Extracting Medical Image-Related Entities from Spanish Electronic Health Records using NER Methods

Alexander Platas[1,3,*], Marcos Merino Prado[1,*], Elena Zotova[1,*], Mikel Pérez de Mendiola[2], Montse Cuadros Oller[1], Karen López-Linares[1], María Gálvez[2], Cristina Barba[2], Antón Asla[2]

[1]Vicomtech Foundation, Basque Research and Technology Alliance (BRTA), Mikeletegi 57, 20009 Donostia-San Sebastián, Spain
{aplatas, mmerino, ezotova, mcuadros, klopez}@vicomtech.org

[2]Serikat - Consultoría y Servicios Tecnológicos, Ercilla 19, 48009 Bilbao, Spain

[3]Department of Languages and Computer Systems, University of the Basque Country (UPV-EHU), Paseo Manuel deLardizabal, 1, Donostia/San-Sebastián, 20018, Spain

[*]These authors contributed equally.

Abstract

This paper presents a novel corpus in Spanish tailored for the extraction of medical image-related entities from radiological reports using Named Entity Recognition (NER) methods. The dataset was created by aggregating and refining multiple existing corpora, focusing on entities that can be visually interpreted in associated medical images. This resource aims to bridge the gap between natural language processing and computer vision in the biomedical domain. The study evaluates various NER methods, including encoder-only, encoder-decoder, and decoder-only architectures. It explores fine-tuning, zero-shot, and few-shot In-Context Learning (ICL) strategies to determine the most effective approach for entity extraction. The resulting dataset is publicly available[1].

Keywords: Biomedical NER, LLM, clinical NLP, Spanish corpus

1. Introduction

This paper introduces a new Spanish medical dataset for Named Entity Recognition (NER), focused on entities that are visually discernible in radiological images. By aligning textual annotations with medically visual concepts, the dataset facilitates a range of multimodal applications, including image-text retrieval, feature alignment, and cross-modal interpretation, bridging the gap between medical NLP and computer vision.

We benchmark NER models of different architectures—encoder-only, encoder-decoder, and decoder-only—covering proprietary and open-source models of various scales, applying fine-tuning and In-Context Learning (ICL). Through experiments, we analyze entity extraction performance and the strengths and limitations of each approach. All technical details—per-class metrics, prompts, hyperparameters, and the dataset—are available on GitHub[1]. The main contributions are:

- A curated Spanish corpus annotated with image-related medical entities, supporting future research in medical language processing and multimodal approaches.
- A comprehensive benchmark of NER models, analyzing trade-offs between architecture, learning method, size and accuracy.

[1]https://github.com/Vicomtech/SMINER

The paper is structured as follows: Section 2 reviews related work on medical NER and language models. Section 3 details the dataset construction. Section 4 explains the experimental setup and methodologies. Section 5 presents and discusses results. Finally, Section 6 concludes the paper and outlines future directions.

2. Related Work

The recognition and classification of medical image-related entities are deeply intertwined with several broader research trajectories. More general research is associated with NER in electronic health records (EHRs). Early approaches to NER in EHRs in Spanish were implemented with Conditional Random Fields (CRF) (Santiso et al., 2017), neural networks, such as BiLSTM architecture (Santiso et al., 2021), or BiLSTM combined with CRF on top (Weegar et al., 2019). By 2020, almost all NER systems in Spanish shifted to BERT-based models fine-tuned for sequence tagging (Akhtyamova, 2020; Tamayo et al., 2022), adding more pre-trained variants of BERT, such as clinical and biomedical Spanish models (Báez et al., 2022; Villaplana et al., 2023; Goenaga et al., 2023), and adding multi-head classification (García-Pablos et al., 2020; Jonker et al., 2024).

Specific image-related NER is represented by entity detection in medical image reports. Datasets

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 10569–10578
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

such as the SpRadIE (513 anonymized radiology reports) provide essential resources for training and evaluating NER models in Spanish radiology reports (Cotik et al., 2021). NER approximations in Spanish radiology reports are also being led by transformer-based models. Hybrid systems combining BETO, a Spanish-specific BERT model, with dictionary-based approaches and cross-lingual word alignment were used by Suárez-Paniagua et al. (2021) (Suárez-Paniagua et al., 2021). Godoy et al. (2023) (Godoy et al., 2023) also uses an approach based on a transformer model, fine-tuned on a corpus, which consists of mammographic radiological reports annotated in laterality, location, and the finding. Distant metastasis annotated corpus based on computed tomography reports was created by Ahumada et al. (2024) (Ahumada et al., 2024). The authors used Bidirectional Long Short-Term Memory (BiLSTM) combined with CRF layers for detecting metastasis mentions.

Recent research focuses on generative models. For example, encoder-decoder multilingual T5 pre-trained on clinical and biomedical corpora (García-Ferrero et al., 2024) achieves state-of-the-art results in Spanish biomedical NER tasks. Early experiments with decoder-only LLMs for Spanish NER show competitive results (García-Barragán et al., 2024), though further research is needed.

