MeSHClass-ES and AnatEM-ES: Open Resources for Spanish Biomedical NLP

Santiago Martínez Novoa[1], Lina María Gomez Meza[1], Juan Camilo Prieto[1], Ruben Manrique[1]
[1]Universidad de los Andes, Bogotá, Colombia
{s.martinezn, l.gomez1, jc.prietoa, rf.manrique}@uniandes.edu.co

Abstract

Despite Spanish being one of the most widely spoken languages in the world, biomedical NLP resources and systematic evaluations remain limited relative to English. We address this gap by constructing and releasing two Spanish biomedical corpora: (1) **MeSHClass-ES**, a 29,063 abstract bilingual corpus translated from PubMed with Opus-MT, and (2) **AnatEM-ES**, the AnatEM anatomical entity corpus translated with a chunk-level LLM-based pipeline that jointly preserves BIO annotations across 13,849 entity mentions. Both corpora achieve a mean COMET score of 0.73 despite using different translation systems. We benchmark nine encoder models spanning general-domain Spanish, domain-specific, and multilingual architectures for both tasks. RigoBERTa-2.0 leads both tasks (micro-F1 classification 0.69, tied with SciBETO-large; NER F1 0.66). Both domain pretraining and model capacity drive performance, with the gap slightly more pronounced for NER (4-point spread) than classification (3-point spread). XLM-RoBERTa-large emerges as a competitive multilingual baseline. A parallel evaluation of four open-weight decoders (7–9B) reveals a task-dependent encoder-decoder gap: QLoRA-adapted Gemma-2-9B reaches 88% of the best encoder on classification (micro-F1 .61 vs .69), but for NER every decoder configuration we tested stays at or below 40% of the best encoder F1. We release both corpora on the HuggingFace Hub[1], translation pipelines, and evaluation code on GitHub.[2]

1 Introduction

Biomedical NLP has advanced significantly through transformer-based models trained in large English corpora, allowing effective entity recognition (Lee et al., 2019), clinical prediction (Huang et al., 2020), and concept normalization (Wei et al., 2015). However, these advances remain largely inaccessible to Spanish-speaking communities: while English has 11.35 models per million speakers on HuggingFace, Spanish has only 2.30 (García Subies et al., 2024). This disparity is particularly acute in the biomedical domain, where dedicated models such as BioBERT (Lee et al., 2019), PubMedBERT (Gu et al., 2021) and BioGPT (Luo et al., 2022) exist for English, but Spanish equivalents are limited to a handful of initiatives (Carrino et al., 2021b,a).

Two questions remain open for Spanish biomedical NLP. First, which of the available Spanish and multilingual encoder architectures is best suited for biomedical tasks, and how much does domain-specific pretraining matter? Second, can open-weight instruction-tuned LLMs offer a viable alternative without task-specific fine-tuning?

We answer these questions on two complementary tasks: broad-topic MeSH classification (no open Spanish corpus exists at PubMed scale) and fine-grained anatomical NER. Existing native Spanish biomedical NER corpora (PharmaCoNER, DisTEMIST, SympTEMIST) target drugs, diseases, and symptoms; none provide the 12-type anatomical granularity of AnatEM (Pyysalo and Ananiadou, 2014), which makes translation the only route to an anatomy-focused Spanish NER benchmark at this granularity.

We address these questions with three contributions:

1. **Two Spanish biomedical corpora** built with translation approaches suited to each downstream task: MeSHClass-ES (44,604 abstracts translated with Opus-MT; 29,063 with MeSH labels of major-topics used for classification) and AnatEM-ES (1,212 documents, LLM-

[1]MeSHClass-ES: https://huggingface.co/datasets/Flaglab/pubmed_mesh_spanish AnatEM-ES: https://huggingface.co/datasets/Flaglab/anat-em-es

[2]https://github.com/SantiagoM99/spanish-medical-nlp

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 617–629

based fragment translation with joint preservation of BIO-label and human verification) for NER. Both are COMET-assessed and are released in bilingual format.

2. **A systematic encoder benchmark** of nine models that span general-domain Spanish, domain-specific, and multilingual architectures for both tasks.

3. **Evaluation of a decoder** of four open-weight LLMs (7–9B parameters) in zero-shot, few-shot, retrieval-augmented, and QLoRA settings, demonstrating that current decoders cannot substitute for fine-tuned encoders.

Our work aligns with the growing calls for the evaluation of non-English biomedical NLP (Neves et al., 2023; García Subies et al., 2024) and contributes reproducible resources to the community.

2 Related Work

Biomedical NLP in Spanish. The development of Spanish biomedical NLP has been driven by two lines of work. On the resource side, the Barcelona Supercomputing Center released RoBERTa-BioMed-ES, pretrained on a 1B-token biomedical and clinical corpus (Carrino et al., 2021b). The underlying Spanish Biomedical Crawled Corpus (Carrino et al., 2021a) remains one of the few large-scale open resources for the domain. On the evaluation side, shared tasks such as PharmaCoNER (Gonzalez-Agirre et al., 2019), MEDDOCAN (Marimon et al., 2019), and DisTEMIST/SympTEMIST (Miranda-Escalada et al., 2022) have produced annotated corpora for clinical NER. A recent survey by García Subies et al. (2024) benchmarked 17 corpora across Spanish clinical encoder models, confirming the value of domain-specific pretraining, but noting the persistent gap between clinical and biomedical evaluation resources. RigoBERTa-2.0 (Vaca Serrano et al., 2022) adapted the XLM-RoBERTa-large parameters (∼560M) to Spanish through continued pretraining in a curated Spanish corpus and was subsequently extended to the clinical domain as RigoBERTa-Clinical (Subies et al., 2025), trained in the ClinText-SP corpus of Spanish clinical text. SciBETO (Flaglab, 2024) is the first encoder pretrained in Spanish scientific text, compiling 1.9M documents (18.2B tokens) from Colombian and international open-access repositories and training RoBERTa-base (125M) and large (355M) variants.

Benchmarking biomedical models. In English, comprehensive benchmarks such as BLURB (Gu et al., 2021) and recent LLM evaluations (Chen et al., 2025) have established baselines for NER, relation extraction, classification, and answer questions. Gade et al. (2025) evaluated 12 model architectures in 61 biomedical corpora for zero-shot relation triplet extraction, finding that even the best LLMs achieve F1 below 0.5, while fine-tuned encoders still dominate in their training corpora. For non-English settings, BioMistral (Labrak et al., 2024) included multilingual evaluation but relied on auto-translated English benchmarks, leaving open questions about performance on domain-translated tasks.

