Procesamiento del Lenguaje Natural, Revista nº 76, marzo de 2026, pp. 225-238 recibido 05-12-2025 revisado 17-01-2026 aceptado 09-02-2026

Discriminative Benchmarking of Spanish Language Models: Findings from the ODESIA Challenge 2024

Evaluación Discriminativa de Modelos de Lenguaje en Español: Resultados del ODESIA Challenge 2024

Alejandro Benito-Santos,[1] Roser Morante,[1] Adrián Ghajari,[1] Iker García-Ferrero,[2] Robiert Sepúlveda-Torres,[3] German Rigau,[2] Rodrigo Agerri,[2] Juan Pablo Consuegra-Ayala,[3] Ernesto L. Estevanell-Valladares,[3] Fabio Yáñez-Romero,[3] Miquel Canal-Esteve,[3] Yoan Gutiérrez,[3] Rafael Muñoz-Guillena,[3] Manuel Palomar,[3] Eva Sánchez Salido,[1] Guillermo Marco,[1] Andrés Fernández García,[1] Víctor Fresno,[1] Enrique Amigó,[1] Laura Plaza,[1] Jorge Carrillo-de-Albornoz,[1] Miguel Lucas,[4] and Julio Gonzalo[1]

[1]NLP&IR Group – UNED, Spain
[2]HiTZ Center – Ixa, University of the Basque Country UPV/EHU
[3]Dep. of Software and Computing Systems, University of Alicante, Spain
[4]Llorente y Cuenca SLU, Madrid, Spain

{al.benito,r.morant,aghajari,evasan,gmarco,afernandez}@lsi.uned.es
{vfresno,enrique,lplaza,jcalbornoz,julio}@lsi.uned.es, mlucas@llyc.global
{iker.garciaf,rodrigo.agerri,german.rigau}@ehu.eus
{robiert.sepulveda,juan.consuegra,ernesto.estevanell,fabio.yanez,mikel.canal}@ua.es
{ygutierrez,rafael,mpalomar}@dlsi.ua.es

Abstract: This paper presents the results from the 2024 ODESIA Challenge, a public competition aimed at benchmarking natural language processing (NLP) systems in Spanish across ten discriminative tasks using a standardized methodology based on private, held-out test sets. Results show the winning system (Qwen2.5-14B) prevailed due to structural advantages in extractive Question Answering, whereas encoders outperformed LLMs in other tasks such as sequence labeling and soft classification. We conclude that, while generative models may dominate reasoning-heavy tasks involving long contexts, encoder architectures obtain on-par or even better performance in many other discriminative scenarios, challenging the assumption that massive scale universally supersedes specialized architectural design.
Keywords: Benchmarking, discriminative tasks, encoders vs. decoders, Spanish NLP.

Resumen: Presentamos los resultados del ODESIA Challenge 2024, una competición abierta basada en conjuntos de prueba privados orientada a evaluar sistemas de procesamiento del lenguaje natural (PLN) en español en diez tareas discriminativas. El sistema ganador, un LLM (Qwen2.5-14B), destacó por su rendimiento en *extractive Question Answering*, mientras que los *encoders* superaron a los LLM en tareas como *sequence labeling* y *soft classification*. Concluimos que, aunque los grandes modelos generativos pueden dominar tareas de razonamiento con contextos largos, los *encoders* logran un rendimiento comparable o superior en muchos escenarios discriminativos, poniendo en tela de juicio la creencia de que el tamaño de un modelo es un factor más decisivo que el emplear una arquitectura especializada en este tipo de tareas.
Palabras clave: Benchmarking, tareas discriminativas, encoders vs. decoders, PLN en español.

1 Introduction

We present the outcomes and a technical account of the *2024 ODESIA Challenge*, a benchmarked evaluation of Natural Language Processing (NLP) systems for Spanish framed within the ODESIA (Observation Space for Artificial Intelligence in Spanish)

ISSN 1135-5948 DOI 10.26342/2026-76-15

project[1] (2022-2025). The project's main goal is to measure the development gap in Artificial Intelligence between English and Spanish across four key areas: state of the art of language technologies; market solutions; level of technology adoption; and user experience. To measure the gap in the state of the art, benchmarks have been created with private (not publicly shared), held-out test partitions in Spanish and English. In this context, the challenge presented in this paper was conceived as a large-scale evaluation campaign aimed at assessing the performance, robustness, and adaptability of NLP systems for Spanish across ten discriminative tasks. The challenge was launched to serve two primary purposes: First, it sought to measure the current capabilities of modern NLP systems—particularly large language models (LLMs)—when required to solve multiple heterogeneous discriminative tasks using a single, unified methodological approach. Second, it aimed to stimulate research on Spanish NLP by providing a controlled evaluation environment with private test sets, preventing training data contamination and allowing for reliable comparisons across models.[2]

By covering a variety of linguistic phenomena, domains, and annotation paradigms—including tasks based on rater disagreement (i.e. tasks where the objective is to predict the full label distribution of human annotators rather than a single point estimate) (Uma et al., 2021b)—the challenge provided a comprehensive benchmark for evaluating system generalization and cross-task consistency. This article documents the challenge setup and provides a detailed description of the systems for which complete technical reports were submitted—specifically, the solutions from the teams *IXA_taldea* and *GPLSI*. These two teams, along with the organizers' baseline, constitute the systems for which full methodological transparency is available, enabling an accurate and meaningful comparison. The best-performing system, submitted by *IXA_taldea* and based on the *Qwen2.5-14B-Instruct* model, achieved a macro average of **0.6306**, outperforming all other participating systems on the private test data.

While the results confirm that large language models excel in reading comprehension and semantic reasoning, they also demonstrate that traditional encoder architectures often outperform billion-parameter models in other discriminative tasks. Consequently, we argue that massive scale is not a silver bullet, and that specialized discriminative designs remain essential for solving many resource-efficient, high-precision tasks.