3. Dataset Construction

To fulfill the objectives of this study, the corpus required token-level annotations for entities relevant to medical imaging research. To construct the dataset, we assessed the use of six distinct datasets from the Barcelona Supercomputing Center (BSC), each comprising the same set of 1,000 Spanish-language documents manually annotated in BRAT format for different topics. Each dataset was analyzed to identify relevant annotations for this study, with the goal of selecting and filtering only those entities that can be perceived in medical images (e.g., tumor), while excluding entities that are not visually observable (e.g., diabetes).

- **MEDDOCAN** (Marimon et al., 2020): includes 22,795 annotations related to demographic information about the patient, healthcare center, and consultation, such as patient name, age, sex, address, and dates. Only the Age and Sex classes were retained, as they can be reasonably inferred from medical images. During inspection, annotation errors were detected and manually corrected. After curation, the final dataset comprised 1,963 entities. Unlike the other datasets, these documents contained a header preceding the report; it was removed and entity offsets were adjusted to preserve alignment with the text.
- **MedProcNER** (López et al., 2023): comprises 14,684 entities related to medical procedures. To address this task, MediPhi, a LLM specialized in the medical domain, was employed. Each entity, along with its textual context, was provided to the model through specific prompts. MediPhi classified each entity into one of three categories based on whether it referred to a medical imaging procedure: Positive, Negative, or Ambiguous. Subsequently, Gemini 2.5 Pro was used to further classify the Ambiguous cases and to provide justifications for each decision. Finally, a manual review of the classifications and explanations from model was conducted to ensure their accuracy and consistency. This filtering process resulted in a final set of 3,645 annotations. This pipeline is represented in Figure 1.

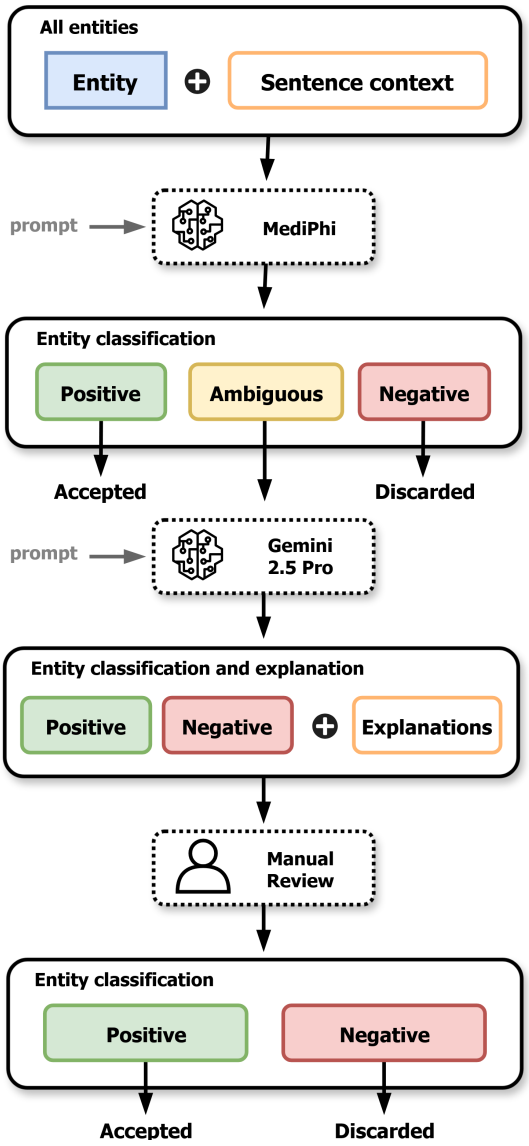

Figure 1: Dataset filtering pipeline, for Disease and Procedure entities related to medical imaging.

- **DisTEMIST** (Miranda-Escalada et al., 2022): a dataset containing 10,663 entities referring to various diseases and conditions. To exclude those not observable through medical imaging procedures, the same filtering approach applied to MedProcNER was employed, which resulted in a final set of 6,862 entities.
- **PharmaCoNER** (Gonzalez-Agirre et al., 2020), **CodiEsp** (Miranda-Escalada et al., 2020) and **SympTEMIST** (López et al., 2024) provide annotations for medications and chemical compounds, diseases, procedures and diagnoses, and symptoms, respectively. Although based on the same clinical documents as the previously mentioned corpora, they were excluded from this study because their annotations are either not related to the image or overlap with entities in the other datasets.

As a result of these processes, we constructed a dataset consisting of 1,000 Spanish medical documents divided into training (70%), validation (15%), and test (15%) subsets. The dataset contains only textual radiology reports (no images), annotated with clinical entities corresponding to imaging findings. Table 1 summarizes the main statistics of the corpus, including the number of tokens and annotated entities across each split.

