A Comparative Study of Approaches to Anonymization of Clinical Free Text in Spanish

Florencia Brunello[1,2], **Serena Villata**[2], **Laura Alonso Alemany**[1], **Milagro Teruel**[1]
[1]Universidad Nacional de Córdoba, [2]Université Côte d'Azur
{florenciabrunello}@unc.edu.ar

Abstract

The anonymization of clinical free-text records is a prerequisite for enabling the secondary use of healthcare data while preserving patient privacy. This challenge is particularly acute for Spanish clinical text, where annotated resources are scarce and practitioners lack clear empirical guidance on which technological approaches are more adequate to their particular restrictions and capabilities.

In this work, we present a controlled comparative study of representative anonymization paradigms for Spanish clinical narratives, including a baseline rule-based approach, a general-purpose large language model under prompt-based inference, an off-the-shelf industrial NLP toolkit (spaCy) and comparable neural sequence labeling architectures. To ensure a fair and contamination-aware evaluation, particularly given the opacity of pretrained model training data, we introduce a synthetic clinical dataset designed to mirror the structural properties of established benchmarks while avoiding direct overlap.

Across heterogeneous evaluation settings, recurrent neural network architectures, particularly the off-the-shelf spaCy toolkit, consistently achieve the best balance between effectiveness, computational efficiency, and deployment feasibility. In contrast, the evaluated large language model and rule-based approach exhibit limited robustness in this domain. We further observe that training task-specific embeddings end-to-end yields stronger generalization than incorporating pretrained representations.

Although limited to Spanish and to representative instances of each paradigm, the study identifies stable performance tendencies across datasets. These results provide actionable guidance for institutions seeking to implement anonymization pipelines under realistic infrastructure constraints, and substantially narrow the technological search space for practitioners operating in under-resourced clinical NLP settings.

This work contributes reproducible evaluation procedures, synthetic benchmark resources, and empirical evidence for privacy-preserving clinical NLP in Spanish—an area where comparative studies remain limited.

Keywords: Clinical Text Anonymization, Spanish Clinical NLP, De-identification Methods Comparison

1. Introduction and Motivation

Clinical narratives such as discharge summaries, progress notes, and radiology reports contain rich, fine-grained information that is essential for clinical research, quality assessment, and healthcare optimization (Stubbs et al., 2015). However, these documents frequently include personally identifiable information (PII), making their use subject to strict legal, ethical, and regulatory constraints, including frameworks such as GDPR (European Parliament and Council of the European Union) and HIPAA (U.S. Congress, 1996).

De-identification, the removal or transformation of personally identifiable information to prevent re-identification of patients and workers, constitutes a core component of clinical data anonymization. It enables controlled data sharing across institutions and research initiatives while mitigating privacy risks. Although anonymization cannot eliminate all re-identification risks, it plays a central role in enabling the secondary use of healthcare data and supporting reproducible clinical research.

Despite its importance, anonymization remains a technically demanding task. Even manual de-identification has been shown to be time-consuming, cognitively demanding, and subject to inter-annotator variability (Dorr et al., 2006). Clinical text is characterized by high variability across institutions and specialties, frequent spelling errors, unconventional abbreviations, domain-specific shorthand, and elliptical phrasing. These properties make automatic anonymization particularly challenging and sensitive to domain shifts.

Over the years, a wide range of approaches has been proposed, including rule-based systems, statistical sequence models, and neural architectures. More recently, large pretrained language models (LLMs) have been applied to clinical de-identification, promising improved generalization and reduced need for feature engineering (Staab et al., 2024; Liu et al., 2025). At the same time, industrial NLP toolkits such as spaCy offer readily deployable pipelines that lower the entry barrier for institutions seeking practical solutions.

However, practitioners, particularly in under-resourced language settings, often face a fundamental question: which paradigm should be prioritized under realistic infrastructure and regulatory constraints? While English clinical NLP bene-

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 270–280

fits from multiple comparative evaluations, Spanish clinical text remains comparatively under-studied. Annotated corpora are scarce, cross-dataset analyses are limited, and empirical guidance for technology selection is lacking. As a result, institutions frequently must make high-stakes technological decisions with limited evidence.

Evaluation practices further complicate this landscape. Performance is commonly measured on a small number of public benchmarks, such as i2b2 (Stubbs et al., 2015) and MEDDOCAN (**?**). For pretrained LLMs in particular, the opacity of training data raises concerns about benchmark contamination and potential information leakage, which may lead to overly optimistic performance estimates and hinder fair cross-paradigm comparison.

Robust evaluation therefore requires not only multi-dataset validation but also methodological safeguards against data leakage, especially when assessing large pretrained models. Additionally, clinical documentation varies substantially across institutions, languages, and document types. Systems that perform well on a single benchmark may fail to generalize in real-world deployment settings. For practitioners, this variability underscores the need for evaluation protocols tailored to local usage scenarios.

