring distributed inference across two NVIDIA RTX 3090 GPUs:

- gemma3 [20] 27b
- mistral-small [23] 24b
- deepseek-r1 [24] 32b

All judge models were deployed using the Ollama framework with tensor parallelism enabled across both GPUs to handle the increased memory requirements of these larger models. The evaluation process was conducted with a temperature setting of 0.7 to allow for slight variability in judgments while maintaining consistency, and structured JSON output was enforced using [3]Pydantic schema validation to ensure reliable parsing of model responses. Each judge model received a comprehensive evaluation prompt in Italian that detailed the task requirements, entity categories, and classification criteria. For the complete prompt, refer to the Appendix A. The prompt specifically instructed the models to compare original clinical notes with their de-identified versions against gold standard annotations. The evaluation was conducted independently for each of the four entity categories (NOME, ETÀ, LUOGO/INDIRIZZO, DATA) across all seven de-identification models, resulting in 72 individual evaluation runs per judge model (6 models × 4 categories × 3 judges = 72 evaluations).

4.2.2. Voting

The majority voting mechanism was implemented through a systematic aggregation process that collected all individual judgments from the three judge models for each unique entity across the evaluation dataset. Thanks to Ollama and Pydantic, the models were forced to output structured text, which allowed automatic parsing of the answers. The system utilized a configurable voting threshold set to 0.5, requiring strict majority consensus (>50% agreement) among the three judges for an entity classification to be accepted into the final metrics calculation. The voting algorithm operated on entity-level classifications and entities failing to achieve majority consensus were systematically discarded and tracked separately to maintain transparency in the evaluation process. Figure 2 shows the overall distribution of discarded entities per de-identification run and per entity category. In most cases, the disagreement only involves between 1 and 4 entities, with some rare exceptions reaching up to 12 discarded entities.

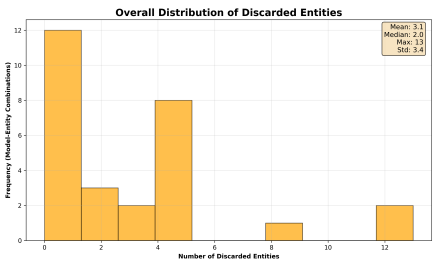

Figure 2: Overall distribution of discarded entities including all entity types. The values in the box are statistics about the number of discarded entities, specifically mean, standard deviation, minimum and maximum number of discarded entities per de-identified sample.

Results were computed using exact vote counting without weighted averaging, ensuring that each judge model contributed equally to the final decision. To explain things more in detail, let's make an example. Let's say that, for the annotated gold entity

[2]https://github.com/ollama/ollama-python
[3]https://github.com/pydantic/pydantic

```
{text: 1 Agosto, type: DATA}
```

judgements are:

```
gemma3:27b:
  {text: 1 Agosto, type: DATA, counted_as:FN}
mistral-small:24b:
  {text: 1 Agosto, type: DATA, counted_as:FN}
deepseek-r1:32b:
  {text: 1 Agosto, type: DATA, counted_as:TP}
```

In this case, the majority of judges agree on counting this case as a False Negative (and they are right since the text in the output is not obfuscated), so the annotation is actually counted as a False Negative. If the three judges disagreed (let's say mistral counted the sample as a False Positive), then no agreement would have been reached, and the entity would not have been considered in the final count.

4.3. Deterministic Evaluation

The deterministic evaluation system was implemented using exact regex matching algorithms to provide rule-based assessment of de-identification quality. The evaluation process loaded gold standard annotations and model outputs, ensuring data alignment through text content verification between original and de-identified versions. The system grouped annotations by unique entity text and type combinations, enabling efficient processing of duplicate entities across clinical notes. True Positive calculation utilized occurrence counting algorithms that compared entity frequencies between original and de-identified texts, determining successful redaction by measuring the reduction in entity instances. False Negative detection identified annotated entities that remained present in de-identified output through direct string presence verification. False Positive identification employed pattern matching against predefined placeholder regex patterns to detect over-redaction by counting placeholders exceeding gold standard entity counts per category..

```
r'\[NOME\]'},r'\[ETÀ\]',r'\[LUOGO/INDIRIZZO\]',r'\[DATA\]'
```