Split	Reports	Tokens	Entities
Train	701	396.14 ± 182.99	37.72 ± 25.58
Dev	150	409.13 ± 207.27	39.53 ± 27.77
Test	149	386.72 ± 195.28	35.49 ± 23.81
Full	1,000	396.69 ± 188.57	37.66 ± 25.67

Table 1: Average number of tokens, entities, and documents per split, with their standard deviation.

The dataset includes token-level annotations for four distinct classes: Age (*Edad*), Sex (*Sexo*), Disease (*Enfermedad*), and Procedure (*Procedimiento*). The split was stratified to ensure that the class distribution remained balanced and representative across all subsets. The overall distribution of these classes, as well as the number of annotated entities within each split, are summarized and presented in Table 2.

Class	Number of entities			
	Train	Dev	Test	Total
Age	764	172	164	1,100
Sex	609	122	132	863
Disease	4,811	1,097	954	6,862
Procedure	2,529	581	535	3,645
Total	**8,713**	**1,972**	**1,785**	**12,470**

Table 2: Class distribution of the final dataset.

To facilitate the implementation of various methods and models, the dataset is publicly available[2] in both *BRAT* and *BIO* formats.

4. Experiments

The experimental evaluation was divided into three distinct approaches based on the architecture of models employed: encoder-only, encoder-decoder, and decoder-only. The following sections detail the methodologies used for each approach.

4.1. Encoder-Only models

Encoder-only models are a well-established choice for various tasks involving medical texts and were therefore selected as the baseline for the experiments. To capture a representative range of model characteristics among the many existing ones, three architectures were fine-tuned:

- Multilingual BERT (Devlin et al., 2018)
- Biomedical RoBERTa (Carrino et al., 2021)
- EriBERTa (de la Iglesia et al., 2023)

Additionally, three GLiNER models were evaluated in a zero-shot setting to assess their out-of-the-box performance on this task:

- GLiNER-X (Stepanov and Shtopko, 2024)
- GLiNER-BioMed large (Yazdani et al., 2025)
- NuNER-span (Bogdanov et al., 2024)

Table 3 summarizes their most relevant architectural and training characteristics.

As seen in Table 1, some documents in the dataset might exceed the token limit of 512 imposed by these models. In order to overcome the input size limitation present in these models, a sliding window was implemented. This technique is applied prior to inputting the text into the model and segments it into chunks that match the model's maximum input capacity. The chunks are then fed sequentially to the model. To minimize the loss of context during this process, a degree of overlap equal to half the token limit for each architecture is maintained between adjacent chunks.

During training, various hyperparameter combinations were evaluated to identify the optimal model configurations. Specifically, different values for learning rate, batch size, and weight decay were tested. In all experiments, the number of training epochs was fixed at 20. Details of the hyperparameters used in the fine-tuning of the best-performing model are specified in the supplementary materials hosted in our GitHub repository[2].

[2] https://github.com/Vicomtech/SMINER

Model	Base model	Medical domain	Languages	Usage
Multilingual BERT	BERT	✗	Spanish + 101 more	Fine-tuning
Biomedical RoBERTa	RoBERTa	✓	Primarily Spanish	Fine-tuning
EriBERTa	RoBERTa	✓	Spanish + English	Fine-tuning
GLiNER-BioMed-large	GLiNER	✓	Primarily English	Zero-shot
NuNER-span	GLiNER	✗	Primarily English	Zero-shot
GLiNER-X-large	GLiNER	✗	Spanish + 19 more	Zero-shot

Table 3: Encoder-only models used and their distinct characteristics.

4.2. Encoder-Decoder Models (T5)

We experiment with Medical mT5-large (García-Ferrero et al., 2024), as it has shown high performance in medical tasks in Spanish. Fine-tuning of the T5 model requires data preprocessing to match its text-to-text paradigm. We format the corpus into input-output pairs, where the input is a structured prompt and the output is an expected response containing image-related entities. We also split data into sentences to make the pairs fit into the maximum sequence length of the model. The resulting size is shown in Table 4.

Splits	Train	Dev	Test	**Total**
Sentences	10,716	2,532	2,454	**15,702**

Table 4: Dataset split and number of sentences in each part, for Medical-T5 training.

The prompt used consists of a prefix, which is an instruction to extract named entities, followed by a colon, and a sentence, as shown in Example 1. In all examples, the Spanish original is presented alongside its English translation in gray. For the output, we design two formats: (a) a list of extracted spans with their entity labels separated by a semicolon, shown in Example 2; (b) a tagged sentence, where tags are the defined named entities, given in Example 3. The sequences in the List format of are notably shorter than those in the Tag format, as they include only the entities.

Example 1. ***Input***

Extract named entities: Varón de 38 años, con antecedentes de enfermedad de Crohn e ingresado en dos ocasiones por episodios de obstrucción intestinal.

Extract named entities: 38-year-old male, with a history of Crohn's disease and hospitalized on two occasions for episodes of intestinal obstruction.