2 Challenge Infrastructure: the ODESIA Leaderboard

The challenge relied on the infrastructure provided by the ODESIA Leaderboard, a web-based evaluation platform designed to benchmark language models by comparing their outputs against a gold standard. The leaderboard differs from traditional multilingual benchmarks by strictly adhering to design principles that allow for a fair and direct comparison between Spanish and English capabilities (Benito-Santos, Ghajari, and Fresno, 2025; Sánchez Salido et al., 2025).

In this regard, and to ensure data independence, the English and Spanish partitions of datasets hosted on the leaderboard were not generated through translation or alignment. Instead, they were constructed independently in each language following identical guidelines and maintaining comparable data volumes. To ensure a fair comparison between languages, the difficulty of each language portion was calibrated by incorporating language-agnostic baselines and specific linguistic gap metrics. Crucially, to prevent data leakage and avoid *data contamination* (Sainz et al., 2024), the CORE part of the benchmark consists of newly created datasets with private test partitions.

The evaluation logic is handled by a specialized open-source library (PyEVaLL) [3] capable of managing complex evaluation contexts, such as the Learning with Disagreement (LeWiDi) paradigm (Uma et al., 2021b). It also supports a diverse array of metrics to mitigate the limitations associated with relying on single performance indicators (Ruder, 2021). The benchmark structure categorizes tasks into two distinct groups:

- **Core Tasks:** Comprising five datasets

[1] https://odesia.uned.es/

[2] https://leaderboard.odesia.uned.es/leaderboard/challenge

[3] https://github.com/UNEDLENAR/PyEvALL

that derive into ten specific tasks of different nature. These tasks utilize private test sets to guarantee zero contamination.

- **Extended Tasks:** Comprising 4 widely used public datasets (such as MLDoc (Schwenk and Li, 2018) or MultiCONER (Malmasi et al., 2022)) for which model contamination cannot be ruled out.

The ODESIA Challenge required each participating team to develop a *single* system capable of tackling all ten tasks of ODESIA-CORE [4] under a unified framework. This constraint was a defining characteristic: systems could make task-specific adjustments—such as prompt wording, hyperparameters, preprocessing choices, or fine-tuning procedures—but the underlying model architecture or methodological paradigm had to remain consistent across all tasks. The **ODESIA CORE tasks** for Spanish include the following tasks (see summary in Table 1):

- The **DIPROMATS 2023** (Moral et al., 2023) tasks, performed on a dataset of tweets issued by diplomats from four world powers (the European Union, Russia, China and the United States), annotated according to the propaganda techniques used to convey a particular image of their countries and competitors. Task 1 on propaganda identification involves a binary classification, while Tasks 2 and 3 on coarse- and fine-grained propaganda classification involve multilabel classification.
- The **EXIST 2022** (Rodríguez-Sánchez et al., 2022) tasks address sexism detection and categorization on a dataset of tweets annotated according to whether they contain sexist messages or not, and what type of sexism they express.
- The **EXIST 2023** (Plaza et al., 2023) tasks are similar to the EXIST 2022 tasks, but they are performed on a different dataset annotated within the Learning with Disagreement (LeWiDi) paradigm (Uma et al., 2021a) and contains also annotations for the identification of the source of the sexist content. Apart from sexism detection and sexism categorization, the source detection task was added.
- The **DIANN 2023** (Fabregat, Martínez-Romo, and Araujo, 2018) task on disability detection is a sequence labeling task performed on a dataset of biomedical abstracts annotated with disabilities.
- **SQAC 2024** is an extractive text comprehension task for question-answering systems, performed on a dataset of popular science articles from the research institution CSIC annotated with questions and answer spans following the guidelines of the SQAC dataset.

More information about the datasets and tasks can be found in their corresponding publications. The tasks in ODESIA-CORE are intentionally varied in nature including sequence labeling, binary classification, multiclass and hierarchical classification, multilabel classification, and LeWiDi tasks. This diversity ensured that systems had to demonstrate versatility across different linguistic and modeling challenges. The requirement for a unified methodology inevitably favored models capable of flexible prompting, robust generalization, and adaptation to heterogeneous label structures.

Participants had access only to the public training and development sets provided within each of the ten tasks included in the challenge. The evaluation process was fully automated through the ODESIA Leaderboard platform, which validated predictions, computed metrics, and maintained public rankings.

Systems were required to share a common architectural base. Examples of acceptable systems included, for example 1) a single encoder model fine-tuned separately per task; 2) A single generative LLM used via prompting, optionally with task-specific prompt templates; 3) An ensemble of fixed models used consistently across tasks; 4) A retrieval-augmented generative architecture with task-dependent corpora. Conversely, unacceptable systems included using unrelated LLMs per task, employing mixed paradigms (e.g., rule-based + neural + statistical) without a shared foundation, or submitting fully independent models for each task.

[4] To ensure a fair, contamination-free comparison across systems, the Extended Tasks were excluded from the challenge.

Dataset	Domain	Task	Abstract Task	Metric
DIANN 2023	Biomedical	1 Disability detection	Sequence labeling	F1
DIPROMATS 2023	Politics	1 Propaganda identification	Binary classification	Macro-F1
		2 Propaganda cat. coarse	Multiclass multilabel classification	Macro-F1
		3 Propaganda cat. fine	Multiclass multilabel classification	Macro-F1
EXIST 2022	Social	1 Sexism detection	Binary classification	Macro-F1
		2 Sexism categorization	Multiclass classification	Macro-F1
EXIST 2023	Social	1 Sexism detection	Binary soft classification	Soft-ICM
		2 Source intention	Hierarchical multiclass soft classification	Soft-ICM
		3 Sexism categorization	Hierarchical multiclass multilabel soft classification	Soft-ICM
SQAC 2024	Scientific	1 Extractive QA	Span extraction	F1

Table 1: Summary of datasets and tasks in the ODESIA challenge benchmark.