In this paper, we present a controlled and contamination-aware comparison of representative anonymization paradigms for Spanish clinical text. Specifically, we evaluate: (i) a baseline rule-based approach, (ii) general-purpose large language model under prompt-based inference, (iii) an off-the-shelf industrial NLP toolkit (spaCy), and (iv) fully trainable neural sequence labeling architectures based on recurrent neural networks. To mitigate benchmark contamination and ensure a fair comparison, we introduce two synthetic datasets that reproduce the structural and statistical characteristics of clinical narratives without replicating existing corpora. We complement these resources with evaluation on the recently released CARMEN-I dataset of naturally occurring Spanish clinical records (Farre Maduell et al., 2024).

By combining synthetic and natural datasets under uniform experimental conditions, this study aims to identify stable performance tendencies across paradigms rather than isolated benchmark scores. Our goal is not to claim universal superiority of any single architecture, but to provide empirically grounded guidance that narrows the technological search space for institutions implementing anonymization pipelines in Spanish clinical settings.

The remainder of this paper is structured as follows. Section 2 reviews related work in clinical de-identification, with special attention to Spanish-language resources. Section 3 describes the datasets and the rationale behind the synthetic data generation. Section 4 outlines the evaluation methodology, and Section 5 presents the anonymization approaches compared. Section 6 reports the experimental results. Finally, we discuss implications, limitations, and practical recommendations for deployment.

2. Relevant Work

Early work on clinical text anonymization relied primarily on rule-based systems, often implemented using regular expressions and domain-specific dictionaries (Neamatullah et al., 2008; Uzuner et al., 2007). These approaches offer transparency, deterministic behavior, and low computational cost, but they typically suffer from limited coverage, brittle pattern matching, and poor generalization to unseen writing styles or institutions because they are typically based on the concrete cases seen by the people who develop the systems, thus tailored to specific practices, and also because it seems impossible to have a rule-based system cover all use cases, as the resulting set of rules would most probably be inconsistent, with contradictory rules (Meystre et al., 2010). Similar tendencies have been observed in Spanish clinical text, where handcrafted rules are effective for highly unambiguous, regular entities such as dates or identification numbers, but struggle with ambiguity (specially abbreviations), lexical variability and informal writing (González et al., 2019).

Machine learning approaches, especially sequence-aware approaches like Conditional Random Fields (CRFs) and recurrent neural networks, later became dominant due to their ability to model contextual dependencies and reduce manual rule engineering, and, with deep learning, also reducing feature engineering (Dernoncourt et al., 2017; Lavergne et al., 2010). In the context of Spanish, shared tasks such as MEDDOCAN fostered the development of supervised sequence labeling systems for clinical de-identification (Martínez et al., 2019). Many top-performing systems in MEDDOCAN relied on bidirectional LSTM architectures, often combined with a CRF decoding layer, confirming the effectiveness of BiLSTM-CRF models for Spanish clinical named entity recognition (Saluja et al., 2019). These architectures build on earlier work demonstrating the effectiveness of deep-learning-based sequence labeling for NER (Lample et al., 2016; Ma and Hovy, 2016), and have since become a strong baseline or state-of-the-art for clinical anonymization tasks.

More recently, pretrained language models based on the Transformer architecture have been applied to de-identification, including for Spanish

medical text. Multilingual models such as mBERT and XLM-R, as well as Spanish-specific models like BETO, have shown promising results on clinical and biomedical NLP tasks (Cañete et al., 2020; Conneau et al., 2020; Moreno-Barea et al., 2025; Subies et al., 2025).

Within the MEDDOCAN challenge, several systems reported improved performance through the incorporation of contextualized embeddings and Transformer-based architectures (Martínez et al., 2019), highlighting the value of pretrained components in low-resource clinical settings. More recently, approaches leveraging large pretrained language models (LLMs) have been showing promising performance (Staab et al., 2024; Liu et al., 2025). In particular, prompt-based approaches have emerged as an attractive alternative, offering minimal implementation effort and rapid deployment, an appealing option for clinical institutions without specialized machine learning expertise.

However, these approaches present practical and methodological challenges. LLM-based systems entail substantially higher computational costs, which may be difficult to accommodate in secure, on-premise clinical infrastructures. In addition, because pretrained models rely on large and partially opaque training corpora, it is not always possible to determine whether reported benchmark performance reflects robust generalization or incidental exposure to similar data during pretraining (Bommasani et al., 2021). Consequently, performance reported on widely used shared-task datasets may not directly translate to new institutional contexts.

More broadly, anonymization systems frequently experience performance degradation when applied across institutions, document types, or stylistic conventions (Stubbs et al., 2017). This issue is particularly pronounced in Spanish, where publicly available resources remain scarce and often originate from a single institutional source (González et al., 2019).

These limitations motivate the present work, which conducts a controlled comparative evaluation across heterogeneous datasets to provide more reliable and practitioner-relevant evidence for Spanish clinical text anonymization.

3. Datasets

The main purpose in determining the datasets to be used in this evaluation was to ensure fair evaluation of the different approaches, trying to avoid data leakage from public datasets to large pretrained models, and to provide a low-effort methodology to create datasets that can be tailored to the specific characteristics of particular deployment contexts.