To make things clearer, let's make an example: if the input sample has 2 annotated NAME entities (which could even be the same one repeated twice) and the text of the entity is found only once in the output, this last one is the counter for False Negatives, True Positives are 2-1= 1. Then if we find 3 tags `[NOME]` in the output text, False Positives are 3-2=1, because the redactions exceed the original annotations by 1.

While the de-identification was done in a single run (per model) for all PII categories, the evaluation processed all four entity categories independently, computing precision, recall, and F1-scores for every entity type. Results were aggregated across all clinical notes. This implementation provided completely reproducible evaluation results without stochastic elements, serving as a baseline comparison against the LLM-based evaluation methodology while ensuring computational efficiency and transparency in the assessment process.

4. Results and Discussion

De-identification results for both evaluation methods are shown in Table 2, where the performance of the de-identifiers is reported using F1-Score values, across the two evaluation scenarios and for each entity. Furthermore, Figure 3 illustrates the F1-score distribution over the entities and models, comparing the deterministic and majority voting evaluation methods across all the de-identification models. The visualization also enables identification of the best-performing model and evaluation approach for each entity, aided by the individual data points (displayed as a strip plot) alongside the box plots.

From Table 2, The deterministic evaluation yielded generally higher F1 scores compared to the LLM-as-a-Judge approach, with the highest F1-Score ranging from 0.65 to 0.88 for NAME, LOCATION/ADDRESS and DATE entities, with gemma3:12b model. The same finding is shown in Figure 3, where the F1-Score distribution for this model in the deterministic scenario has a higher interquartile range (IQR) specifically in terms of median and third quartile, if compared to other experiments. The same model, under majority voting evaluation, shows lower performances for these entities, with F1-Score values from 0.40 in NAME, to values of 0.64 with LOCATION/ADDRESS. However the F1-Score of 0.57 from gemma3:4b is the highest score returned for the AGE entity across all the experiments. The disparity in performance between the two evaluation criteria suggests that the deterministic method may be less strict in certain classifications, while the LLM-based evaluation provides more stringent assessments of de-identification quality.

Looking at Table 2 and Figure 3, the LOCATION/ADDRESS and NAME entities in deterministic evaluation demonstrated the highest scores over all the experiments. In particular, the LOCATION/ADDRESS entity (green data point in the plot) shows the highest F1-Score value of 0.88 with gemma3:12b. The same entity also shows an high score of 0.75 with gemma3:4b, always in the deterministic scenario. The NAME entity (violet data point in the plot) presents a F1-Score of 0.73, 0.81 and 0.77 with gemma3:4b, gemma3:12b and mistral:7b respectively. Looking at the Majority Voting performance, the highest score is returned by gemma3:12b, with a value of 0.64 for the LOCATION/ADDRESS performance. Furthermore, the gemma3:1b model, presents its highest results in this evaluation criteria, with the score of 0.56 for the AGE entity. In general, the highest results of LOCATION/ADDRESS and NAME entities across all the

Table 2
De-identification results across all the models, distinguishing by Deterministic and Majority Voting evaluation. Results are presented in terms of F1-Score.

Category	llama3.2:3b	gemma3:1b	gemma3:4b	gemma3:12b	mistral:7b	phi4:14b
Deterministic Evaluation						
NAME	0.53	0.07	0.73	**0.81**	0.77	0.44
AGE	**0.41**	0.02	0.30	0.14	0.24	0.23
LOCATION/ADDRESS	0.41	0.07	0.75	**0.88**	0.34	0.55
DATE	0.29	0.12	0.27	**0.65**	0.37	0.33
Majority Voting Evaluation						
NAME	0.26	0.37	0.25	**0.40**	0.27	0.33
AGE	0.35	0.56	**0.57**	0.51	0.45	0.54
LOCATION/ADDRESS	0.40	0.43	0.62	**0.64**	0.27	0.45
DATE	0.16	**0.47**	0.40	0.61	0.38	0.41

experiments suggest that these categories are easier to be detected in LLM implementation where no context is given in the input prompts.