Example 2. ***Output Medical-mT5-large + List***

Varón | SEXO; 38 años | EDAD; enfermedad de Crohn | ENFERMEDAD; obstrucción intestinal | ENFERMEDAD

Male | SEX; 38 years | AGE; Crohn's disease | DISEASE; intestinal obstruction | DISEASE

Example 3. ***Output Medical-mT5-large + Tag***

<SEXO>Varón</SEXO> de <EDAD>38 años</EDAD>, con antecedentes de <ENFERMEDAD> enfermedad de Crohn</ENFERMEDAD> e ingresado en dos ocasiones por episodios de <ENFERMEDAD>obstrucción intestinal</ENFERMEDAD>.

<SEX>Male</SEX> of <AGE>38 years</AGE>, with a history of <DISEASE>Crohn's disease</DISEASE> and hospitalized on two occasions for episodes of <DISEASE>intestinal obstruction</DISEASE>.

Some sentences are not annotated with any label. In this case, the List model generates the phrase "NO ENTITIES" and the Tag model generates just a copy of the input sentence.

4.3. Decoder-Only Models (LLMs)

For the experiments with Large Language Models (LLMs), we conducted tests using In-Context Learning (ICL), followed by fine-tuning a smaller model. We selected commercial models from different providers and open-source models of varying sizes. The open-source models included both small models (~4B) and large models (~30B). Notably, two of the three open-source models were specifically pre-trained on medical data to better handle domain-specific terminology and entities. Regarding commercial models, we selected three state-of-the-art models recognized for achieving top performance on a range of natural language understanding and generation benchmarks. The full list of models used is shown in Table 5.

For zero-shot experiments, we directly used the test set reports along with task instructions. In the few-shot setting, we incorporated 5 examples from the training set into the prompt. Due to the context length limitations of LLMs, we split texts into sentences of approximately 100 tokens and later merged the predictions.

During evaluation, we frequently encountered cases where the models altered the text tokenization, resulting in an excess or shortage of tags that complicated the evaluation. This mismatch prevented proper comparison, as the number of predicted and ground-truth labels must match for each text. To address this issue, we detokenized the input text and then re-tokenized the model output to align with the original tokenization, as shown in Figure 2. We employed the same tagging format as *Medical-mT5-large* defined in Example 3, where each entity is enclosed within XML-like tags.

Large Language Model	Size	Open source	Context window	Max. output tokens	Release date
MediPhi (Corbeil et al., 2025)	3.8B	✓	128k	16k	May, 2025
Qwen 3 (Qwen Team, 2025)	4B	✓	128k	32k	May, 2025
Qwen 3 (Qwen Team, 2025)	30B	✓	128k	32k	May, 2025
MedGemma (Sellergren et al., 2025)	4B	✓	128k	8k	May, 2025
MedGemma (Sellergren et al., 2025)	27B	✓	128k	8k	May, 2025
Gemini 2.5 Pro (Gemini Team, 2025)	-	✗	1M	64k	Jun, 2025
GPT-5 (OpenAI, 2025)	-	✗	400k	128k	Aug, 2025
Claude 4.5 Sonnet (Anthropic, 2025)	-	✗	200k	64k	Sep, 2025

Table 5: Evaluated LLMs. "-" indicates unknown model size.

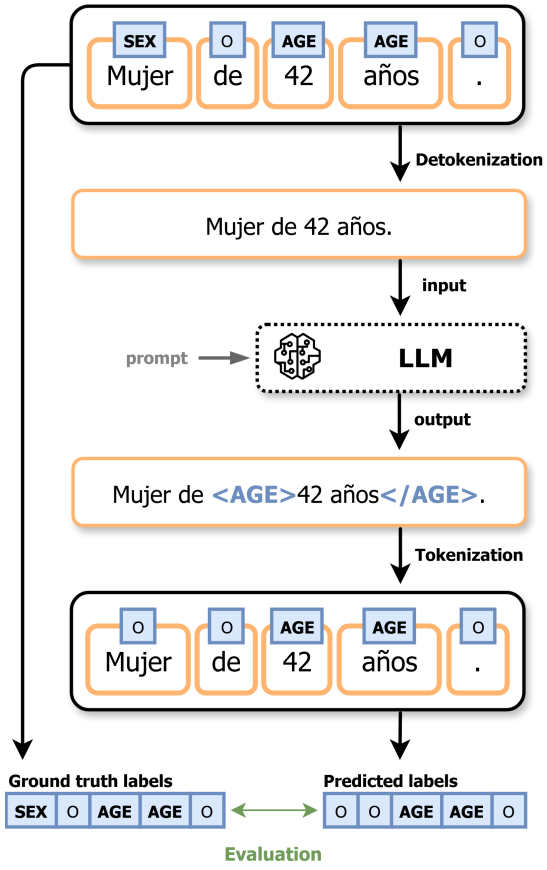

Figure 2: LLM-based NER evaluation pipeline.