3 Evaluation Framework

The evaluation protocol of the ODESIA leaderboard was designed to promote fairness, prevent overfitting, and ensure methodological transparency.

Metrics. Different tasks required different evaluation metrics:

- **Macro-F1** was used for standard multilabel and multiclass settings (DIPROMATS 2023 and EXIST 2022)
- Normalized **Soft-ICM** (Amigo and Delgado, 2022) was used for tasks requiring the prediction of the full human label distribution (i.e., EXIST 2023).
- **F1** was used for DIANN 2023 (sequence labeling) and SQAC 2024 (extractive QA).

Final score. The official ranking metric was the **macro-average** of each system's scores across the ten tasks.

4 Baselines

To provide a robust comparative framework for the ODESIA Challenge, the *UNED* team developed a set of strong baselines covering two distinct modeling paradigms: traditional encoder-only transformers and modern decoder-only LLMs. The objective was to benchmark the tasks using a unified discriminative methodology, treating both architectures as feature extractors with classification heads rather than relying on generative prompting.

The baseline generation pipeline[5] was designed to standardize the fine-tuning process while respecting the computational and architectural differences between the model families.

Encoder Baselines. For the encoder-based baselines, a range of Spanish-centric and multilingual models were evaluated, including `RoBERTa-base-BNE`, `RoBERTa-large-BNE` (Gutiérrez-Fandiño et al., 2022), and `XLM-RoBERTa` (Conneau et al., 2020). These models underwent standard full fine-tuning, where all parameters were updated during training. The classification or token labeling heads were initialized on top of the final hidden states. A grid search was conducted to optimize performance, focusing on lower learning rates characteristic of these architectures ($\{1e^{-5}, 3e^{-5}, 5e^{-5}\}$) and batch sizes of $\{16, 32\}$.

Decoder Baselines. To assess the discriminative capabilities of Generative AI, baselines were also computed using state-of-the-art decoder models such as `Llama-3.1-8B`, `Mistral-7B-v0.3`, and `Qwen2.5-7B`. Unlike the encoders, these massive models required a Parameter-Efficient Fine-Tuning (PEFT) strategy. The system utilized Low-Rank Adaptation (LoRA) (Hu et al., 2021), injecting adapters ($r = 16, \alpha = 8$) into the attention mechanisms while keeping the backbone frozen.

[5]Code is available at `https://github.com/ale0xb/odesia_benchmark`.

Crucially, the causal language modeling head was replaced by a linear layer, forcing the model to perform direct classification rather than text generation.

4.1 Unified Task Adaptation

Despite the architectural differences, both model types shared a common adaptation strategy for the specific challenges of the ODESIA benchmark:

Standard Classification. For standard text classification tasks, such as binary or multiclass categorization (e.g., EXIST 2022, DIPROMATS 2023), the systems used a sequence classification head. The models were trained to minimize cross-entropy loss between the predicted logits and the target class. For multilabel scenarios, the problem was framed as multiple binary classifications, using a sigmoid activation function and binary cross-entropy loss to predict the presence or absence of each label independently.

Learning with Disagreement (LeWiDi). For tasks involving soft labels where inter-annotator disagreement is a key feature (e.g., EXIST 2023), the approach was modified to handle probability distributions directly. Instead of converting soft labels into a single hard ground truth, a custom training loop was implemented to minimize the Binary Cross Entropy with Logits loss between the model's output logits and the soft label distribution provided in the dataset. This allows the models to learn and predict the uncertainty associated with each instance, preserving the richness of the original annotations.

Token Classification and Question Answering. Sequence labeling tasks, such as named entity recognition (DIANN 2023), were addressed using a token classification head that assigned a label to each token in the input sequence. The pipeline included a mechanism to align the model's subword tokenization with the original word-level labels, ensuring accurate span prediction. Similarly, for extractive question answering (SQAC 2024), the systems predicted the start and end logits of the answer span within the context, effectively treating it as a specialized token classification problem.

4.2 Grid Search

For each model, the best configuration was determined through a systematic grid search. It is important to note that the hyperparameter space for these decoder-only models differed significantly from that of standard encoder baselines. Due to the use of LoRA and the distinct optimization dynamics of large decoders, higher learning rates were found to be effective. The grid search explored learning rates in the range of $\{1e^{-4}, 3e^{-4}, 1e^{-3}\}$, considerably higher than the typical $\{1e^{-5}, \ldots, 5e^{-5}\}$ range used for fully fine-tuned encoders. Other tuned parameters included batch sizes of $\{8, 16, 32\}$ and weight decay values of $\{0.0, 0.01\}$. Results for all the tested models can be consulted upon Table 5 of the Appendix. Baselines were selected from the best-performing decoder (`Qwen2.5-7B`, Baseline 1) and encoder (`XLM-RoBERTa-large`, Baseline 2) models.

5 Challenge Results

The challenge was officially launched during a dedicated session at SEPLN 2024, held in Valladolid on 26 September 2024, and concluded on 2 February 2025. In total, 26 teams registered for participation, of which 6 successfully completed a full submission, that is, they submitted results for all 10 tasks in ODESIA CORE. These teams were: IXA (Ixa Group from the University of Basque Country EHU), BSC (Language Technologies Laboratory at the Barcelona Supercomputing Centre, Spain), GPLSI (Language Processing and Information Systems Group at the University of Alicante, Spain), UMUTeam (Universidad de Murcia, Spain), IIC (Instituto de Ingeniería del Conocimiento, Spain), and UDA-LIDI (Universidad del Azuay, Ecuador).

Most submissions employed large decoder-only generative LLMs, although the challenge rules allowed encoder-based, hybrid, or retrieval-augmented approaches as long as consistency was maintained. Table 2 shows the macro-average results per task (only the best submission of each team is shown).