MeSH classification. Automatic MeSH indexing has been driven by the BioASQ challenge series (Tsatsaronis et al., 2015), with representative approaches including DeepMeSH, which couples learned sparse and dense representations for large-scale multi-label indexing (Peng et al., 2016), and self-attentive architectures such as MeSHProbeNet (Xun et al., 2019). All of these efforts are targeted at English PubMed. No comparable Spanish evaluation setup has been released, which MeSHClass-ES addresses.

Translation-based corpus construction. Neural machine translation has been used to create biomedical resources in low-resource languages, with Helsinki-NLP Opus-MT models (Tiedemann and Thottingal, 2020) serving as a common backbone. For span-annotated corpora, label preservation throughout translation is typically handled by annotation projection (Yarowsky et al., 2001), which decouples translation from alignment and can fail in domain text where aligner coverage is poor. Recent work has systematically compared these approaches in clinical settings: Gaschi et al. (2023) benchmarked cross-lingual transfer against translate-train and translate-test strategies for clinical NER in French and German, showing that translation-based methods can match cross-lingual transfer but require careful alignment design; Lanz and Pecina (2025) applied a similar translation pipeline with answer projection for clinical question answering across six languages, finding that multilingual models can compete with monolingual domain-specific ones. We instead use an LLM-based chunk translation that preserves BIO tags via instruction, trading explicit alignment for the model's joint handling of translation

and tagging (§3.2). Previous work has validated translated biomedical corpora through BLEU (Papineni et al., 2002) and more recently COMET (Rei et al., 2020), a learned metric that correlates more strongly with human judgments for domain text; we adopt COMET-based validation following these recommendations.

3 Dataset Construction

We construct and release two Spanish biomedical corpora, each built with a translation approach suited to its downstream task. MeSHClass-ES consists of plain abstracts with document-level MeSH labels: sentence-level neural MT with Opus-MT is sufficient, cheap at 44,604 document scale, and leaves labels untouched. AnatEM-ES is token-annotated with BIO tags, which makes naive sentence MT insufficient: we use a chunk-level LLM pipeline (GPT-4o) that jointly translates the text and preserves the BIO labels via instruction, avoiding explicit cross-lingual alignment at the cost of a more expensive per-document call. Figure 1 illustrates both pipelines. Both corpora are distributed in bilingual format, retaining the original English text alongside the Spanish translations, enabling future re-translation with improved MT systems without repeating the source collection and curation pipeline. Both corpora will be released on the HuggingFace Hub.

3.1 MeSHClass-ES corpus

Source selection. We retrieve open-access abstracts in Spanish from PubMed via the NCBI E-utilities API, combining three complementary search strategies to overcome the 9,999-result API limit: year-by-year queries (1988–2025), queries over 20 Spanish-language biomedical journals, and queries over 35 MeSH subject headings. All queries require `spanish[Language]`, `hasabstract`, and open-access filters. From 78,733 initial hits, we retrieve 49,039 unique PMIDs, filtered to yield a final corpus of **44,604 articles** with complete metadata (title, abstract, MeSH descriptors, journal, publication year). Of these, **29,063** carry at least one major-topic MeSH descriptor (`MajorTopicYN="Y"`) and form the classification benchmark; the remaining articles are retained in the bilingual release but excluded from classification, as non-major descriptors are too broad to serve as supervisory labels. The top journals and the full collection funnel are reported in Appendix C.

Translation. We translate abstracts and titles into Spanish using `Helsinki-NLP/opus-mt-en-es` (Tiedemann and Thottingal, 2020) sentence by sentence. We chose Opus-MT for its open-source availability, reproducibility, and established use in prior translation-based corpus construction (Gaschi et al., 2023). We did not compare alternative MT systems (e.g., NLLB, MadLad) for MeSHClass-ES; however, the convergence of COMET scores between Opus-MT and GPT-4o on AnatEM-ES (both 0.73) despite their vastly different architectures and training data suggests that absolute translation quality is bounded by domain difficulty rather than MT system choice (Section 6). The bilingual release format further mitigates this limitation by enabling re-translation with improved systems without repeating the collection pipeline. All 44,604 articles were successfully translated (166 were already in Spanish and bypassed translation). The mean abstract length expands from 193.5 words in English to 224.2 words in Spanish ($\sim$1.16$\times$ expansion), consistent with typical EN$\rightarrow$ES translation ratios.

Translation quality. We evaluated a random sample of 200 abstracts with COMET (`Unbabel / wmt22-comet-da`) (Rei et al., 2020), a neural metric trained on human quality judgments. Our Opus-MT translations achieve a mean score of 0.73 (SD = 0.08; median = 0.74; range: 0.22–0.85), with 38% above 0.75 and 6% above 0.80. These values indicate moderate quality; following Freitag et al. (2022), we use COMET rather than BLEU as it provides more reliable quality estimates for domain text. The low proportion above 0.80 reflects the inherent difficulty of biomedical translation, where specialized terminology, nominal compounds, and abbreviations challenge general-domain MT. We discuss downstream implications in Section 6.

Corpus statistics. Figures 2–3 show the temporal distributions of the MeSH category. Health Care (N) and Techniques (E) dominate; Anatomy (A), Humanities (K), and Technology/Industry (J) are the rarest, producing a 9:1 imbalance ratio.