Four datasets were used for this comparison of anonymization approaches: the publicly available MEDDOCAN corpus, which has been available since 2019, and thus most probably used for pre-training of large language models, the 2025 CARMEN-I corpus, which has most probably not been used for training of models, and additionally presents a big variety of texts, and finally two synthetic datasets.

Sensitive entities are annotated following the MEDDOCAN Protected Health Information (PHI) guidelines (**?**), using a set of 28 entity types covering personal identifiers, healthcare professionals, institutions, locations, dates, contact information, identifiers, and other sensitive attributes.

3.1. MEDDOCAN

The MEDDOCAN dataset was released in the context of the MEDDOCAN anonymization of medical texts shared task organized within IberLEF 2019 (**?**). It consists of 1,000 synthetic clinical case reports manually selected and curated by medical experts and health documentalists, and enriched with realistic instances of protected health information derived from discharge summaries and medical genetics records (**?**).

3.2. Carmen-I

The CARMEN-I corpus consists of anonymized clinical narratives derived from electronic health records collected at Hospital Clínic de Barcelona between March 2020 and March 2022(**??**). For this study, we worked only with the 1,697 documents in Spanish. It comprises radiology reports (IR, 961 documents), discharge summaries (IA, 573 documents), transfer reports (IT, 154 documents), exitus reports (IE, 5 documents), and clinical course notes (CC, 4 documents).

3.3. Synthetic datasets

We created two synthetic datasets that reproduce the properties of the MEDDOCAN corpus, to better assess the performance of pre-trained components by ruling out data leakage, that is, that the test data had been used for training, in a manner similar to Kim et al. (2024).

The SPG dataset was generated using the Synthetic Patient Generator (SPG), an open-source software module developed for this work and released under the MIT license[1]. The generator produces fully synthetic clinical documents in Spanish, including protected health information such as names, addresses, dates, phone numbers, and identifiers, reproducing the MEDDOCAN style.

[1] repository URL withheld for double-blind review

Anonymization annotations were produced automatically by the generator in BRAT and XML formats.

Generation of reports was based on random sampling functions from predefined lexical resources containing person names, institution names, professions, and contact information. While this approach ensures full control over the generated content, it results in limited linguistic variability, as it relies exclusively on random selection from static lists and does not incorporate paraphrasing or syntactic rewriting mechanisms.

To increase linguistic variability and better approximate real clinical narratives, the SPG dataset was extended using the Llama 3.1 large language model. Starting from the original SPG documents, clinical reports were expanded to include richer medical details and longer narratives, while preserving the original patient and administrative sections.

The generation process followed a carefully designed one-shot prompting strategy enforcing a strict document structure and prohibiting meta-comments or safety disclaimers. Despite these constraints, the raw outputs required extensive manual curation. Documents containing refusals, meta-text, missing sections, or invalid structure were removed or corrected. After this curation process, a total of 448 extended clinical reports were retained.

4. Evaluation Methods

4.1. Dataset Partitioning

Each dataset was divided into three subsets: training (*train*), development (*dev*), and evaluation (*test*). For MEDDOCAN, CARMEN-I and SPG, a 50/25/25 split was used, following common practice in de-identification benchmarks. Due to the smaller size of SPG Extended after curation, an 80/10/10 split was applied. Table 1 summarizes the partition sizes and proportions for each dataset.

Dataset	Train	Dev	Test
MEDDOCAN	500 (50%)	250 (25%)	250 (25%)
CARMEN-I	847 (50%)	425 (25%)	425 (25%)
SPG	500 (50%)	250 (25%)	250 (25%)
SPG Extended	358 (80%)	45 (10%)	45 (10%)

Table 1: Train, development, and test splits for each dataset, in number of documents.

For CARMEN-I, the split was performed using stratified sampling by document type to preserve the original distribution of report categories (e.g., radiology, discharge summaries) across partitions. This stratification reduces bias toward dominant document types and supports better generalization.

4.2. Evaluation Metrics

All approaches were evaluated against gold standard annotations at the character level. Accuracy and F1-scores were obtained by macro-averaging the score for each class, leaving out the class of non-entities (the "O" class), which is a majority and tends to skew results.

Confusion matrices were also analyzed, to obtain a more fine-grained analysis of systematic prediction errors.

5. Approaches to anonymization

5.1. Rule-Based Anonymization Using Domain Knowledge

As a representative symbolic approach, we evaluated the set of regular expressions developed within the ARPHAI initiative[2]. These rules encode domain knowledge for detecting sensitive information in Spanish clinical texts and do not require any training data. They were designed to process free-text records of primary care in the public health system of Argentinean province La Rioja.

The system comprises two types of patterns: (i) context-aware expressions that detect entities directly from textual cues and their surrounding context, such as dates, national identification numbers, passport numbers, email addresses, and license numbers; and (ii) dictionary-based patterns that rely on curated lists of known entities, such as street names and hospital names.