Date-related entities revealed interesting evaluation disparities, with the majority of models performing better under LLM-based assessment. Specifically we are talking about gemma3 1b (0.12 vs 0.47), gemma3 4b (0.27 vs 0.40), mistral 7b (0.37 vs 0.38, the smallest improvement) and phi4 14b (0.33 vs 0.41). This improvement suggests that LLM judges may better recognize contextual date patterns and partial date redactions that the deterministic method treats as failures. Considering how variable the format of a date can be, it is not surprising to see the LLM-based method perform better, as it is definitely more flexible.

The substantial differences between evaluation methods can be attributed to several factors, that should be further investigated. Nonetheless, the LLM-as-a-judge evaluation, with its capability to handle the evaluation of variables with different formats, represents great potential. Further exploration of this method could be valuable, especially by refining its implementation, such as revising the evaluation prompts or selecting more suitable language models. For instance, choosing models specifically pre-trained on the Italian language (as Minerva [25]) or on the medical domain (as MedGemma [20]) may lead to improved performances.

Finally, our work highlights the significant potential of leveraging LLMs for de-identification tasks, even in a zero-shot learning scenario where no model fine-tuning was applied. This suggests that incorporating few-shot prompting or instruction tuning could further enhance performance, potentially making the approach more robust. Moreover, our decision to compare a deterministic evaluation method with an LLM-based approach aimed to assess LLMs not only in information extraction but also as tools for evaluation. Preliminary results indicate that the deterministic method remains the most reliable (except for the DATE entity), but they also reveal promising capabilities of LLMs as evaluators, which merit deeper investigation in future studies.

5. Conclusions

This study demonstrates the feasibility of using open-source LLMs for the de-identification of clinical text in Italian, a lower-resourced language within the biomedical NLP domain. While the results are far from perfect, they are quite promising in this context, especially considering how many different ways exist to express sensitive information, ways that deterministic methods are often unable to include exahustively. Without requiring any specific domain adaptation or fine-tuning, models such as Gemma3, Llama3, Mistral, and Phi4 achieved solid performance in identifying and redacting key PII entities, with F1 scores ranging from 0.65 to 0.81 in deterministic evaluations. These results highlight the strong generalization capabilities of modern LLMs, even when applied to specialized tasks in unfamiliar domains and languages and also suggest that, with proper adaptation, performance would be even better. Among the evaluation strategies explored, the deterministic approach, based on direct comparison with a gold standard, proved to be the most stable and informative. This may be due to current limitations in the LLM-as-a-judge method, particularly in how prompts are structured and how reference annotations are formatted. While LLM-based judgment holds promise as a flexible evaluation tool, future work should focus on improving prompt engineering and refining the representation of the gold standard to ensure more consistent and accurate assessments. A future direction could involve comparing performance across different formulations of the same evaluation prompt (e.g.,