3.2 AnatEM-ES corpus

Translation with joint annotation preservation. The AnatEM corpus (Pyysalo and Ananiadou, 2014) (1,212 documents, 13,701 anatomical entity mentions across 12 types) is distributed in token-per-line BIO format, which makes naive transla-

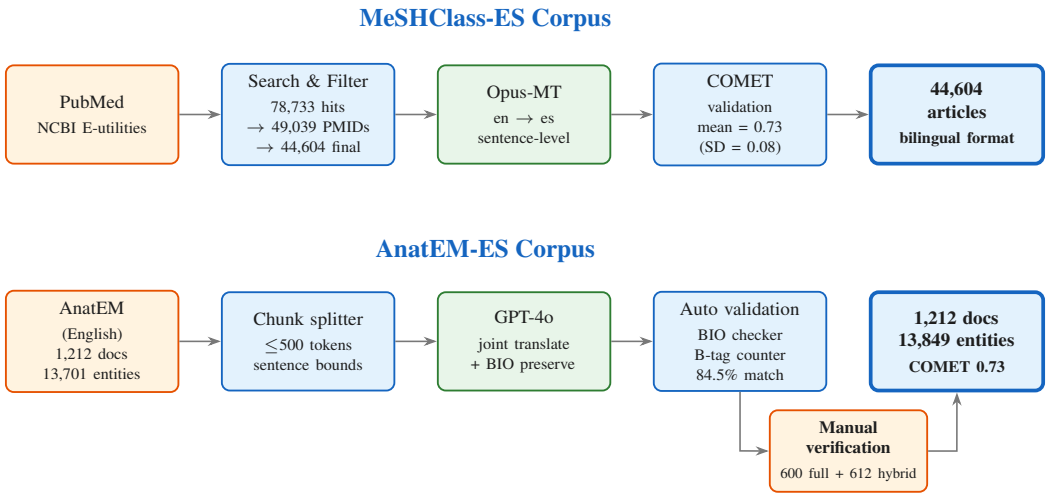

Figure 1: Construction pipelines for MeSHClass-ES (top) and AnatEM-ES (bottom). The two corpora use different translation approaches matched to their annotation requirements: sentence-level neural MT for plain-text abstracts, and chunk-level LLM-based translation with joint BIO-label preservation for token-annotated documents.

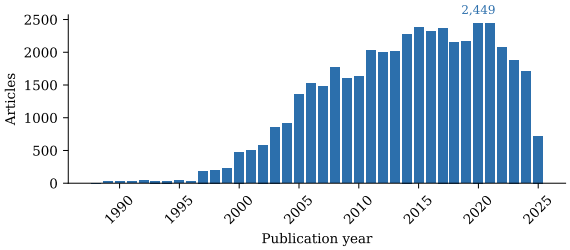

Figure 2: Publication year distribution of MeSHClass-ES articles (1988–2025). Peaks around 2015 and 2020–2021 reflect growth in open-access Spanish biomedical publishing.

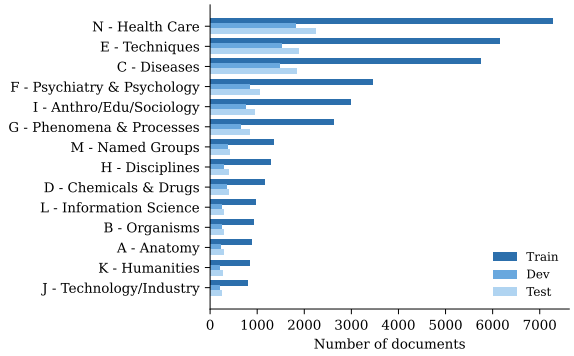

Figure 3: MeSH Level-1 category distribution across train/dev/test splits. Categories V (Publication Type) and Z (Geography) are excluded.

tion problematic: splitting by token loses linguistic context, while translating plain text loses the alignment between tokens and labels. Standard annotation projection via word aligners (Dou and Neubig, 2021) assumes roughly monotonic token correspondences, which breaks down when EN→ES translation introduces word reordering or inserts prepositions within entity spans (e.g., "blood samples" → "muestras de sangre", where the inserted preposition "de" must receive an I-tag despite having no source-side counterpart). We therefore adopt a chunk-level LLM-based pipeline using GPT-4o via the OpenAI Batch API that sidesteps explicit alignment by handling translation and re-labeling as a single operation. Each document is split into chunks of up to 500 tokens at natural sentence boundaries (preserving document integrity), and the model is instructed via a domain-specific prompt (Appendix A) to (i) translate the text into fluent Spanish, (ii) keep the original BIO labels attached to the translated tokens, and (iii) preserve the 12 entity types. This trades aligner-based guarantees for instruction-following fidelity, at the cost of relying on the LLM's ability to jointly manage translation and tagging. We do not compare the two strategies head-to-head; such a comparison would be a valuable direction for future work.

Validation. Two automatic validators check all outputs: a BIO consistency checker (orphan I-tags, type mismatches, invalid formats) and a B-tag count comparator against the English original. The translated corpus contains **13,849 entity mentions** (+1.1% vs. English), with exact counts matching in 1,024/1,212 documents (84.5%); typical discrepancies are occasional duplicate segments or minor label-formatting variants. One author manually verified the first 600 documents for BIO conven-

Entity Type	Train	Dev	Test
Anatomical_system	67	7	39
Cancer	1,469	513	1,129
Cell	2,459	666	1,447
Cellular_component	413	138	287
Developing_anat_str.	45	20	35
Immaterial_anat_ent.	116	50	97
Multi-tissue_str.	884	274	559
Organ	413	146	306
Organism_subdiv.	166	62	109
Organism_subst.	338	102	244
Pathological_form.	182	63	151
Tissue	464	129	260
Total	7,016	2,170	4,663

Table 1: AnatEM-ES entity type distribution (606/202/404 documents in train/dev/test).

tions, entity typing, and fluency; the remaining 612 followed a hybrid protocol prioritizing documents with the largest number of flagged inconsistencies, applying LLM-assisted correction at scale, and a second manual pass on critical cases.

Translation quality. All 1,212 documents are scored with COMET (`Unbabel/wmt22-comet-da`); AnatEM-ES reaches mean 0.73 (SD 0.06; range 0.35–0.99; 36% above 0.75, 8% above 0.80). These scores are nearly identical to MeSHClass-ES (0.73) despite using a different MT system (GPT-4o vs. Opus-MT), suggesting a domain-level quality ceiling rather than a system-specific limitation.

Entity distribution. Table 1 shows the entity type distribution in AnatEM-ES. Cell mentions dominate (35%), followed by Cancer (22%) and Multi-tissue_structure (13%). Rare types such as Developing_anatomical_structure (45 train mentions) and Anatomical_system (67 train) present challenges for fine-grained NER evaluation.