5.2. Recurrent Neural Networks: BiLSTM-CRF

As a classical machine learning approach, we adopt the BiLSTM-CRF model implemented by Saluja et al. Saluja et al. (2019) for the MEDDOCAN shared task. The architecture follows the end-to-end sequence labeling approach proposed by Ma and Hovy (2016), composed of three main layers: CNNs, BiLSTMs, and a CRF. First, a Convolutional Neural Network (CNN) is used to extract character-level representations from words, allowing the model to capture morphological information like prefixes and suffixes. These character vectors are concatenated with pre-trained word embeddings and fed into a Bi-directional Long Short-Term Memory (BiLSTM) network, which models contextual information by processing the sequence in both forward and backward directions. Finally, the output vectors from the BiLSTM are passed to a Conditional Random Field (CRF) layer on top to jointly decode the optimal label sequence.

[2]https://www.ciecti.org.ar/arphai/

We hypothesize that, in the clinical domain and specifically for EHRs, the word representation component plays a critical role. Medical notes are frequently filled with typographical errors, non-standard abbreviations, and variations in word forms that rule-based systems often struggle to capture. By extracting morphological information (such as prefixes and suffixes) rather than looking for exact whole-word matches, the CNN layer allows the model to identify medical terms even when they are misspelled or appear as variations not seen during training. This design alleviates vocabulary sparsity and improves robustness to the noisy and informal nature of clinical text, compared to models that depend solely on word-level embeddings.

The model is trained in a fully supervised manner with the training partition of each dataset, using the validation partition to select the best hyperparameters.

The BiLSTM and CRF layers in our sequence labeling model are initialized with random weights; however, we also experiment with incorporating pretrained word embeddings. The hypothesis guiding this comparison is that embeddings learned from very large, general-purpose corpora encode linguistic and semantic regularities that may not be captured when training solely on smaller annotated datasets, even if they are domain-specific. We used the following:

- **PlanTL** word embeddings pre-trained for Spanish biomedical text with the Scielo corpus (Soares et al., 2019)
- **Flair** word embeddings for Spanish NER for the Flair toolset (Akbik et al., 2018)
- **spaCy** word embeddings for Spanish NER for the spaCy toolset (Honnibal et al., 2020)

5.3. Off-the-shelf Named Entity Recognition

In addition to custom models, we evaluated a widely used Named Entity Recognition (NER) tool, spaCy (Honnibal et al., 2020). It is a production-oriented natural language processing library that provides neural network-based models for text annotation. Its NER component is also based in a BiLSTM architecture, but it builds its representations using features derived from token vectors produced by the tok2vec layer and learned sequence patterns.

We initialized all the components of the spaCy model randomly using the default configuration provided by the library, which resulted in a training of the parameters *from scratch*.

Using spaCy offers several practical advantages. First, it is an off-the-shelf framework with a well-tested architecture similar to the previous one, which simplifies implementation and experimentation. Second, the tok2vec layer generates contextual token representations via a trainable embedding and more complex encoding subnetworks, potentially yielding richer features. However, the experiment is justified as further evaluation is necessary, since neural architectures trained from scratch may not generalize as effectively to new domains without sufficient labeled data.

5.4. Large Language Models

We evaluated a large language model approach using Meta-Llama-3.1-8B-Instruct, a transformer-based model optimized for instruction following. Experiments were conducted on GPU infrastructure at the *Centro de Computación Avanzada of the Universidad Nacional de Córdoba*, using mixed-precision inference to reduce computational costs.

Entity recognition was performed via prompt-based annotation, instructing the model to generate in-line XML annotations directly within the clinical text, as shown in Figure 1. Multiple prompting strategies were explored, including zero-shot and one-shot variants. Zero-shot prompting yielded inconsistent and poorly formatted outputs, whereas one-shot prompts with carefully curated examples significantly improved annotation quality.

Given the model's context window limitation (8192 tokens, including input and output), the prompt and input text were dynamically adjusted using the official tokenizer to ensure complete generation. Approximately half of the available context was reserved for the output to prevent truncation.

The Llama-based approach does not require task-specific training or fine-tuning, relying instead on prompt engineering to adapt the pretrained model to the anonymization task.

6. Analysis of Results

Table 2 shows macro-averaged accuracy and F1 scores computed on the test partition of each dataset. As discussed, all metrics exclude the O (outside) class in order to better highlight differences in the detection and classification of named entities.

It should be noted that, during the spaCy experiments, a small proportion of entities could not be processed due to formatting inconsistencies in the data. As a result, the reported scores are not strictly comparable. In most cases, the number of affected entities was below 5% and distributed among different labels, except in the case of the MEDDOCAN dataset and for label S̈EXO_SUJETO_ASISTENCIA, where it reached almost 10%. This issue led to a noticeable decrease in macro-averaged accuracy for that dataset,

```
<document>
  <sentence id="1">
    <TAG TYPE="NOMBRE_SUJETO_ASISTENCIA">Juan Perez</TAG>
    fue atendido el
    <TAG TYPE="FECHAS">12/03/2023</TAG>
    remitido por <TAG TYPE="NOMBRE_PERSONAL_SANITARIO">Ignacio</TAG>
    <TAG TYPE="NOMBRE_PERSONAL_SANITARIO">Navarro Cuellar</TAG>
  </sentence>
</document>
```

Figure 1: Example of XML-formatted clinical text with annotated entities.