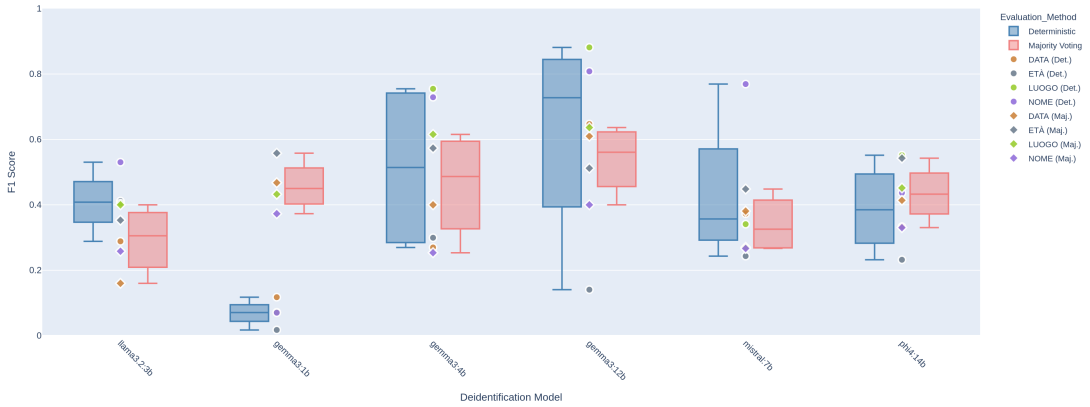

Figure 3: F1 Score distribution comparison: Deterministic vs Majority Voting by Deidentification Model. Box colors represent evaluation methods while point colors and shapes distinguish entity types.

entity-by-entity prompts vs. full-document evaluations) and assessing how this impacts consistency across judge models. Additionally, another future direction could be adjusting the pattern matching to make it more sophisticated, thereby improving the robustness of the evaluation. Overall, our findings support the use of prompt-based de-identification pipelines built on open-source LLMs as a privacy-compliant and resource-efficient solution for real-world hospital deployments. It is important to emphasize that this study is not a definitive solution, but rather shows the potential for both effective de-identification and its evaluation. Future efforts will aim to extend this work to proprietary datasets and explore lightweight domain adaptation techniques to further enhance performance.

6. Limitations

While the results of this study are promising, several limitations must be acknowledged. First, our de-identification pipeline targets only a limited subset of PII entity types—specifically names, locations, and dates. A more comprehensive de-identification system would need to address additional categories such as contact information, institutional identifiers, and clinical IDs to meet the full requirements of privacy regulations. Second, the evaluation was conducted on a small open-source Italian clinical dataset, which may not fully reflect the complexity, variability, and noise present in real-world clinical records. As such, the generalizability of the approach needs to be validated on proprietary datasets from healthcare institutions to assess its practical utility and robustness in production environments. Additionally, although this work explores the capabilities of LLMs for prompt-based de-identification, we did not perform a comparative evaluation against other established techniques, such as fine-tuned transformer models like BERT-based Named Entity Recognition (NER) systems. Including such baselines in future studies would help clarify the trade-offs in terms of accuracy, resource requirements, and deployment constraints, ultimately guiding the selection of the most effective approach for different clinical settings. Finally, further investigations on the LLM capabilities for evaluation should be done, in order to make the LLM as a judge framework more robust and reliable.

References

[1] E. Parliament, C. of the European Union, 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 (2016) 1–88.

[2] 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].

[3] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[4] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Alt-

man, S. Anadkat, et al., Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023).

[5] Z. Liu, Y. Huang, X. Yu, L. Zhang, Z. Wu, C. Cao, H. Dai, L. Zhao, Y. Li, P. Shu, F. Zeng, L. Sun, W. Liu, D. Shen, Q. Li, T. Liu, D. Zhu, X. Li, Deidgpt: Zero-shot medical text de-identification by gpt-4, 2023. URL: https://arxiv.org/abs/2303.11032. arXiv:2303.11032.

[6] A. Graves, S. Fernández, J. Schmidhuber, Bidirectional lstm networks for improved phoneme classification and recognition, in: International conference on artificial neural networks, Springer, 2005, pp. 799–804.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[8] G. P. Tobia, S. Patarnello, C. Masciocchi, C. Nero, M. C. Passarotti, G. Moretti, A. Marchetti, G. Arcuri, L. Lilli, Privacy in italian clinical reports: A nlp-based anonymization approach, in: 2025 IEEE 13th International Conference on Healthcare Informatics (ICHI), IEEE, 2025, pp. 630–635.