The overall ranking was led by Team IXA using **Qwen2.5-14B-Instruct**, achieving a macro-average of 63.06. They were closely followed by the BSC team (62.37) using a powerful custom encoder-based approach (*xlm_roberta_cpt*), demonstrating that discriminative models remain highly competitive against generative AI in this benchmark. The GPLSI team secured the fourth spot

Model	Team	Avg.
Qwen2.5-14B-Instruct	IXA	**63.06**
xlm_roberta_cpt	BSC	62.37
Hermes-3-Llama-3.1-8B	IXA	61.63
Llama_3.1-8B-Instruct	GPLSI	60.12
Qwen2.5-7B	Baseline 1	58.58
xlm-roberta-large	Baseline 2	58.73
XLM-RoBERTa-large-v3	UMUTeam	54.62
Gemma-2B-IT	IXA	54.56
RigoBERTa	IIC	52.64
DeepSeek_Llama3.1	UDA-LIDI	51.63

Table 2: ODESIA Challenge 2024 average results.

(60.12), outperforming the official Qwen2.5-7B baseline.

Breaking down the performance by task (see Table 3) reveals interesting specializations. In the **EXIST** datasets, Team GPLSI's *Llama_3.1-8B* excelled in sexism categorization and identification, achieving the highest scores in EXIST 2022 Task 2 (62.03) and EXIST 2023 Task 1 (70.29). However, for tasks involving the LeWiDi paradigm (EXIST 2023 Tasks 2 and 3), the encoder-based BSC model proved superior, achieving the highest Soft-ICM scores.

Table 3 also highlights a clear dichotomy between task types. In the Biomedical Named Entity Recognition task (**DIANN 2023**), encoder models dominated: the specialized *RigoBERTa* model (Serrano et al., 2022) from Team IIC achieved a remarkable top score of **81.23**, followed by the BSC encoder (77.90), with generative models trailing behind. Conversely, in the Question Answering task (**SQAC 2024**), the generative Qwen2.5-14B (IXA) outperformed all other systems by a wide margin (72.65 vs. 62.70 for the next best), a key factor that ultimately secured IXA's overall victory.

6 Participant Approaches

While six teams successfully completed the full ODESIA-CORE benchmark, this section focuses on the systems for which detailed technical reports were made available for analysis: *IXA_taldea* and *GPLSI*.

The selection of these two approaches offers a scientifically valuable contrast, as they represent distinct architectural paradigms within the LLM landscape. Team IXA employed a **Unified Text-to-Text Framework**, treating all tasks as generative instruction-following problems, whereas Team GPLSI explored a **Retrieval-Augmented Generation (RAG)** pipeline, attempting to leverage external context from training data. In the following subsections, we dissect the engineering decisions, prompting strategies, and architectural constraints adopted by these teams to provide qualitative context to the quantitative results reported above.

6.1 Team IXA: Unified Text-to-Text Framework

The submission [6] by the HiTZ Center - Ixa (University of the Basque Country EHU) leverages auto-regressive LLMs. In particular, they utilize pre-trained instruction-based models trained to generate responses conditioned on task descriptions or user prompts. Consequently, they transform all tasks in the ODESIA Challenge into a chat-style text-to-text format. Their approach is compatible with any auto-regressive text-generation language model, and their code implementation supports arbitrary models out-of-the-box.

6.1.1 Text-to-Text Conversion

All tasks are converted into a chat-style text-to-text format. The team developed a prompt that clearly defines the task and instructs the annotation of a single unlabeled input. For each task, they compiled detailed annotation guidelines describing the task and its labels. These guidelines, ranging from 170 to 300 words, are summarized from task descriptions originally provided in the corresponding papers or task websites.

Subsequently, a set of few-shot examples is randomly sampled from the training split. During both training and inference, new few-shot examples are dynamically generated for each input. This dynamic selection ensures coverage of at least one few-shot example per label in each task. Specifically, they use 3 few-shot examples for SQAC 2024, 8 for DIANN 2023, and 20 for all other tasks. The unlabeled input to be annotated by the model is appended at the end of the prompt.

The tasks' labels are represented using a structured JSON schema. The JSON format provides flexibility, allowing accommodation of diverse label types, and aligns well with the

[6] Code available at `https://github.com/hitz-zentroa/Odesia-Struct`

		EXIST 2022		EXIST 2023			DIPROMATS 2023			DIANN 2023	SQAC 2024
Model	**Team**	**E22-1**	**E22-2**	**E23-1**	**E23-2**	**E23-3**	**DP-1**	**DP-2**	**DP-3**	**DI-1**	**SQ-1**
Qwen2.5-14B-Instruct	IXA	80.27	60.65	67.74	45.52	41.11	**83.60**	55.30	**49.31**	74.42	**72.65**
xlm_roberta_cpt	BSC_models	78.16	60.04	66.93	**46.54**	**43.80**	81.66	**57.56**	48.37	77.90	62.70
Hermes-3-Llama-3.1-8B	IXA	**80.65**	59.07	66.88	45.85	38.63	82.03	55.15	48.45	71.71	67.91
Llama_3.1-8B-Instruct	GPLSI	79.89	**62.03**	**70.29**	45.15	39.30	82.74	53.79	43.83	58.64	65.54
Qwen2.5-7B	Baseline 1	74.90	57.65	66.17	43.57	39.29	82.82	55.72	45.28	64.55	54.99
Xlm roberta large	Baseline 2	76.63	55.93	65.64	44.14	39.95	81.86	53.43	45.27	78.55	45.89
XLM-RoBERTa-large-v3	UMUTeam	74.52	55.40	54.41	43.84	36.09	82.24	54.25	45.81	59.67	40.00
Gemma-2B-IT	IXA	75.48	52.62	62.57	40.12	29.20	81.09	52.83	43.03	61.29	47.38
RigoBERTa	IIC	74.90	59.57	57.30	25.28	40.15	81.33	55.94	46.70	**81.23**	4.04
DeepSeek_Llama3.1	UDA-LIDI	75.86	50.77	61.54	39.45	30.60	75.34	45.25	36.87	47.52	53.12