3.3 Corpus quality characterization

We characterize both corpora as silver-standard resources with respect to translation quality. For MeSHClass-ES, the supervisory labels are gold-standard (assigned by trained NLM indexers to the original English articles) and are not affected by translation, but the Spanish text is machine-translated without systematic human post-editing. Level-1 MeSH categories are derived by taking the first letter of each MeSH descriptor's tree number (e.g., C01.539 maps to category C, Diseases). For AnatEM-ES, entity annotations have been verified through a combination of full manual review (600 documents) and a hybrid protocol targeting flagged inconsistencies (612 documents); while this exceeds typical silver-standard corpora, residual annotation noise from the translation process likely persists in a subset of documents. We encourage users to consider the bilingual format for cross-referencing against the original English annotations when high-fidelity labels are required.

3.4 Task formulations

MeSH classification. Multi-label classification over 14 MeSH Level-1 categories (A through N, excluding Geography and Publication Type). We restrict the task to the 29,063 articles with at least one MajorTopic MeSH descriptor (`MajorTopicYN="Y"`), since non-major descriptors are too broad to serve as supervisory labels, and apply a random split (18,600 / 4,650 / 5,813 train/dev/test, approximately 64/16/20; mean 1.96 labels per document). All models receive title + abstract concatenation.

AnatEM NER. Token-level BIO tagging over 12 anatomical entity types, using the original AnatEM splits: 606 / 202 / 404 train/dev/test documents.

4 Experimental Setup

4.1 Models

Encoder fine-tuning. We evaluated nine encoders in three categories: *general-domain Spanish* (BETO (Cañete et al., 2023), BERTIN (de la Rosa et al., 2022), and RigoBERTa-2.0 (Vaca Serrano et al., 2022)); *specific-domain* (SciBETO-base/large (Flaglab, 2024), RigoBERTa-Clinical (Subies et al., 2025), RoBERTa-Bio-Clinical-ES (Carrino et al., 2021b)); and *multilingual* (XLM-RoBERTa base / large (Conneau et al., 2020)). We use `AutoModelForSequenceClassification` for classification and `AutoModelForTokenClassification` with BIO labels and first-piece sub-word alignment for NER.

Decoder zero/few-shot. Four open-weight instruction-tuned LLMs: Gemma-2-9B-it (Team et al., 2024), Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Qwen2.5-7B-Instruct (Qwen et al., 2025), and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), all loaded in 4-bit NF4 quantization. For classification, we use zero-shot and few-shot (3 examples) prompts in Spanish. For NER, we additionally test the k-NN few-shot retrieval

Hyperparameter	Search space
Learning rate	{1e-5, 2e-5, 3e-5}
Batch size	{16, 32}
Max epochs	10
Warmup ratio	0.06
Weight decay	0.1

Table 2: Encoder fine-tuning hyperparameter search space, applied to both MeSHClass-ES and AnatEM-ES.

(k = 5) using FAISS with `all-MiniLM-L6-v2` embeddings (Reimers and Gurevych, 2019), and optional self-verification (Wang et al., 2023) in which the model re-examines its own predictions. All prompts are provided in the Appendix A.

PEFT. We fine-tuned all four decoders with QLoRA (Dettmers et al., 2023) (r = 16, α = 32, dropout=0.05) in attention projection layers, NF4 quantization, AdamW 8-bit. The NER QLoRA prompt emits a linear space-separated BIO tag sequence (one tag per input token), matched between training and inference; this differs from the JSON format used by zero-shot and few-shot prompts (see Appendix A).

4.2 Training configuration

For encoder fine-tuning, we grid-search over the hyperparameters in Table 2 (the same space for both tasks), with maximum sequence length 512, AdamW FP16, sigmoid + threshold 0.5 for classification. Checkpoints are selected by validation loss and test F1 is reported once per model from the best-validation-loss configuration, so the test split is not used for model selection. QLoRA runs use 3 epochs, effective batch size 16 (4× gradient accumulation), lr = 2e-4. Decoder inference uses greedy decoding with `max_new_tokens` 100 (cls) / 512 (NER).

4.3 Evaluation metrics

Classification: F1-score (micro and macro); micro precision and recall are also reported for decoders. NER: entity-level precision, recall, and F1-score via `seqeval` (Nakayama, 2018) with strict span matching.

5 Results

5.1 Multi-label classification

Table 3 presents encoder classification results. All models fall within a narrow 3-point micro-F1 range (0.67–0.69), suggesting that the task is moderately tractable for fine-tuned encoders. RigoBERTa-2.0 leads both metrics (micro-F1 0.692, macro-F1 0.584), tied on micro-F1 by SciBETO-large (0.692, macro-F1 0.577); XLM-RoBERTa-large follows at micro-F1 0.688. Among the domain-specific models, RigoBERTa-Clinical is the closest to the leaders on macro-F1 (0.580), consistent with its clinical pretraining giving it broader coverage of tail categories. BERTIN and BETO occupy ranks 5 and 7 of 9, suggesting that classification relies more on topic-level semantics than on domain vocabulary.

Model	Mi-F1	Ma-F1
RigoBERTa-2.0	**.692**	**.584**
SciBETO-large	.692	.577
XLM-R (large)	.688	.578
RigoBERTa-Clinical	.681	.580
BERTIN	.679	.573
SciBETO-base	.678	.561
BETO	.675	.553
RoBERTa-Bio-Clin.	.675	.557
XLM-R (base)	.666	.556

Table 3: Encoder multi-label MeSH classification results (test set, T+A input). Best per column in **bold**; RigoBERTa-2.0 and SciBETO-large tie at the reported precision on micro-F1 (0.6921 vs. 0.6917 at four decimals).

Table 4 presents decoder results. The zero-shot best decoder (Gemma-2-9B, micro-F1 .458) reaches 66% of the best encoder. Few-shot prompting hurts micro-F1 for three of four decoders (Gemma −3 pts, Llama −4, Qwen −8), with Mistral as the exception (+3 pts from .355 to .384); macro-F1, in contrast, improves mildly across all four (+1 to +5 pts), indicating that fixed exemplars help tail-category coverage even as they hurt overall precision by anchoring models toward the exemplar label combinations. QLoRA adaptation tells a very different story: Gemma-2-9B reaches micro-F1 .608 / macro-F1 .500, within 8 points of the best encoder (micro-F1 .692) and corresponding to ∼88% of encoder performance; Mistral-7B follows at .535 / .450. Llama (.427) and Qwen (.416) trail on micro-F1, but all four models show sharply improved macro-F1 (+14–26 points over zero-shot), indicating that PEFT helps with tail-category coverage across the board. Model choice matters substantially under QLoRA: the 9B Gemma and 7B Mistral outperform Llama and Qwen by 11–19 micro-F1 points with the same training recipe, narrowing the encoder–decoder gap considerably for classification.