	SPG		Extended SPG		MEDDOCAN		CARMEN-I	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
REGEX	0,094	0,092	0,081	0,092	0,155	0,123	0,004	0,001
BiLSTM	**0,942**	**0,884**	0,382	0,591	**0,864**	0,694	**0,702**	0,535
BiLSTM-PlanTL	0,817	0,820	0,472	0,560	0,632	0,690	0,347	0,420
BiLSTM-Flair	0,900	0,781	0,363	0,507	0,847	0,679	0,665	0,335
BiLSTM-spaCy	0,877	0,775	0,360	0,503	0,782	0,623	0,619	0,225
spaCy	0,846*	**0,880***	**0,588**	**0,675**	0,693	**0,770**	0,561	**0,657**
Llama 3.1	0,710	0,712	0,425	**0,674**	0,393	0,455	0,091	0,098

Table 2: Macro average accuracy and F1 excluding the non-entity class, for different approaches to anonymization: regular expressions (REGEX); different flavors of recurrent neural networks: BiLSTM fully trained on the training portion of each corpus, including the embeddings layer, and with different pre-trained embeddings (BiLSTM-PlanTL, BiLSTM-Flair, BiLSTM-spaCy); off-the-shelf spaCy fully trained on the training portion of each corpus; and off-the-shelf Llama 3.1. with prompts. Evaluation on three synthetic datasets, the public MEDDOCAN and the newly synthesized SPG and Extended SPG, and the naturally occurring CARMEN-I. Results marked with * correspond to a subset of data

although the F1 score remained largely unaffected. Given the exploratory nature of this study, we consider the results sufficiently reliable to identify general performance trends and to provide practitioners with meaningful guidance when selecting an anonymization approach.

As a first approach to these results, Figure 3 shows the differences in our main metric of interest, macro F1. Learning-based approaches substantially outperform REGEX (first bar in each group) across all datasets. In fact, for the only natural dataset, with real life variability, CARMEN-I, REGEX performs very poorly, reflecting its limited coverage and inability to generalize beyond explicit patterns. This could be expected given that these rules have been based on the study of a specific dataset, but it is sobering to determine empirically the very little portability this kind of approach, and how it requires an important adaptation effort for each new domain.

It is also very interesting to see how a generic pretrained LLM (last bar per group) performs very poorly for the natural dataset. In this case, we have a very important remark to make: the only case where Llama 3.1 performed comparably to other approaches was for Extended SPG, the synthetic dataset that had been created... using Llama 3.1. But such good performance in this dataset cannot be extrapolated to good performance on other datasets, or, more importantly, on a natural dataset like CARMEN-I. Thus, in working with synthetic datasets to assess performance of different technological alternatives, one must be very careful with the circularities of using an LLM to analyze data produced by the same or similar LLM.

The most important conclusion from this study is that **recurrent neural network architectures perform better for the task of automatic anonymization of clinical text**. The **best approach is off-the-shelf spaCy** (second-to-last bar per group), probably because

- all the components have been trained from scratch specifically for the dataset, and
- the architecture constructs a more complex representation of words and strings than classic BiLSTM, with a rich representation component transforming different kinds of embeddings to provide the input for the main recurrent component.

Indeed, it can be seen that training the embedding layers in the training dataset improves performance with respect to using pre-trained components, even if these have been trained in much

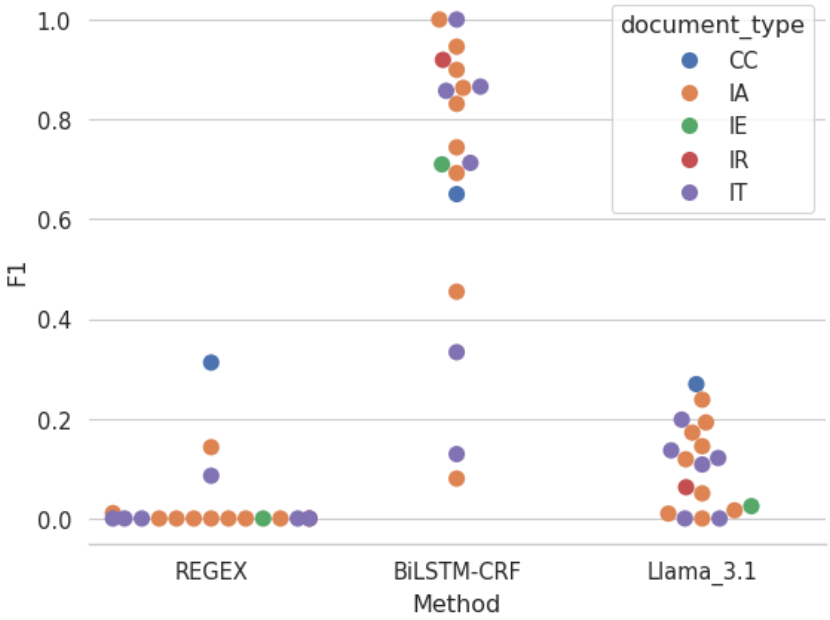

Figure 2: Distribution of performance of different anonymization approaches across different document types in the CARMEN-I datasets. Colors represent coarse types of documents, each dot represents a subtype of the coarser document types. The detail for each subtype of document will be provided in the camera-ready version of the paper, as an Appendix.