Table 3: ODESIA Challenge 2024 results across the set of 10 CORE tasks: EXIST 2022 (E22), EXIST 2023 (E23), DIPROMATS 2023 (DP), DIANN 2023 (DI), and SQAC 2024 (SQ).

familiarity of current large language models. For text classification tasks, the output is a list of labels. In the EXIST 2023 task, the model outputs a score per label; thus, they use a dictionary where keys represent label names, and values are floats indicating the scores. The model predicts these scores fully autoregressively as text tokens, without additional handling. For sequence labeling tasks, the output consists of labeled spans grouped by each label category. Finally, for question-answering tasks, the output is represented as a string containing the answer to the question. To encode inputs, they employ each model's default chat template. The prompt is encoded as user input, whereas the corresponding output is encoded as system output. During inference, the model receives only the input prompt along with the token indicating the start of the system response and subsequently generates the predicted labels.

6.1.2 Training

Team IXA adopts a multitask training approach, where a single model is trained jointly across all tasks. Preliminary experiments showed that training dedicated expert models for each task resulted in inferior performance. To balance the data distribution and prevent tasks with larger datasets from dominating the training process, they limit each task to a maximum of 10,000 training examples per epoch. Only the Spanish versions of the datasets were used for training. The standard Next Token Prediction (NTP) loss is used for model training. To ensure that the loss contribution from the guidelines' tokens does not overshadow the actual output tokens, the loss is calculated exclusively on the output tokens. They perform full fine-tuning on all model parameters. Although they experimented with Low-Rank Adaptation (LoRA) (Hu et al., 2021), which reduces VRAM usage and computational resources, the results from LoRA were slightly lower compared to full fine-tuning. DeepSpeed ZeRO3 (Rajbhandari et al., 2020) is utilized to distribute the model parameters, gradients, and optimizer state across eight Nvidia A100 80GB GPUs. They employ the AdamW optimizer (Loshchilov and Hutter, 2019) with a batch size of 64, a learning rate of 5×10^{-6}, and a cosine scheduler with a 10% warm-up ratio. Training is conducted for a total of 3 epochs.

6.1.3 Constrained Decoding

During inference, the model must generate valid JSON outputs containing predictions for each task. To ensure the generated outputs adhere strictly to valid JSON formats and to avoid hallucinations or incomplete annotations, constrained decoding is employed. Each task output format is defined using a Pydantic JSON schema. Throughout the autoregressive token generation process, next-token predictions are constrained exclusively to tokens that comply with the defined JSON schema, setting the probability of non-conforming tokens to zero. To implement this, they utilize the Constrained Decoding functionality provided by the Outlines library (Willard and Louf, 2023). This approach has been successfully applied by IXA team members to other sequence labeling tasks such as QA and Argument Component Detection in the medical domain (Sviridova et al., 2024; García-Ferrero et al., 2024).

6.1.4 Models

They experimented with four distinct models, ranging in size from 2B to 14B pa-

rameters: Gemma (Team, 2024), a 2B-parameter model; Hermes-3-Llama-3.1, an 8B-parameter model (Teknium, Quesnelle, and Guang, 2024) fine-tuned from LLama-3.1 (Grattafiori and others, 2024); and the 14B-parameter version of Qwen2.5 (Qwen et al., 2025).

6.1.5 Performance Analysis

Their approach, utilizing the Qwen2.5 14-billion parameter model, achieves the highest overall performance. They surpass the encoder-only model XLM-RoBERTa, submitted by the BSC_models team, demonstrating that the text-to-text method combined with state-of-the-art LLMs can outperform traditional encoder-only models. However, it is noteworthy that the performance of XLM-RoBERTa remains highly competitive while requiring significantly fewer computational resources.

The Hermes-3-Llama-3.1 model, with 8 billion parameters, performs on average 1.43 points lower than the 14-billion parameter Qwen2.5 model. On the other hand, the smaller Gemma 2-billion parameter model yields significantly lower results, indicating that smaller text-to-text models struggle with structured tasks such as those in the ODESIA CORE benchmark.

This approach achieves superior results in the question answering task by a wide margin. This success is attributed to the intrinsic proficiency of LLMs in question-answering tasks and their substantial context length. Interestingly, in the EXIST 2023 task, the method of autoregressively predicting probability scores as textual values yields competitive results compared to encoder-only models, which directly extract probabilities from classification layers.

The DIANN 2023 task is the one where this approach performs the least effectively compared to other team submissions. This lower performance is attributed to the small size of the training dataset for this specific task, which likely becomes diluted among the larger volumes of data from other tasks during multitask training.

6.2 Team GPLSI: Retrieval-Augmented Generation

The team from the University of Alicante (GPLSI) took a Retrieval-Augmented Generation (RAG) approach[7]. They proposed a system designed to improve task-based response generation by utilizing an instruction model enriched with contextual information retrieval and dynamic prompt construction. Their approach integrates augmented task-specific prompts, a knowledge retrieval mechanism, a modular prompt assembly process, and a response parsing module.

6.2.1 Approach Overview

The system refines a pre-trained model into an instructed model through a sequential pipeline:

1. **Task Augmentation and Prompt Generation:** Predefined tasks undergo augmentation to improve generalizability. Each task is used to generate specific prompts for instruction tuning.
2. **Knowledge Retrieval and Contextualization:** For new task examples, an Information Retrieval module fetches relevant contextual documents from a structured knowledge base to enhance the task context.
3. **Prompt Assembly Module:** Retrieved context is integrated into the prompt. The module supports various strategies (zero-shot, k-shot, chain-of-thought) to adapt to different input structures.
4. **Response Generation and Parsing:** The instructed model generates a response based on the assembled prompt, which is then parsed to ensure alignment with task requirements.