Setting	Model	Mi-P	Mi-R	Mi-F1	Ma-F1
Zero-shot					
	Gemma-2-9B	.352	**.654**	**.458**	**.282**
	Llama-3.1-8B	.398	.459	.426	.178
	Qwen2.5-7B	**.401**	.438	.418	.216
	Mistral-7B	.259	.564	.355	.190
Few-shot (3 ex.)					
	Gemma-2-9B	**.325**	**.615**	**.426**	**.289**
	Llama-3.1-8B	.317	.497	.387	.217
	Qwen2.5-7B	.292	.412	.342	.234
	Mistral-7B	.284	.591	.384	.244
QLoRA					
	Gemma-2-9B	**.619**	**.597**	**.608**	**.500**
	Llama-3.1-8B	.564	.343	.427	.367
	Qwen2.5-7B	.593	.320	.416	.358
	Mistral-7B	.611	.475	.535	.450

Table 4: Decoder classification results (T+A input, greedy decoding). Best per column in **bold**.

Model	P	R	F1
RigoBERTa-2.0	**.661**	.661	**.661**
RigoBERTa-Clinical	.638	**.664**	.651
XLM-R (large)	.640	.655	.647
SciBETO-large	.638	.648	.643
RoBERTa-Bio-Clin.	.644	.622	.633
BETO	.624	.630	.627
BERTIN	.620	.622	.621
XLM-R (base)	.616	.621	.618
SciBETO-base	.612	.623	.618

Table 5: Encoder NER results (entity-level, strict span match) on the translated AnatEM-ES test set. Best per column in **bold**.

5.2 Named entity recognition

Table 5 shows the NER encoder results. The spread is 4 F1 points, slightly larger than for classification. RigoBERTa-2.0 achieves the best F1 (0.661), ahead of RigoBERTa-Clinical (0.651) and XLM-RoBERTa-large (0.647). SciBETO-large (0.643) and RoBERTa-Bio-Clinical-ES (0.633) form the next tier, both benefiting from domain-focused pretraining. BETO (0.627), BERTIN (0.621), XLM-RoBERTa-base and SciBETO-base (both 0.618) trail behind. While RigoBERTa-2.0 leads both tasks, the mid-table reshuffles substantially: SciBETO-large drops from a tie for 1st on classification to 4th on NER, and RigoBERTa-Clinical rises from 4th to 2nd, reinforcing that task-specific evaluation is essential.

Table 6 presents decoder NER results. The encoder-decoder gap is even more pronounced than in classification. Gemma-2-9B achieves the best decoder F1 (0.27), roughly 1.8× better than other decoders but still only 40% of the best encoder. Contrary to expectations, k-NN retrieval-augmented prompting substantially *degrades* performance: across all four decoders, recall collapses from .13–.30 to $\leq$.012 while precision stays moderate (.17–.27), suggesting that retrieved examples encourage pattern copying rather than entity boundary detection, a failure mode amplified by translation-induced span inconsistencies. Adding a self-verification step raises precision sharply (.23–.57) but does not recover recall, so F1 stays at $\leq$.023 for every decoder. QLoRA-adapted decoders collapse entirely (F1 $\leq$ .004 for all four models), indicating that the generative BIO-labeling formulation is fundamentally unsuited for 7–9B parameter decoders under these training conditions; the gap between the best and worst QLoRA decoder is within rounding error, so model choice does not rescue the formulation here.

Setting	Model	P	R	F1
Zero-shot				
	Gemma-2-9B	**.243**	**.295**	**.267**
	Llama-3.1-8B	.117	.211	.151
	Qwen2.5-7B	.132	.126	.129
	Mistral-7B	.115	.129	.122
k-NN few-shot (k=5)				
	Gemma-2-9B	.228	**.012**	**.024**
	Llama-3.1-8B	**.266**	**.012**	.023
	Qwen2.5-7B	.211	.007	.013
	Mistral-7B	.171	.004	.009
k-NN + self-verification				
	Gemma-2-9B	.471	**.012**	**.023**
	Llama-3.1-8B	**.567**	.004	.007
	Qwen2.5-7B	.357	.006	.013
	Mistral-7B	.230	.004	.008
QLoRA				
	Gemma-2-9B	.002	**.006**	.003
	Llama-3.1-8B	.000	.000	.000
	Qwen2.5-7B	.000	.001	.001
	Mistral-7B	**.005**	.004	**.004**

Table 6: Decoder NER results (entity-level, strict span match). Best F1 per column in **bold**. k-NN uses FAISS with `MiniLM-L6-v2` embeddings over the training set.

6 Analysis and Discussion

Domain pretraining helps, but the advantage is task-dependent. For classification, the spread among encoders is 3 points (micro-F1 .67–.69), with general-domain base models like BETO and BERTIN performing near the domain-specialized ones. For NER, the spread is 4 points (.62–.66), with the top group (RigoBERTa-2.0, RigoBERTa-Clinical, XLM-RoBERTa-large, SciBETO-large) clearly separating from base-sized models. Classi-

fication relies more on topic-level semantics available in general pretraining; NER benefits from both model capacity and domain-specific entity coverage. Practically, choosing RigoBERTa-2.0 over BERTIN yields a 4-point F1 gain on NER but only about 1 point on classification, meaning model selection matters more for token-level tasks. We note a confound: in our model set capacity and domain pretraining are partially correlated — all four NER top performers are large models, and the only base-sized domain model (RoBERTa-Bio-Clinical-ES) ranks 5th. We cannot cleanly separate the two factors with the current pool, and a stronger claim would require a large biomedical-pretrained Spanish encoder of comparable size, which is not yet available.

XLM-RoBERTa-large is a strong multilingual baseline. Despite lacking Spanish-specific or scientific pretraining, XLM-RoBERTa-large ranks 3rd on classification (micro-F1 .688, within 0.5 points of the leaders) and 3rd on NER (F1 .647, 1.4 points behind RigoBERTa-2.0). Its competitive macro-F1 (.578) in classification indicates that cross-lingual pretraining on 100 languages provides robust coverage of tail categories. This makes it a practical default when domain-specific Spanish models are unavailable.