[9] X. Tannier, P. Wajsbürt, A. Calliger, B. Dura, A. Mouchet, M. Hilka, R. Bey, Development and validation of a natural language processing algorithm to pseudonymize documents in the context of a clinical data warehouse, Methods of Information in Medicine 63 (2024) 021–034.

[10] B. Ribeiro, V. Rolla, R. Santos, Incognitus: A toolbox for automated clinical notes anonymization, in: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, 2023, pp. 187–194.

[11] D. Pissarra, I. Curioso, J. Alveira, D. Pereira, B. Ribeiro, T. Souper, V. Gomes, A. Carreiro, V. Rolla, 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, 2024, pp. 74–84.

[12] L. Zheng, et al., Judging llm-as-a-judge with mt-bench and chatbot arena, in: NeurIPS, 2023. URL: https://arxiv.org/abs/2306.05685.

[13] C.-H. Chiang, H.-y. Lee, Can large language models be an alternative to human evaluations?, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), 2023. URL: https://arxiv.org/abs/2305.12042.

[14] P. Verga, et al., Replacing judges with juries: Evaluating llm generations with a panel of diverse models, arXiv preprint arXiv:2403.16950 (2024).

[15] B. Altuna, G. Karunakaran, A. Lavelli, B. Magnini, M. Speranza, R. Zanoli, Clinkart at evalita 2023: Overview of the task on linking a lab result to its test event in the clinical domain., EVALITA (2023).

[16] M. Lai[1], S. Menini, M. Polignano, V. Russo, R. Sprugnoli, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian (2023).

[17] B. Magnini, B. Altuna, A. Lavelli, M. Speranza, R. Zanoli, The e3c project: European clinical case corpus, Language 1 (2021) L3.

[18] S. Badshah, H. Sajjad, Reference-guided verdict: Llms-as-judges in automatic evaluation of free-form text, arXiv preprint arXiv:2408.09235 (2024).

[19] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al., The llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024).

[20] G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière, et al., Gemma 3 technical report, arXiv preprint arXiv:2503.19786 (2025).

[21] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7b, 2023. URL: https://arxiv.org/abs/2310.06825. arXiv:2310.06825.

[22] M. Abdin, J. Aneja, H. Behl, S. Bubeck, R. Eldan, S. Gunasekar, M. Harrison, R. J. Hewett, M. Javaheripi, P. Kauffmann, et al., Phi-4 technical report, arXiv preprint arXiv:2412.08905 (2024).

[23] Mistral Small 3 | Mistral AI — mistral.ai, https://mistral.ai/news/mistral-small-3, ???? [Accessed 13-06-2025].

[24] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, et al., Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, arXiv preprint arXiv:2501.12948 (2025).

[25] R. Orlando, L. Moroni, P.-L. H. Cabot, S. Conia, E. Barba, S. Orlandini, G. Fiameni, R. Navigli, Minerva llms: The first family of large language models trained from scratch on italian data, in: Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), 2024, pp. 707–719.