6.2.2 RAG Architecture

The GPLSI team implemented a RAG system (Lewis et al., 2020) where the training sets for each task serve as the knowledge base. This decision was based on findings that using the model's own training data during retrieval enhances performance (Izacard et al., 2022; Borgeaud et al., 2022).

Dataset and Chunking. Unlike conventional RAG implementations that chunk documents, GPLSI treated each training instance as a discrete document. This preserved the full context of the examples, which were generally short enough to be tokenized without fragmentation.

[7] Code is available at `https://github.com/gplsi/Odesia_Challenge`

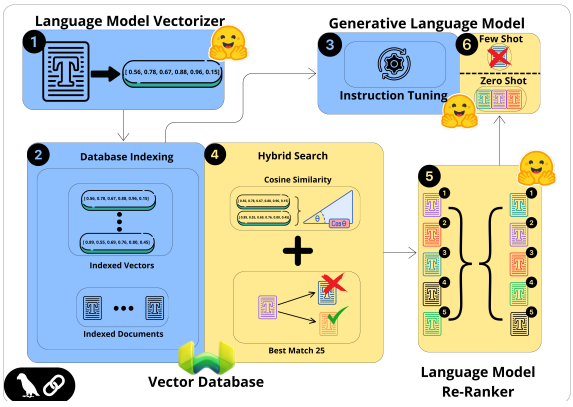

Figure 1: Architecture used for GPLSI's RAG system.

Vector Database and Hybrid Search. The team utilized **Weaviate** (B.V., 2022) for its robust hybrid search capabilities. They implemented a hybrid search strategy combining:

- **Dense Vector Search:** Using cosine similarity with a threshold of 0.6.
- **BM25:** A keyword-based search algorithm (Robertson and Walker, 1994) that scores documents based on Term Frequency and Inverse Document Frequency.

These methods were combined using the `rankedFusion` algorithm (Weaviate, 2024), defaulting to a parameter of 0.75 to favor vector similarity.

Vectorization and Reranking. Embeddings were generated using the `sentence-transformers/all-MiniLM-L6-v2` model (Reimers and Gurevych, 2020). To refine retrieval precision, the top results were reranked using the `BAAI/bge-reranker-v2-m3` model (BAAI, 2025), chosen for its strong multilingual capabilities.

6.2.3 Prompt Assembly

The prompt assembly module constructs structured prompts consisting of `prompt_start` (instructions), `prompt_guide` (few-shot examples), and `prompt_end` (response format). A system prompt is also included to define the model's persona (e.g., "You are an expert linguist...").

Specific modifications were made for certain tasks:

- **DIANN 2023:** Several output formats were tested, ranging from direct BIO tagging (V1) to span extraction (V3) and context-enriched extraction (V4), where RAG was used to identify diseases and provide them as context.
- **EXIST 2023:** Functions were implemented to compute and normalize label distributions for the LeWiDi paradigm, ensuring the output matched the probabilistic format required by the metric.

6.2.4 Experimental Setup and Results

The GPLSI team conducted experiments using **Llama 3.2 (3B)** and **Llama 3.1 (8B)** models, fine-tuned using Fully Sharded Data Parallel (FSDP) on NVIDIA A100 GPUs. They explored ten experimental configurations varying the model size, k-shot value (0, 5, 10), and NER output format (Table 4).

Exp	Model	Few-shot	NER Format
Exp_1	Llama_3.2-3B-Inst	0	V_1
Exp_6	Llama_3.1-8B-Inst	0	V_2
Exp_8	Llama_3.1-8B-Inst	0	V_4
Exp_{10}	Llama_3.1-8B-Inst	10	V_4

Table 4: Selected Experimental Configurations (GPLSI).

Their results revealed that, while the system was designed to leverage retrieval, the inclusion of the RAG module often had a negative effect compared to the pure instruction-tuned baseline. The team concluded that for adapting a model to these specific tasks, starting from a strong instruction-tuned model (0-shot) was more effective than injecting retrieved examples, as the added context often introduced noise or computational bottlenecks without corresponding accuracy gains.

Consequently, their best performing configuration (Exp 6 and 8) relied on the **zero-shot setting** without the RAG component. This setup achieved competitive results, particularly in binary classification tasks like Sexism Detection and Propaganda Identification, and secured the top spot in the Sexism Categorization task.

7 Discussion

The ODESIA Challenge results highlight the transition from discriminative to generative

paradigms in Spanish NLP. Analyzing the leaderboard reveals distinct methodological patterns where different architectures excel in specific domains.

Generative Advantage in QA. The ODESIA Challenge results highlight the transition from discriminative to generative paradigms in Spanish NLP. A critical analysis reveals that Team IXA's overall victory was heavily driven by the **SQAC 2024** (Extractive QA) task. In this specific domain, the generative *Qwen2.5-14B* approach achieved a score of **72.65**, outperforming the best encoder system (62.70) by nearly 10 points. While encoder-based models treat QA as a rigid span-extraction problem (predicting start/end tokens), generative models approach it as natural reading comprehension. This structural advantage compensated for tighter margins in other tasks, effectively acting as the deciding factor in the final macro-average ranking.

Encoders remain the Standard for Sequence Labeling and Soft Classification. Despite the dominance of LLMs, encoder-only architectures demonstrated superior performance in tasks requiring precise span boundaries or probability calibration.