Encoder-decoder gap depends on task and adaptation. Zero-shot, Gemma-2-9B reaches 66% of the best encoder on classification and 40% on NER. The asymmetry reflects task nature: LLMs handle coarse topic classification reasonably but struggle with precise span identification. PEFT changes the picture sharply on classification (QLoRA Gemma 88%, Mistral 77%) but not on NER, where every decoder configuration stays below F1 .27.

Few-shot prompting degrades micro-F1 but helps macro-F1. Few-shot examples hurt micro-F1 for three of four decoders (Gemma -3, Llama -4, Qwen -8), with Mistral the exception ($+3$). The fixed exemplars (label combinations {C,D}, {C,N}, {A,C,E,M}) drive precision drops larger than recall drops in the three degrading cases, consistent with over-prediction of exemplar categories. Macro-F1, in contrast, improves across all four models ($+1$ to $+5$), suggesting that even biased exemplars surface tail categories that would otherwise be ignored. Static few-shot examples can thus be counterproductive for micro-F1 in multi-label settings unless the exemplars span the full label space.

QLoRA works for classification but fails for generative NER. For classification, QLoRA improves macro-F1 by $+14$–26 points over zero-shot across all four decoders, and Gemma/Mistral close most of the encoder gap in micro-F1. For NER, all four collapse to F1 $\leq .004$ despite training and evaluating in the same BIO format (§A): models emit all-O predictions (Llama, Qwen) or malformed tag strings (Gemma, Mistral). The generative BIO formulation requires one tag per token with valid B/I dependencies, which at 7–9B and 3-epoch, r=16 QLoRA none of the four decoders learn; longer training, higher-rank adapters, or a span-extraction output (e.g., JSON) would likely be needed. This aligns with English findings (Gade et al., 2025) but shows a wider gap, likely amplified by limited Spanish instruction-tuning data.

Translation quality likely caps absolute performance. COMET scores (mean 0.73 for both corpora, consistent across Opus-MT and GPT-4o) suggest a domain-level quality ceiling that bounds absolute F1, especially on NER where entity boundaries must align across languages. Since all models process the same translated text, relative rankings should be less affected. A cross-benchmark comparison on a native Spanish corpus would be the natural next step to verify this; however, no existing Spanish biomedical NER corpus covers anatomical entities at the granularity of AnatEM-ES. The closest available resource, PharmaCoNER, targets pharmacological entities and clinical text, making it unsuitable as a direct comparison point. This gap is precisely the motivation for the present work.

Decoder NER error regimes. Gemma-2-9B leads the decoder pool on NER (F1 .27, $\sim$1.8$\times$ the next best), likely reflecting its larger effective capacity, instruction-tuning data, or biomedical coverage. Manual inspection on a 449-sentence sample shows that the encoder–decoder gap is qualitative, not just quantitative. Zero-shot decoders over-predict: Llama emits 8,393 predictions for 4,669 gold mentions, and Mistral hallucinates entities in 51 sentences with no anatomical content (e.g., labeling "Cuba"/"Florida" as `Organism_subdivision`). Even on correct spans, all four models systematically mistype them (e.g., "arteria coronaria derecha" as `Organ` instead of `Multi-tissue_structure`). The k-NN and QLoRA settings exhibit the opposite regime: 89–

97% of sentences receive zero predictions and recall drops to 0.003–0.012, consistent with pattern-copying of retrieved exemplars and all-O collapse under QLoRA.

7 Conclusion

We released MeSHClass-ES (44,604 abstracts; 29,063 with major-topic MeSH labels) and AnatEM-ES (1,212 documents, 13,849 entities), two Spanish biomedical corpora built with translation pipelines matched to each task. RigoBERTa-2.0 leads both benchmarks, domain pretraining and capacity drive performance (more so on NER), and XLM-RoBERTa-large is a strong multilingual baseline. The encoder–decoder gap is task-dependent: QLoRA-adapted Gemma-2-9B reaches 88% of the best encoder on classification, but NER remains out of reach for all four 7–9B decoders. The bilingual corpus format enables re-translation as MT improves, without repeating collection.

Limitations

Our corpora are constructed via machine translation rather than from native Spanish biomedical text. COMET evaluation indicates moderate translation quality for both corpora (mean 0.73 for both MeSHClass-ES and AnatEM-ES, with 6–8% of documents above the 0.80 threshold), reflecting the difficulty of translating biomedical terminology regardless of the MT system used; this quality ceiling likely caps absolute downstream performance. For AnatEM-ES, the LLM-based translation with joint BIO preservation avoids explicit cross-lingual alignment but relies on the LLM's instruction-following fidelity: 15.5% of documents show entity-count discrepancies, and while manual verification covered the first 600 documents in full plus the most critical cases in the remaining 612, residual annotation noise likely persists. We evaluated two translation systems (Opus-MT and GPT-4o) but did not compare them head-to-head. Our encoder benchmark covers a grid over three learning rates and two batch sizes; Bayesian optimization could improve individual results. The decoder evaluation is limited to 7–9B parameter models; larger models may narrow the gap.

Ethical Considerations

All data are derived from publicly available PubMed abstracts and the open AnatEM corpus. Patient data or personally identifiable information are not involved. Our translation pipelines use Opus-MT (open-source) and the OpenAI Batch API (GPT-4o); the output is released together with the translation scripts. Machine-translated medical text should not be used for clinical decision-making without expert review.

References

Casimiro Pio Carrino, Jordi Armengol-Estapé, Ona de Gibert Bonet, Asier Gutiérrez-Fandiño, Aitor Gonzalez-Agirre, Martin Krallinger, and Marta Villegas. 2021a. Spanish biomedical crawled corpus: A large, diverse dataset for spanish biomedical language models. *Preprint*, arXiv:2109.07765.

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

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2023. Spanish pre-trained bert model and evaluation data. *Preprint*, arXiv:2308.02976.

Qingyu Chen, Yan Hu, Xueqing Peng, Qianqian Xie, Qiao Jin, Aidan Gilson, Maxwell B. Singer, Xuguang Ai, Po-Ting Lai, Zhizheng Wang, Vipina K. Keloth, Kalpana Raja, Jimin Huang, Huan He, Fongci Lin, Jingcheng Du, Rui Zhang, W. Jim Zheng, Ron A. Adelman, and 2 others. 2025. Benchmarking large language models for biomedical natural language processing applications and recommendations. *Nature Communications*, 16(1):3280.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *Preprint*, arXiv:1911.02116.