The models were tested with a temperature of 0 to ensure deterministic outputs, except for GPT-5, which only allows a minimum temperature of 1. After some trials, we found that using a Spanish prompt (aligned with the dataset language) led to better results.

For supervised fine-tuning, we used the best of the smallest models (Qwen 3 4B). Fine-tuning is performed in 16-bit precision for 5 epochs. Details of the hyperparameters used in the fine-tuning are specified in in our GitHub repository[3]. After training, the model is evaluated in a zero-shot setting.

[3] https://github.com/Vicomtech/SMINER

5. Results

To ensure a fair evaluation across all models, a standardized evaluation script was developed. This script utilizes token-level annotations in the IOB2 format to compute overall accuracy, precision, recall, and F1-score, as well as class-specific metrics. Tables 6, 7 and 8 present a summary of the overall best results achieved by each model.

Language Model	Zero-shot		
Decoder-only	P	R	F1
MediPhi	37.89	11.71	17.88
MedGemma 4B	**64.73**	9.40	16.41
Qwen 3 4B	32.33	15.08	20.57
MedGemma 27B	45.87	16.54	24.32
Qwen 3 30B	32.19	14.24	19.74
Gemini 2.5 Pro	49.08	**64.60**	55.78
Claude 4.5 Sonnet	63.13	49.52	55.50
GPT-5	57.03	55.03	**56.01**
Encoder-only	P	R	F1
NuNER	**66.98**	19.86	30.64
GLiNER-BioMed	49.51	31.06	38.17
GLiNER-X	47.41	**32.41**	**38.50**

Table 6: Precision (P), Recall (R), and F-Score (F1) in the zero-shot scenario. Underlined scores indicate the best overall result; **bold** denotes the best model within the group.

Among the LLMs in the zero-shot setting, models with a larger number of parameters tend to achieve superior results. GPT-5 attains the highest F1-score, offering a balanced trade-off between precision and recall, while Gemini and Claude perform slightly worse overall. In contrast, MediPhi, MedGemma, and Qwen show lower recall, suggesting limited generalization in zero-shot settings. Among encoder-only models, despite their smaller size compared to LLMs, they achieve competitive results, outperforming models ranging from 4B to 30B parameters. These results highlight that the largest LLMs lead in zero-shot biomedical NER.

Language Model	Few-shot		
Decoder-only	P	R	F1
MediPhi	27.96	19.41	22.92
MedGemma 4B	33.86	20.60	25.61
Qwen 3 4B	25.50	29.37	27.30
MedGemma 27B	43.60	43.73	43.66
Qwen 3 30B	33.77	36.07	34.88
Gemini 2.5 Pro	51.04	**64.60**	57.03
Claude 4.5 Sonnet	**64.95**	57.46	**60.97**
GPT-5	59.49	55.54	57.45

Table 7: Precision (P), Recall (R), and F-Score (F1) in the zero-shot scenario. **Underlined** scores indicate the best overall result;

Few-shot ICL approaches yield more substantial performance gains in smaller models than zero-shot approach, whereas larger LLMs show only marginal improvements. Interestingly, Claude 4.5 Sonnet surpasses all others in few-shot mode, outperforming GPT-5. Nonetheless, even with few-shot prompting, smaller LLMs continue to underperform compared to larger LLMs in zero-shot settings. Overall, the performance of all LLMs remains considerably lower than that of encoder-only and encoder-decoder models, indicating that ICL alone remains insufficient for domain-specific biomedical entity recognition.

Language Model	Fine-tuning		
Decoder-only	P	R	F1
Qwen 3 4B	**68.16**	**36.75**	**47.75**
Encoder-only	P	R	F1
mBERT	65.15	63.25	64.18
BioRoBERTa	**68.21**	**71.24**	**69.69**
EriBERTa	65.99	70.43	68.14
Encoder-Decoder	P	R	F1
Medical mT5 + List	81.97	**82.96**	**82.40**
Medical mT5 + Tag	**81.98**	81.87	81.90

Table 8: Precision (P), Recall (R), and F-Score (F1) in the fine-tuning scenario. Underlined scores indicate the best overall result; **bold** denotes the best model within the group.

Fine-tuned encoder-only models demonstrate a clear advantage over all zero- and few-shot models. BioRoBERTa achieves the highest F1, showing that domain-specific pretraining combined with fine-tuning remains highly effective for biomedical tasks. EriBERTa follows closely, confirming the robustness of fine-tuned transformer encoders for structured extraction. Within decoder-only architectures, fine-tuning also proves beneficial: Qwen 3 4B improves its F1 by more than 20 points and trails the best-performing LLM evaluated in a zero-shot setting (GPT-5) by only 10 points.

These results highlight that supervised domain adaptation can significantly strengthen smaller generative models. Even with more limited architectural capacity, fine-tuned decoder models can narrow the gap with larger LLMs and become competitive for structured information extraction.