- *NER:* In the disability detection task (DIANN 2023), the encoder-based *RigoBERTa* achieved the highest score (**81.23**), significantly beating the best LLM (74.42). This confirms that token-classification heads remain more precise than generative prompts for fine-grained entity extraction.
- *Uncertainty:* In EXIST 2023, which involves soft labels (predicting annotator disagreement), the *BSC XLM-R* encoder outperformed the largest LLMs in the more complex tasks (T2 and T3). A possible explanation is that encoders, capable of directly optimizing loss functions against probability distributions, appear better suited for modeling uncertainty than LLMs, which are inherently designed for next-token prediction.

LLMs Excel at Semantic Judgment. In high-level binary classification tasks requiring semantic reasoning—specifically **Propaganda Identification** and **Sexism Detection**—instruction-tuned LLMs consistently edged out encoders. The massive world knowledge embedded in models like Qwen or Llama provides an advantage in detecting complex, high-level linguistic phenomena compared to encoders trained strictly on fine-tuning data.

Instruction Tuning vs. RAG. The experiments by Team GPLSI provided a valuable negative result regarding RAG for these tasks. While they implemented a robust RAG pipeline, they found that direct instruction tuning (0-shot) was more effective. RAG adds complexity that appears beneficial primarily when external knowledge is strictly required; for self-contained classification tasks, retrieving similar examples often introduces noise or computational bottlenecks without corresponding accuracy gains.

Scale and Cross-Lingual Transfer. Performance scaled predictably with size ($14B > 8B > 7B > 2B$). Moreover, the victory of *Qwen2.5*—a model with strong multilingual capabilities but primarily Chinese/English roots—suggests that cross-lingual transfer in modern LLMs is exceptionally robust. This potentially diminishes the need for Spanish-specific pre-training (Agerri and Agirre, 2023), as general-purpose multilingual LLMs show high adaptability to Spanish cultural and linguistic nuances when properly instruction-tuned.

8 Conclusion

The ODESIA Challenge 2024 successfully benchmarked the state of Spanish NLP, revealing a transition period where generative and discriminative paradigms coexist. While massive generative models offer superior in a wide range of tasks, specialized encoders remain efficient powerhouses in many classification tasks at a much lower economical and environmental cost. Future research lines should prioritize the refinement of these discriminative models for high-precision (e.g., complex hierarchical soft classification) and resource-constrained scenarios, positioning them as indispensable complements to generative AI rather than obsolete predecessors.

Acknowledgements

This work has been funded by the European Union - NextGenerationEU through the 'Recovery, Transformation and Re-

silience Plan', by the *Ministerio para la transformación digital y de la función pública* and by UNED via cooperation agreement C039-21OT. However, the views and opinions expressed are solely those of the author(s) and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them. HiTZ's authors are partially funded by the following MCIN/AEI/10.13039/501100011033 projects: (i) DeepKnowledge (PID2021-127777OB-C21) and by ERDF, EU; (ii) DeepMinor (CNS2023-144375) and European Union NextGenerationEU/PRTR. GPLSI´s authors are partially supported by Project COOLANG.TRIVIAL (PID2021-122263OB-C22), HEART-NLP.UA (PID2024-156263OB-C22), FEDER granted funding for CIDEGENT (CIDEXG/2023/13), and by the Ministerio para la Transformación Digital y de la Función Pública, funded by the EU – NextGenerationEU, within the framework of the project "Desarrollo de Modelos ALIA", under the "Plan Nacional de Tecnologías del Lenguaje - ENIA 2024 y PRTR, NextGeneration EU, Resol. SEDIA 19.08.2024."

References

Agerri, R. and E. Agirre. 2023. Lessons learned from the evaluation of Spanish Language Models. *Procesamiento del Lenguaje Natural*, 70:157–170.

Amigo, E. and A. Delgado. 2022. Evaluating Extreme Hierarchical Multi-label Classification. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5809–5819, Dublin, Ireland, May. Association for Computational Linguistics.

BAAI. 2025. Bge-reranker-v2-m3: A lightweight reranker model with strong multilingual capabilities. Hugging Face Model Hub, 2.

Benito-Santos, A., A. Ghajari, and V. Fresno. 2025. Robust Estimation of Population-Level Effects in Repeated-Measures NLP Experimental Designs. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 33076–33089, Vienna, Austria, July. Association for Computational Linguistics.

Borgeaud, S., A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G. van den Driessche, J.-B. Lespiau, B. Damoc, A. Clark, D. de Las Casas, A. Guy, J. Menick, R. Ring, T. Hennigan, S. Huang, L. Maggiore, C. Jones, A. Cassirer, A. Brock, M. Paganini, G. Irving, O. Vinyals, S. Osindero, K. Simonyan, J. W. Rae, E. Elsen, and L. Sifre. 2022. Improving language models by retrieving from trillions of tokens.

B.V., S. T. 2022. Weaviate: A cloud-native, modular, real-time vector search engine. `https://weaviate.io`. Version X.X.

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July. Association for Computational Linguistics.

Fabregat, H., J. Martínez-Romo, and L. Araujo. 2018. Overview of the DIANN task: Disability annotation task. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018*, volume 2150 of *CEUR Workshop Proceedings*, pages 1–14. CEUR-WS.org.

García-Ferrero, I., R. Agerri, A. Atutxa Salazar, E. Cabrio, I. de la Iglesia, A. Lavelli, B. Magnini, B. Molinet, J. Ramirez-Romero, G. Rigau, J. M. Villa-Gonzalez, S. Villata, and A. Zaninello. 2024. MedMT5: An open-source multilingual text-to-text LLM for the medical domain. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, editors, *Proceedings of the 2024 Joint International Conference on*

Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 11165–11177, Torino, Italia, May. ELRA and ICCL.

Grattafiori, A. et al. 2024. The llama 3 herd of models.

Gutiérrez-Fandiño, A., J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, C. Armentano-Oller, C. Rodriguez-Penagos, A. Gonzalez-Agirre, and M. Villegas. 2022. MarIA: Spanish Language Models. *Procesamiento del Lenguaje Natural*, 68(0):39–60, March.