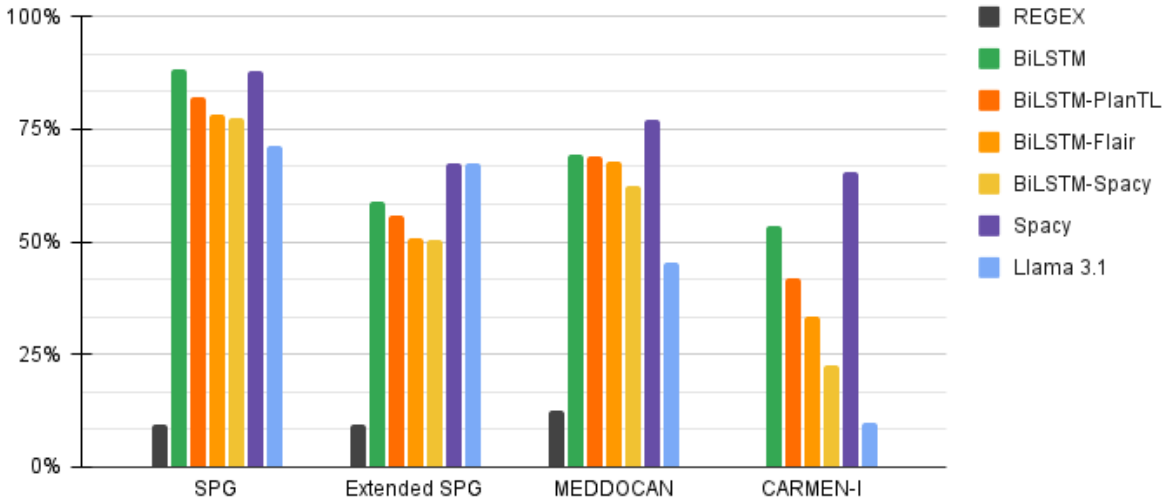

Figure 3: F1 macro of each approach on the different datasets, excluding the non-entity class. The detail can be seen in Table 2.

bigger corpora or a domain specific one. We can see that the BiLSTM that was fully trained on the datasets (second bar per group) performs consistently better than the BiLSTM approaches using pre-trained embeddings (3 middle bars per group). Even when using embeddings trained specifically in biomedical texts in Spanish, represented by the darkest orange bar never reach the performance of the fully trained BiLSTM. The only exception is the MEDDOCAN corpus, where performance of the different BiLSTM approaches is very similar. This can possibly be attributed to data leakage, since the MEDDOCAN corpus has been publicly available since 2019 and has been probably used for training these embeddings. More importantly, the drop in performance for approaches with pre-trained word embeddings is more acute for the CARMEN-I natural corpus, where the superiority of fully trained approaches can be clearly appreciated, followed by the relative better performance of the embeddings

pre-trained in biomedical domain.

Given the heterogeneity of the CARMEN-I corpus, we further analyzed performance by document type. Results in Figure 2 show that BiLSTM-CRF consistently outperforms REGEX and Llama 3.1 across most categories, although its performance is not uniform. As can be expected, the number of examples for a given document type strongly influences performance: indeed, the most notable degradation occurs for *IT_EXPLORACION_CLINICA*, a low-frequency document type with atypical date expressions, for which learned patterns do not generalize well.

7. Discussion and Conclusions

This study presents a controlled, cross-dataset comparison of representative anonymization paradigms for Spanish clinical text under uniform experimental conditions. Beyond reporting performance scores, our goal was to derive practically relevant insights for deployment in real clinical environments.

Across datasets, recurrent neural architectures, more concretely, CNN-BiLSTM-CRF, and in particular the spaCy implementation consistently outperformed both a rule-based system and a prompt-based approach with a large language model. The advantage was especially clear on heterogeneous, naturally occurring data such as that in the CARMEN-I dataset, which better reflects institutional variability.

From a deployment perspective, these models offer a favorable balance between effectiveness, computational cost, and infrastructural feasibility. They can be trained and executed on moderate hardware, including secure on-premise servers, whereas large language models require substantially greater computational resources and, if they are to be fine-tuned for a specific domain, skilled expertise. The prompt-based approach to a generic LLM, without fine-tuning, demonstrated limited robustness and weaker generalization to unseen natural data. Under realistic infrastructure constraints, moderately sized neural sequence labeling models therefore appear to be a more sustainable solution for Spanish clinical anonymization.