Finally, encoder-decoder architectures dominate all settings. The Medical mT5 + List model achieves the best scores across metrics, with the Medical mT5 + Tag model close behind. These results highlight that task-specific supervision combined with architecture explicitly suited for structured sequence labeling yields the highest performance. This suggests that models designed to both understand the input and generate structured outputs are particularly well suited for this task, outperforming encoder-only and decoder-only alternatives.

5.1. Discussion

1. **Encoder-decoder architectures obtain the best results.** The top-performing model was Medical-mT5 + List, likely due to the shorter target sequences, since this format outputs only the extracted entities while omitting the remaining tokens, unlike the + Tag format which assigns a label to every token. However, encoder-only models achieved strong results at a much lower computational cost.

2. **Few-shot ICL significantly improves performance.** Few-shot and zero-shot experiments were conducted, and the few-shot approaches consistently outperformed the zero-shot ones, particularly in models ranging from 4B to 30B parameters. Table 9 summarizes the average F1 score improvement achieved by the LLMs, divided by model size.

Model	0-shot	5-shot
MediPhi 4B	17.88	22.92 +5.04
MedGemma 4B	16.41	25.61 +9.20
Qwen 3 4B	**20.57**	**27.30** +6.73
MedGemma 27B	**24.32**	**43.66** +19.34
Qwen 3 30B	19.74	34.88 +15.14
Gemini 2.5 Pro	55.78	57.03 +1.25
Claude 4.5 Sonnet	55.50	**60.97** +4.96
GPT-5	**56.01**	57.45 +1.44
Mean	35.75	43.60 +7.85

Table 9: Comparison of F1 scores between zero-shot and few-shot across different models. Underlined score means the best overall result, **bold** font stands for the best model inside the group, and green indicates the difference in F-score.

3. **Fine-tuning yields substantial gains.** The fine-tuned Qwen 3 4B achieved an improvement of 20 F1 points over its base version, surpassing even the 30B model in the few-shot setting, as shown in Table 10. This demonstrates that targeted fine-tuning can effectively bridge or even overcome size-related performance gaps, enhancing domain adaptation and entity recognition accuracy.

Model	Size	Scenario	F1 score
Qwen 3	4B	Zero-shot	20.57
		Few-shot	27.30
		Fine-tuned	**47.75**
Qwen 3	30B	Zero-shot	19.74
		Few-shot	**34.88**

Table 10: Comparison of F-Scores between zero-shot, few-shot and fine-tuned Qwen3 models. Underlined score means the best overall result, **bold** font stands for the best model inside the group.

4. **Domain specificity has a significant impact on encoder-only models.** The best-performing model in this category is BioRoBERTa, which has been specifically trained on clinical texts. It is closely followed by EriBERTa, another model specialized in the medical domain. The considerably larger gap between these two models and the next best one, mBERT, underscores the importance of domain-specific training corpora in achieving superior performance.

5. **Model scale plays a crucial role in the performance of LLMs.** In this case, larger proprietary models achieve the best results, surpassing smaller, domain-specific medical models. The increased parameter number and broader training data of these large-scale LLMs provide them with stronger generalization capabilities, allowing them to outperform specialized models despite lacking explicit medical domain adaptation.

6. **Model performance highly varies from class to class.** As seen in Table 11, the Disease and Procedure classes exhibit the lowest performance across all models, with Disease being particularly challenging to identify. This difficulty likely stems from the high variability and need for highly specific medical knowledge to correctly interpret them. In contrast, the Age and Sex classes are consistently recognized with high scores across all models, as their expressions in text are highly standardized and do not require specialized medical understanding.

7. **Hallucinations remain a persistent challenge in LLMs.** Although the tokenization issue was resolved, smaller models still showed hallucination risks, introducing tags that were neither requested in the prompt nor present in the examples. Example 4 (original in Spanish, with English translation in gray) illustrates a case where the model hallucinates by generating previously unseen tags. This behavior suggests a tendency of smaller architectures to overgeneralize, which introduces uncertainty and limits their reliability in precision-critical tasks.

Example 4. ***Hallucination case in Qwen3 4B***

Paciente de `<EDAD>` *41 años* `</EDAD>` *,* `<FUMADORA>` *fumadora* `</FUMADORA>` *y diagnosticada de* `<ENFERMEDAD>` *lipoma en muslo derecho* `</ENFERMEDAD>`*.*

Patient aged `<AGE>` *41 years* `</AGE>`*,* `<SMOKER>` *smoker* `</SMOKER>` *and diagnosed with* `<DISEASE>` *lipoma in the right thigh* `</DISEASE>`*.*

6. Conclusions and Future Work

This work introduces a Spanish NER corpus, curated from multiple datasets extracting visually grounded medical entities. This resource fills a gap in domain-specific benchmarks for Spanish and enables rigorous evaluation of NER systems in clinical imaging contexts.

Results highlight the importance of having training data, with fine-tuning and few-shot approaches (both data-dependent) achieving the best performance. This is particularly critical in medical domain, where domain-specific data ensures that models capture the precise terminology, and nuances required for reliable predictions. Without relevant domain data, even large models (including proprietary ones) risk producing inaccurate outputs, yielding significantly worse results than when domain-specific data is available.