Javier de la Rosa, Eduardo G. Ponferrada, Paulo Villegas, Pablo Gonzalez de Prado Salas, Manu Romero, and Maria Grandury. 2022. Bertin: Efficient pre-training of a spanish language model using perplexity sampling. *Preprint*, arXiv:2207.06814.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. pages 2112–2128.

Flaglab. 2024. Scibeto collection. https://huggingface.co/collections/Flaglab/scibeto. Accessed: 2026-04-17.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Frederik Gade, Ole Lund, and Marie Lisandra Mendoza. 2025. Benchmarking zero-shot biomedical relation triplet extraction across language model architectures. In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 88–100, Viena, Austria. Association for Computational Linguistics.

Guillem García Subies, Álvaro Barbero Jiménez, and Paloma Martínez Fernández. 2024. A comparative analysis of spanish clinical encoder-based models on NER and classification tasks. *J. Am. Med. Inform. Assoc.*, 31(9):2137–2146.

Félix Gaschi, Xavier Fontaine, Parisa Rastin, and Yannick Toussaint. 2023. Multilingual clinical NER: Translation or cross-lingual transfer? In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 289–311, Toronto, Canada. Association for Computational Linguistics.

Aitor Gonzalez-Agirre, Montserrat Marimon, Ander Intxaurrondo, Obdulia Rabal, Marta Villegas, and Martin Krallinger. 2019. PharmaCoNER: Pharmacological substances, compounds and proteins named entity recognition track. In *Proceedings of the 5th Workshop on BioNLP Open Shared Tasks*, pages 1–10, Hong Kong, China. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2020. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *Preprint*, arXiv:1904.05342.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. BioMistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5848–5864, Bangkok, Thailand. Association for Computational Linguistics.

Vojtech Lanz and Pavel Pecina. 2025. When multilingual models compete with monolingual domain-specific models in clinical question answering. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, pages 69–82, Albuquerque, New Mexico. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6).

Montserrat Marimon, Aitor Gonzalez-Agirre, Ander Intxaurrondo, Heidy Rodriguez, Jose Antonio Lopez Martin, Marta Villegas, and Martin Krallinger. 2019. Automatic de-identification of medical texts in Spanish: the MEDDOCAN track, corpus, guidelines, methods and evaluation of results. In *IberLEF@ SEPLN*, pages 618–638.

Antonio Miranda-Escalada, Luis Gasco, Salvador Lima-López, Eulàlia Farré-Maduell, da Lisa Jaramillo Estrada, Anastasios Nentidis, Anastasia Krithara, Georgios Katsimpras, Georgios Paliouras, and Martin Krallinger. 2022. Overview of distemist at bioasq: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources. In *Conference and Labs of the Evaluation Forum*.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Mariana Neves, Antonio Jimeno Yepes, Aurélie Névéol, Rachel Bawden, Giorgio Maria Di Nunzio, Roland Roller, Philippe Thomas, Federica Vezzani, Maika Vicente Navarro, Lana Yeganova, Dina Wiemann, and Cristian Grozea. 2023. Findings of the WMT 2023 biomedical translation shared task: Evaluation of ChatGPT 3.5 as a comparison system. In *Proceedings of the Eighth Conference on Machine Translation*, pages 43–54, Singapore. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Shengwen Peng, Ronghui You, Hongning Wang, Chengxiang Zhai, Hiroshi Mamitsuka, and Shanfeng Zhu. 2016. DeepMeSH: deep semantic representation for improving large-scale MeSH indexing. *Bioinformatics*, 32(12):i70–i79.

Sampo Pyysalo and Sophia Ananiadou. 2014. Anatomical entity mention recognition at literature scale. *Bioinformatics*, 30(6):868–875.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Preprint*, arXiv:1908.10084.

Guillem García Subies, Álvaro Barbero Jiménez, and Paloma Martínez Fernández. 2025. Clintext-sp and rigoberta clinical: a new set of open resources for spanish clinical nlp. *Preprint*, arXiv:2503.18594.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, and 1 others. 2015. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16(1):138.

Alejandro Vaca Serrano, Guillem Garcia Subies, Helena Montoro Zamorano, Nuria Aldama Garcia, Doaa Samy, David Betancur Sanchez, Antonio Moreno Sandoval, Marta Guerrero Nieto, and Alvaro Barbero Jimenez. 2022. Rigoberta: A state-of-the-art language model for spanish. *Preprint*, arXiv:2205.10233.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. Gpt-ner: Named entity recognition via large language models.

Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu. 2015. Gnormplus: an integrative approach for tagging genes, gene families, and protein domains. *BioMed Research International*, 2015:918710.

Guangxu Xun, Kishlay Jha, Ye Yuan, Yaqing Wang, and Aidong Zhang. 2019. MeSHProbeNet: a self-attentive probe net for MeSH indexing. In *Bioinformatics*, volume 35, pages 3794–3802.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research (HLT)*. Association for Computational Linguistics.

A Prompts

Classification: Zero-shot

```
Eres un experto en indexación biomédica
con conocimiento profundo de MeSH.

Categorías MeSH disponibles:
  A: Anatomía  B: Organismos
  C: Enfermedades
  D: Sustancias Químicas y Fármacos
  E: Técnicas Analíticas, Diagnósticas
     y Terapéuticas
  F: Psiquiatría y Psicología
  G: Fenómenos y Procesos
  H: Disciplinas y Ocupaciones
  I: Antropología, Educación y Sociología
  J: Tecnología, Industria y Agricultura
  K: Humanidades
  L: Ciencias de la Información
  M: Grupos Nombrados
  N: Cuidados de Salud
  V: Características de Publicación

Clasifica el siguiente abstract médico
asignándole los códigos MeSH de nivel 1.
Responde ÚNICAMENTE con los códigos
separados por comas (ejemplo: A, C, D).

Abstract: {text}
Códigos MeSH:
```