A consistent finding across datasets is that fully task-specific training of neural architectures yields stronger generalization than incorporating pretrained embeddings, even when those embeddings are derived from large biomedical corpora. This suggests that institutional specificity plays a central role: local documentation practices, abbreviations, and stylistic conventions are better captured through end-to-end training on representative annotated data. Importantly, effective anonymization does not require massive pretrained models. A well-curated, institution-specific dataset combined with a classical BiLSTM-CRF architecture can provide strong performance with lower operational complexity.

Methodologically, this work highlights the importance of contamination-aware and cross-dataset evaluation. The synthetic datasets introduced here were designed to replicate structural characteristics of clinical text without reproducing existing benchmarks, enabling more controlled comparisons. At the same time, results show that synthetic data must be handled carefully, as models may benefit when evaluated on data generated using related architectures. Complementary evaluation on naturally occurring corpora such as CARMEN-I is therefore essential. Performance differences between MEDDOCAN and CARMEN-I further demonstrate that single-benchmark evaluation is insufficient to guarantee real-world robustness, particularly in Spanish clinical NLP, where resources are scarce and institution-specific.

Taken together, these findings provide indicative but stable tendencies that can guide practitioners. While this study is limited to Spanish and to representative instances of each paradigm, it narrows the technological search space for institutions implementing anonymization pipelines under realistic constraints. In Spanish clinical contexts, investment in moderately sized, task-specific neural sequence labeling models appears to offer the strongest balance between effectiveness, efficiency, and deployability.

Beyond empirical comparisons, this work contributes reproducible evaluation procedures, synthetic benchmark resources, and cross-dataset evidence in an underrepresented language setting. By foregrounding methodological rigor and deployment feasibility, we aim to support more informed, evidence-based decisions in privacy-preserving clinical NLP.

Future work should extend this framework to additional languages and institutional contexts, explore domain-adapted or fine-tuned large language models, and investigate hybrid or lightweight architectures that further balance interpretability, robustness, and infrastructural constraints.

Ethics Statement

This work addresses the anonymization of clinical free-text records with the objective of enabling privacy-preserving secondary use of healthcare data. The study focuses on methodological evaluation and does not involve direct access to identifiable patient information by the authors.

Experiments were conducted using (i) publicly available benchmark datasets, (ii) the CARMEN-I corpus accessed under its Data Usage Agreement,

and (iii) synthetic datasets generated to replicate structural and statistical properties of clinical narratives without reproducing real patient content. The CARMEN-I resource consists of previously anonymized clinical documents distributed under controlled conditions via PhysioNet. All usage complied with the corresponding licensing and data access requirements.

The synthetic datasets were designed to mitigate potential benchmark contamination and reduce privacy risks associated with redistributing sensitive clinical material. Nevertheless, synthetic data cannot entirely eliminate the possibility of unintended memorization effects in large pretrained models. For this reason, we explicitly separate data generation and evaluation pipelines and complement synthetic evaluation with experiments on naturally occurring clinical data.

We acknowledge that anonymization systems are not infallible and may fail to detect certain categories of sensitive information. In real-world deployment, automated de-identification systems should therefore be integrated within broader governance frameworks that include human oversight, identification and charcterization of discriminatory behaviors including bias against subgroups of population, risk assessment, and compliance with applicable data protection regulations.

Finally, this work focuses exclusively on Spanish clinical text, a comparatively under-resourced setting in clinical NLP. By providing reproducible evaluation procedures and comparative evidence, we aim to support responsible technological decision-making in privacy-preserving healthcare applications (López et al., 2023).

Limitations

Several limitations must be acknowledged.

First, this study focuses exclusively on Spanish clinical text. Although this language-specific scope addresses the scarcity of annotated resources and comparative research in Spanish clinical anonymization, it limits the generalizability of our findings to other languages. Differences in documentation practices, annotation schemes, and linguistic features may affect performance, and further studies are needed to assess cross-linguistic robustness.

Second, although the synthetic datasets were carefully designed to resemble clinical narratives, synthetic data cannot fully replicate the linguistic variability, stylistic idiosyncrasies, and implicit conventions of naturally occurring documentation. Consequently, performance estimates obtained on synthetic corpora may differ from those observed in real-world deployment scenarios.

Third, the evaluation of large language models was restricted to a single representative LLM under prompt-based inference, without task-specific fine-tuning. Similarly, the off-the-shelf comparison was limited to spaCy as a representative general-purpose industrial NLP toolkit. While these choices reflect realistic scenarios faced by practitioners, who often evaluate a small number of accessible tools, they do not exhaust the full spectrum of possible architectures, fine-tuning strategies, or tool configurations. Alternative LLMs, domain-adapted transformers, or additional industrial frameworks may yield different absolute performance levels.

Finally, spaCy experiments were affected by preprocessing and formatting issues that prevented processing of a subset of documents. Although the overall performance trends remain stable across datasets and configurations, exact metric comparisons should therefore be interpreted with caution.