Future work will focus on multimodal integration and hallucination mitigation strategies in generative models, while continuing to emphasize the collection and utilization of high-quality, domain-specific datasets.

7. Acknowledgements

This work has been partially funded by the Basque Government under the HAZITEK 2024 Programme (grant number ZE-2024/00030) through the IRUDIA project. The authors acknowledge the essential contributions of the project's partners.

Average F-Score per class				
Small LLMs (< 4B)	Sex	Age	Disease	Procedure
Zero-shot	63.81	63.82	2.78	4.20
Few-shot	68.84	80.37	10.94	20.93
Fine-tuned	74.88	69.66	34.17	57.53
Medium LLMs (27B / 30B)	Sex	Age	Disease	Procedure
Zero-shot	66.16	57.33	9.98	10.28
Few-shot	85.43	84.24	23.74	41.31
Large LLMs (proprietary)	Sex	Age	Disease	Procedure
Zero-shot	82.16	90.58	41.33	63.54
Few-shot	**92.80**	93.38	43.25	69.75
BERT-based models	Sex	Age	Disease	Procedure
Fine-tuned	92.20	**96.15**	56.34	75.18
GLiNER-based models	Sex	Age	Disease	Procedure
Zero-shot	91.66	89.90	15.07	28.88
mT5-based models	Sex	Age	Disease	Procedure
Fine-tuned	92.31	91.58	**67.69**	**77.03**

Table 11: Average F-Score per class across different model types and training configurations. **Underlined** denotes the best scores.

8. Bibliographical References

Ricardo Ahumada, Jocelyn Dunstan, Matías Rojas, Sergio Peñafiel, Inti Paredes, and Pablo Báez. 2024. Automatic detection of distant metastasis mentions in radiology reports in spanish. *JCO Clinical Cancer Informatics*, 8:e2300130.

Liliya Akhtyamova. 2020. Named Entity Recognition in Spanish Biomedical Literature: Short Review and BERT Model. In *2020 26th Conference of Open Innovations Association (FRUCT)*, pages 1–7.

Anthropic. 2025. System card: Claude sonnet 4.5.

Pablo Báez, Felipe Bravo-Marquez, Jocelyn Dunstan, Matías Rojas, and Fabián Villena. 2022. Automatic extraction of nested entities in clinical referrals in spanish. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(3):1–22.

Sergei Bogdanov, Alexandre Constantin, Timothée Bernard, Benoit Crabbé, and Etienne Bernard. 2024. Nuner: Entity recognition encoder pre-training via llm-annotated data.

Casimiro Pio Carrino, Jordi Armengol-Estapé, Asier Gutiérrez-Fandiño, Joan Llop-Palao, Marc Pàmies, Aitor Gonzalez-Agirre, and Marta Villegas. 2021. Biomedical and clinical language models for spanish: On the benefits of domain-specific pretraining in a mid-resource scenario. *arXiv preprint arXiv:2109.03570*.

Jean-Philippe Corbeil, Amin Dada, Jean-Michel Attendu, Asma Ben Abacha, Alessandro Sordoni, Lucas Caccia, Francois Beaulieu, Thomas Lin, Jens Kleesiek, and Paul Vozila. 2025. A modular approach for clinical slms driven by synthetic data with pre-instruction tuning, model merging, and clinical-tasks alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 19352–19374. Association for Computational Linguistics.

Viviana Cotik, Laura Alonso Alemany, Darío Filippo, Franco Luque, Roland Roller, Jorge Vivaldi, Ammer Ayach, Fernando Carranza, Lucas Defrancesca, Antonella Dellanzo, and Macarena Fernández Urquiza. 2021. Overview of CLEF eHealth Task 1 - SpRadIE: A Challenge on Information Extraction from Spanish Radiology Reports. In *CLEF 2021 – Conference and Labs of the Evaluation Forum*, CEUR Workshop Proceedings, Bucharest, Romania. CEUR-WS.org. CLEF eHealth Lab 2021.

Iker de la Iglesia, Aitziber Atutxa, Koldo Gojenola, and Ander Barrena. 2023. Eriberta: A bilingual pre-trained language model for clinical natural language processing. *arXiv preprint arXiv:2306.07373*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Álvaro García-Barragán, Alberto González Calatayud, Oswaldo Solarte-Pabón, Mariano

Provencio, Ernestina Menasalvas, and Víctor Robles. 2024. Gpt for medical entity recognition in spanish. *Multimedia Tools and Applications*, pages 1–20.

Iker García-Ferrero, Rodrigo Agerri, Aitziber Atutxa Salazar, Elena Cabrio, Iker de la Iglesia, Alberto Lavelli, Bernardo Magnini, Benjamin Molinet, Johana Ramirez-Romero, German Rigau, Jose Maria Villa-Gonzalez, Serena Villata, and Andrea Zaninello. 2024. MedMT5: An open-source multilingual text-to-text LLM for the medical domain. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11165–11177, Torino, Italia. ELRA and ICCL.