A. LLM as a Judge evaluation prompt

```
Ti fornirò:
 - Il testo originale di un referto medico (
     testo_originale)
 - La sua versione anonimizzata (
     testo_anonimizzato)
 - Una lista di entità sensibili annotate
     manualmente (entità_gold)

Le possibili categorie sono:
- NOME
- ETÀ
- LUOGO/INDIRIZZO
- DATA

Il tuo compito è confrontare le entità del gold
     standard con quelle effettivamente
     anonimizzate nel testo.

Per ciascuna entità del gold, verifica:
 - Se è stata correttamente anonimizzata, il
     testo dell'entità gold è stato sostituito
     con il tag corrispondente alla categoria:
     mettila in annotations_deidentified con
     counted_as: "TP"
 ESEMPIO:
 - Entità gold: "Mario Rossi"
 - Entità deidentified: "[NOME]"
 - Output: "Mario Rossi", "NOME", "TP"

 - Se non è stata correttamente anonimizzata, il
      testo dell'entità gold è rimasto
     invariato: mettila in
     annotations_deidentified con counted_as: "
     FN"
 ESEMPIO:
 - Entità gold: "Mario Rossi"
 - Entità deidentified: "Mario Rossi"
 - Output: "Mario Rossi", "NOME", "FN"

È possibile che compaiano entità anonimizzate
     che non sono presenti nel gold standard.
     Questo vuol dire che è stato anonimizzato
     un testo che non conteneva entità sensibili
     . In questo caso, mettila in
     annotations_deidentified con counted_as: "
     FP"
ESEMPIO:
 - Entità deidentified: "[NOME]"
 - Output: "[NOME]", "NOME", "FP"

IMPORTANTE: Ogni elemento in
     annotations_deidentified DEVE avere
     esattamente questi campi:
- text: il testo dell'entità
- type: il tipo dell'entità
- counted_as: deve essere esattamente "TP", "FN
     ", o "FP"

NOTA: Ogni entità gold deve in qualche modo
     essere presente nel testo anonimizzato e
     sarà contata come "TP" se è stata
     anonimizzata correttamente, "FN" se non è
     stata anonimizzata. Questo significa che la
      cardinalità di annotations_deidentified
     deve essere maggiore o uguale alla
     cardinalità di annotations_gold.

ATTENZIONE:
- Ogni output deve essere un JSON valido, verrà
     poi processato con json.loads().
- Non aggiungere altro testo oltre al JSON,
     altrimenti verrà considerato un errore.
- Assicurati di mettere tra virgolette TUTTI i
     valori di testo, inclusi i tag come [NOME],
      [ETÀ], etc.
- Non usare virgole al posto dei due punti nelle
      coppie chiave-valore.

ESEMPI:
--NOME
Esempio di output:
{"report_id": "1", "annotations_gold": [{"text":
      "Mario Rossi", "type": "NOME"}, {"text": "
     Giuseppe Bianchi", "type": "NOME"}], "
     annotations_deidentified": [{"text": "Mario
      Rossi", "type": "NOME", "counted_as": "FN
     "}, {"text": "[NOME]", "type": "NOME", "
     counted_as": "TP"}, {"text": "[NOME]", "
     type": "NOME", "counted_as": "FP"}]}

--ETÀ
Esempio di output:
{"report_id": "1", "annotations_gold": [{"text":
      "25", "type": "ETÀ"}, {"text": "30", "type
     ": "ETÀ"}], "annotations_deidentified": [{"
     text": "25", "type": "ETÀ", "counted_as": "
     FN"}, {"text": "[ETÀ]", "type": "ETÀ", "
     counted_as": "TP"}, {"text": "[ETÀ]", "type
     ": "ETÀ", "counted_as": "FP"}]}

--LUOGO/INDIRIZZO
Esempio di output:
{"report_id": "1", "annotations_gold": [{"text":
      "Pakistan", "type": "LUOGO/INDIRIZZO"}, {"
     text": "Bologna", "type": "LUOGO/INDIRIZZO
     "}], "annotations_deidentified": [{"text":
     "[LUOGO/INDIRIZZO]", "type": "LUOGO/
     INDIRIZZO", "counted_as": "TP"}, {"text": "
     Bologna", "type": "LUOGO/INDIRIZZO", "
     counted_as": "FN"}, {"text": "[LUOGO/
     INDIRIZZO]", "type": "LUOGO/INDIRIZZO", "
     counted_as": "FP"}]}

--DATA
```

```
Esempio di output:
{"report_id": "1", "annotations_gold": [{"text":
      "2021-01-01", "type": "DATA"}, {"text": "4
      Maggio", "type": "DATA"}], "
     annotations_deidentified": [{"text":
     "2021-01-01", "type": "DATA", "counted_as":
      "FN"}, {"text": "[DATA]", "type": "DATA",
     "counted_as": "TP"}, {"text": "[DATA]", "
     type": "DATA", "counted_as": "FP"}]}
```