Acknowledgements

This study was partially supported by ARPHAI-CIECTI, a project funded by IDRC (Canada) and SIDA (Sweden). This work used computational resources from UNC Supercómputo (CCAD) – Universidad Nacional de Córdoba (https://supercomputo.unc.edu.ar), which are part of SNCAD, República Argentina.

Bibliographical References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pages 1638–1649. Association for Computational Linguistics.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *Proceedings of PML4DC at ICLR*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, et al. 2020. Unsupervised cross-lingual representation learning at scale. *Proceedings of ACL*.

Franck Dernoncourt, Ji Young Lee, Özlem Uzuner, and Peter Szolovits. 2017. De-identification of patient notes with recurrent neural networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*

(EMNLP), pages 199–204. Association for Computational Linguistics.

David A. Dorr, William F. Phillips, Shyam Phansalkar, Stephanie A. Sims, and John F. Hurdle. 2006. Assessing the difficulty and time cost of de-identification in clinical narratives. *Methods of Information in Medicine*, 45(3):246–252.

European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council.

Eulalia Farre Maduell, Salvador Lima-Lopez, Santiago Andres Frid, Artur Conesa, Elisa Asensio, Antonio Lopez-Rueda, Helena Arino, Elena Calvo, Maria Jesús Bertran, Maria Angeles Marcos, Montserrat Nofre Maiz, Laura Tañá Velasco, Antonia Marti, Ricardo Farreres, Xavier Pastor, Xavier Borrat Frigola, and Martin Krallinger. 2024. CARMEN-I: A resource of anonymized electronic health records in Spanish and Catalan for training and testing NLP tools. *PhysioNet*. Version 1.0.1.

Ana González, David Martínez, Sergio Montalvo, et al. 2019. Overview of the meddocan track: Medical document anonymization. *Proceedings of the IberLEF Workshop*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python. *Zenodo*.

Woojin Kim, Sungeun Hahm, and Jaejin Lee. 2024. Generalizing clinical de-identification models by privacy-safe data augmentation using GPT-4. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21204–21218, Miami, Florida, USA. Association for Computational Linguistics.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *Proceedings of NAACL-HLT*, pages 260–270.

Thomas Lavergne, Olivier Cappé, and François Yvon. 2010. Practical very large scale crfs. *Proceedings of the 48th Annual Meeting of the ACL*, pages 504–513.

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. 2025. Deid-gpt: Zero-shot medical text de-identification by gpt-4.

Sabrina López, Laura Alonso Alemany, Juan Manuel Dias, Laura Ación, and Verónica Xhardez. 2023. Guía práctica para la protección de datos personales en salud. Technical report, Fundar, Buenos Aires.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *Proceedings of ACL*, pages 1064–1074.

David Martínez, Ana González, and Sergio Montalvo. 2019. Meddocan: Medical document anonymization shared task. *IberLEF 2019 Working Notes*.

Stéphane M. Meystre, Jeff Friedlin, Brett R. South, Shengyu Shen, and Matthew H. Samore. 2010. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC Medical Research Methodology*, 10(1):70.

Francisco J. Moreno-Barea, Guillermo López-García, Héctor Mesa, Nuria Ribelles, Emilio Alba, José M. Jerez, and Francisco J. Veredas. 2025. Named entity recognition for de-identifying spanish electronic health records. *Computers in Biology and Medicine*, 185:109576.

I. Neamatullah, M. M. Douglass, L. H. Lehman, A. Reisner, M. Villarroel, W. J. Long, P. Szolovits, G. B. Moody, R. G. Mark, and G. D. Clifford. 2008. Automated de-identification of free-text medical records. *BMC Medical Informatics and Decision Making*, 8(1):32.

Binit Saluja, Ankit Kumar, Ankit Suri, and S Onur. 2019. Anonymization of clinical text in spanish using bilstm-crf. In *Proceedings of IberLEF*.

Felipe Soares, Marta Villegas, Aitor Gonzalez-Agirre, Martin Krallinger, and Jordi Armengol-Estapé. 2019. Medical word embeddings for spanish: Development and evaluation. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 124–133.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2024. Large language models are advanced anonymizers.

Amber Stubbs, Michele Filannino, and Özlem Uzuner. 2017. Practical applications for deep learning in clinical natural language processing. In *Proceedings of the BioNLP 2017 Workshop*, pages 25–34. Association for Computational Linguistics.

Amber Stubbs, Colton Kotfila, and Özlem Uzuner. 2015. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/uthealth shared task

track 1. *Journal of Biomedical Informatics*, 58(Suppl):S11–S19.

Guillem García Subies, Álvaro Barbero Jiménez, and Paloma Martínez Fernández. 2025. ClinteXtsp and rigoberta clinical: a new set of open resources for spanish clinical nlp.

U.S. Congress. 1996. Health insurance portability and accountability act of 1996. Public Law 104-191, 110 Stat. 1936.

Ozlem Uzuner, Ira Goldstein, Yuan Luo, and Isaac Kohane. 2007. Identifying patient information in medical records. *Journal of the American Medical Informatics Association*, 14(5):550–563.