Aitor García-Pablos, Naiara Pérez, and Montse Cuadros. 2020. Vicomtech at CANTEMIST 2020. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020)*, pages 489–498.

Google Gemini Team. 2025. Gemini 2.5 pro.

Eduardo Godoy, Steren Chabert, Marvin Querales, Julio Sotelo, Denis Parra, Carlos Fernández, Diego Mellado, Alejandro Veloz, Scarlett Lever, Favian Pardo, Ayleen Bertini Rojas, Yomar Molina, Claudia Díaz, Rodrigo Ferreira, and Rodrigo Salas. 2023. A named entity recognition framework using transformers to identify relevant clinical findings from mammographic radiological reports. In *2024 14th International Conference on Pattern Recognition Systems (ICPRS)*, page 46.

Iakes Goenaga, Edgar Andres, Koldo Gojenola, and Aitziber Atutxa. 2023. Advances in monolingual and crosslingual automatic disability annotation in spanish. *BMC Bioinformatics*, 24(1):265.

Aitor Gonzalez-Agirre, Antonio Miranda-Escalada, Obdulia Rabal, and Martin Krallinger. 2020. Pharmaconer corpus: gold standard annotations of pharmacological substances, compounds and proteins in spanish clinical case reports. *Zenodo*.

Richard Jonker, Tiago Almeida, Rui Antunes, João Almeida, and Sérgio Matos. 2024. Multi-head CRF classifier for biomedical multi-class named entity recognition on Spanish clinical notes. *Database*, 2024:baae068.

Salvador Lima López, Eulàlia Farré Maduell, Luis Gascó Sánchez, and Martin Krallinger. 2023. Medprocner corpus: Gold standard annotations for clinical procedures information extraction. *Zenodo*.

Salvador Lima López, Luis Gascó Sánchez, Eulalia Farré, Laura Vigil Gimenez, and Martin Krallinger. 2024. Symptemist corpus: Gold standard annotations for clinical symptoms, signs and findings information extraction. *Zenodo*.

Montserrat Marimon, Aitor Gonzalez-Agirre, Ander Intxaurrondo, Heidy Rodríguez, Jose Antonio Lopez Martin, Marta Villegas, and Martin Krallinger. 2020. Meddocan corpus: gold standard annotations for medical document anonymization on spanish clinical case reports. *Zenodo*.

Antonio Miranda-Escalada, Eulàlia Farré, Luis Gasco, Salvador Lima, and Martin Krallinger. 2022. Distemist corpus: detection and normalization of disease mentions in spanish clinical cases. *Zenodo*.

Antonio Miranda-Escalada, Aitor Gonzalez-Agirre, and Martin Krallinger. 2020. Codiesp corpus: gold standard spanish clinical cases coded in icd10 (cie10) - ehealth clef2020. *Zenodo*.

OpenAI. 2025. Gpt-5 is here.

Alibaba Qwen Team. 2025. Qwen3 technical report.

Sara Santiso, Alicia Pérez, and Arantza Casillas. 2021. Adverse drug reaction extraction: Tolerance to entity recognition errors and sub-domain variants. *Computer Methods and Programs in Biomedicine*, 199:105891.

Sergio Santiso, Arantza Casillas, Alicia Pérez, and Maite Oronoz. 2017. Medical entity recognition and negation extraction: Assessment of negex on health records in spanish. In *Bioinformatics and Biomedical Engineering - 5th International Work-Conference, IWBBIO 2017, Proceedings, Part I*, pages 177–188.

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

Ihor Stepanov and Mykhailo Shtopko. 2024. Gliner multi-task: Generalist lightweight model for various information extraction tasks.

Víctor Suárez-Paniagua, Hang Dong, and Arlene Casey. 2021. A multi-BERT hybrid system for Named Entity Recognition in Spanish radiology

reports. In *2021 Working Notes of CLEF: Conference and Labs of the Evaluation Forum, CLEF-WN 2021*, pages 846–856.

Antonio Tamayo, Diego Burgos, and Alexander Gelbukh. 2022. Partner: Paragraph tuning for named entity recognition on clinical cases in spanish using mbert+ rules. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2022), A Coruña, Spain*.

Aitana Villaplana, Raquel Martínez, and Soto Montalvo. 2023. Improving medical entity recognition in spanish by means of biomedical language models. *Electronics*, 12(23).

Rebecka Weegar, Alicia Pérez, Arantza Casillas, and Maite Oronoz. 2019. Recent advances in swedish and spanish medical entity recognition in clinical texts using deep neural approaches. *BMC Medical Informatics and Decision Making*, 19(7):274.

Anthony Yazdani, Ihor Stepanov, and Douglas Teodoro. 2025. Gliner-biomed: A suite of efficient models for open biomedical named entity recognition.