NER: Zero-shot

```
Eres un sistema experto en reconocimiento
de entidades de anatomía médica en español.

Tipos válidos: {entity_types}

Extrae TODAS las entidades del texto.
Responde SOLO con un array JSON:
[{"entity": "texto", "type": "tipo"}]
Si no hay entidades, responde: []

Texto: "{sentence}"
JSON:
```

NER: k-NN few-shot

```
[Same header as zero-shot]

EJEMPLOS (ordenados por similitud):
{retrieved examples from FAISS index}

Texto: "{sentence}"
JSON:
```

AnatEM-ES: Translation prompt (GPT-4o)

```
Eres un experto etiquetador médico
especializado en anatomía. Tu tarea es
traducir este dataset de entidades
anatómicas del inglés al español
manteniendo las etiquetas originales
en inglés.

REGLAS FUNDAMENTALES DE NER:
- B-[etiqueta]: Marca el INICIO de una
  entidad anatómica
- I-[etiqueta]: Marca la CONTINUACIÓN
  de la misma entidad
- O: Marca tokens que NO son entidades
  anatómicas
- NUNCA puede haber I- sin un B- previo
  de la misma etiqueta

INSTRUCCIONES CRÍTICAS:
1. Traduce SOLO las palabras, MANTÉN
   etiquetas en inglés
2. El número de etiquetas B- debe ser
   EXACTAMENTE igual al original
3. Si inviertes orden de palabras,
   ajusta etiquetas coherentemente
4. Si añades preposiciones/artículos
   para naturalidad:
   - Dentro de entidad: usar I- de la
     misma etiqueta
   - Fuera de entidad: usar O
5. No añadas etiquetas a términos que
   no tenían previamente
6. Es posible que haya textos donde hay
   ninguna etiqueta, no es necesario
   crear nuevas etiquetas.

EJEMPLOS CLAVE:

1. Inversión con preposición añadida:
ENTRADA:
cell       B-Organism_substance
lysate     I-Organism_substance
experiment O

SALIDA:
lisado     B-Organism_substance
celular    I-Organism_substance
experimento O

2. Adición de preposición dentro de
   entidad:
ENTRADA:
blood   B-Organism_substance
samples I-Organism_substance

SALIDA:
muestras B-Organism_substance
de       I-Organism_substance
sangre   I-Organism_substance

3. Inversión de términos:
ENTRADA:
ventricular B-Multi-tissue_structure
fibrillation O

SALIDA:
fibrilación  O
ventricular  B-Multi-tissue_structure

VERIFICACIÓN: (OBLIGATORIO)
No añadir etiquetas en términos que no
tenían previamente dicha etiqueta.

TEXTO A TRADUCIR:
{input_text}

TRADUCCIÓN:
```

B Extended Encoder Results

Table 7 reports the precision, recall, and F1 per-entity-type for the three top-ranked encoders on AnatEM-ES. Performance tracks entity frequency: high-frequency types such as Cell (F1 .72–.75) and Cancer (F1 .69–.70) are relatively stable across models, whereas low-frequency types such as Anatomical_system (F1 .27–.49) and Immaterial_anatomical_entity (F1 .37–.40) show both lower scores and greater inter-model variance. RigoBERTa-2.0 stands out in Developing_anatomical_structure (F1 0.33) and RigoBERTa-Clinical achieves the strongest result in Organism_substance (F1 .74), consistent with its clinical-domain pretraining. All three models struggle with Pathological_formation (F1 .42–.45) and Tissue (F1 .48–.50), suggesting that these categories pose annotation- or translation-level ambiguity that domain pretraining alone does not resolve.

C MeSHClass-ES Corpus Statistics

Table 8 summarizes the collection and filtering pipeline. Of the 44,604 final articles, 30,667 were originally in English (requiring translation), 166 were already in Spanish (bypassing translation), and 13,771 had mixed or unspecified language

	RigoBERTa-2.0			RigoBERTa-Clin.			XLM-R (large)		
Entity Type	P	R	F1	P	R	F1	P	R	F1
Anatomical_system	.567	.436	.493	.583	.359	.444	.381	.205	.267
Cancer	.675	.731	.701	.668	.725	.696	.663	.724	.693
Cell	.738	.726	.732	.757	.737	.747	.732	.715	.723
Cellular_comp.	.604	.606	.605	.586	.571	.579	.561	.575	.568
Developing_anat.	.774	.686	.727	.636	.600	.618	.640	.457	.533
Immaterial_anat.	.380	.361	.370	.444	.330	.379	.478	.340	.398
Multi-tissue_str.	.519	.580	.548	.505	.506	.505	.505	.540	.522
Organ	.720	.680	.699	.672	.690	.681	.643	.660	.652
Organism_subdiv.	.587	.495	.537	.584	.477	.525	.741	.395	.515
Organism_subst.	.681	.742	.710	.713	.762	.737	.581	.693	.632
Pathological_form.	.394	.444	.417	.418	.424	.421	.435	.464	.449
Tissue	.440	.562	.493	.434	.577	.495	.433	.542	.481

Table 7: Per-entity-type NER breakdown for the top three encoders. Strict entity-level evaluation via `seqeval`. Overall scores are reported in Table 5.

metadata. Mean abstract length is 193.5 words (English) expanding to 224.2 words (Spanish); 24,856 articles (55.7%) contain structured abstracts.

Stage	**Articles**
Initial search hits	78,733
Unique PMIDs retrieved	49,039
− No abstract	−32
− No MeSH terms	−3,949
− Abstract <50 words	−454
Downloaded corpus	44,604
− No major-topic MeSH	−15,541
Classification corpus	**29,063**

Table 8: MeSHClass-ES collection funnel. The downloaded corpus of 44,604 articles is released in bilingual format; the classification benchmark uses the 29,063 with at least one major-topic MeSH descriptor.

Journal	**Articles**
Revista Médica de Chile	3,701
Rev. Española de Cardiología	2,149
Gaceta Sanitaria	2,117
Archivos Argentinos de Pediatría	2,013
Anales de Pediatría	1,871
Atención Primaria	1,870
Rev. Latino-Am. de Enfermagem	1,816
Nutrición Hospitalaria	1,772
Rev. Española de Salud Pública	1,616
Rev. Chilena de Infectología	1,327

Table 9: Top 10 contributing journals in MeSHClass-ES. The corpus spans journals from Chile, Spain, Argentina, Brazil, and Mexico, reflecting geographic diversity across the Spanish-speaking biomedical community.