Hu, E. J., Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and W. Chen. 2021. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685.

Izacard, G., P. Lewis, M. Lomeli, L. Hosseini, F. Petroni, T. Schick, J. Dwivedi-Yu, A. Joulin, S. Riedel, and E. Grave. 2022. Atlas: Few-shot learning with retrieval augmented language models.

Lewis, P., E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Loshchilov, I. and F. Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Malmasi, S., A. Fang, B. Fetahu, S. Kar, and O. Rokhlenko. 2022. Semeval-2022 task 11: Multilingual complex named entity recognition (multiconer). In *Proceedings of the 16th international workshop on semantic evaluation (SemEval-2022)*, pages 1412–1437.

Moral, P., G. Marco, J. Gonzalo, J. Carrillo-de Albornoz, and I. Gonzalo-Verdugo. 2023. Overview of DIPROMATS 2023: automatic detection and characterization of propaganda techniques in messages from diplomats and authorities of world powers. *Revista Procesamiento del Lenguaje Natural*, 71:397–407.

Plaza, L., J. Carrillo-de-Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, and P. Rosso. 2023. Overview of EXIST 2023 – Learning with Disagreement for Sexism Identification and Characterization. In A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, and N. Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 316–342, Cham. Springer Nature Switzerland.

Qwen, :, A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Tang, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, and Z. Qiu. 2025. Qwen2.5 technical report.

Rajbhandari, S., J. Rasley, O. Ruwase, and Y. He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE.

Reimers, N. and I. Gurevych. 2020. sentence-transformers/all-minilm-l6-v2. `https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2`. Accessed: 17 March 2025. Recognized for balancing computational cost and performance in sentence similarity tasks.

Robertson, S. and S. Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 232–241. ACM.

Rodríguez-Sánchez, F., J. Carrillo-de Albornoz, L. Plaza, D. Spina, J. Gonzalo, and P. Rosso. 2022. Overview of EXIST 2022: sexism identification in social networks. *Procesamiento del Lenguaje Natural*, 69:229–240.

Ruder, S. 2021. Challenges and Opportunities in NLP Benchmarking. http://ruder.io/nlp-benchmarking.

Sainz, O., I. García-Ferrero, A. Jacovi, J. Ander Campos, Y. Elazar, Y. Agirre, Eneko anxd Goldberg, W.-L. Chen, J. Chim, L. Choshen, L. D'Amico-Wong, M. Dell, R.-Z. Fan, S. Golchin, Y. Li, P. Liu, B. Pahwa, A. Prabhu, S. Sharma, E. Silcock, K. Solonko, D. Stap, M. Surdeanu, Y.-M. Tseng, V. Udandarao, Z. Wang, R. Xu, and J. Yang. 2024. Data Contamination Report from the 2024 CONDA Shared Task. In O. Sainz, I. García Ferrero, E. Agirre, J. Ander Campos, A. Jacovi, Y. Elazar, and Y. Goldberg, editors, *Proceedings of the 1st Workshop on Data Contamination (CONDA)*, pages 41–56, Bangkok, Thailand, August. Association for Computational Linguistics.

Sánchez Salido, E., R. Morante, J. Gonzalo, G. Marco, J. Carrillo-de-Albornoz, L. Plaza, E. Amigo, A. F. García, A. Benito-Santos, A. Ghajari Espinosa, and V. Fresno. 2025. Bilingual Evaluation of Language Models on General Knowledge in University Entrance Exams with Minimal Contamination. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6184–6200, Abu Dhabi, UAE, January. Association for Computational Linguistics.

Schwenk, H. and X. Li. 2018. A corpus for multilingual document classification in eight languages. In N. C. C. chair), K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may. European Language Resources Association (ELRA).

Serrano, A. V., G. G. Subies, H. M. Zamorano, N. A. Garcia, D. Samy, D. B. Sanchez, A. M. Sandoval, M. G. Nieto, and A. B. Jimenez. 2022. RigoBERTa: A State-of-the-Art Language Model For Spanish, June.

Sviridova, E., A. Yeginbergen, A. Estarrona, E. Cabrio, S. Villata, and R. Agerri. 2024. CasiMedicos-arg: A medical question answering dataset annotated with explanatory argumentative structures. In *EMNLP 2024*, pages 18463–18475.

Team, G. 2024. Gemma: Open models based on gemini research and technology.

Teknium, R., J. Quesnelle, and C. Guang. 2024. Hermes 3 technical report.

Uma, A., T. Fornaciari, A. Dumitrache, T. Miller, J. Chamberlain, B. Plank, E. Simpson, and M. Poesio. 2021a. SemEval-2021 task 12: Learning with disagreements. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 338–347, Online, August. Association for Computational Linguistics.

Uma, A. N., T. Fornaciari, D. Hovy, S. Paun, B. Plank, and M. Poesio. 2021b. Learning from Disagreement: A Survey. *Journal of Artificial Intelligence Research*, 72:1385–1470, December.

Weaviate. 2024. Fusion algorithm: rankedfusion for hybrid search in weaviate. https://weaviate.io/learn/knowledgecards/fusion-algorithm. Accessed: 17 March 2025.

Willard, B. T. and R. Louf. 2023. Efficient guided generation for large language models.

A Annex 1

	EXIST 2022 1	EXIST 2022 2	DIPROMATS 1	DIPROMATS 2	DIPROMATS 3	DIANN 2023	EXIST 2023 1	EXIST 2023 2	EXIST 2023 3	SQAC 2024	Average
Encoders											
xlm-roberta-base	73.95	49.97	78.94	45.04	26.68	78.19	62.36	42.45	31.95	36.91	52.64
xlm-roberta-large	76.63	55.93	81.86	**53.43**	45.27	**78.55**	65.64	44.14	39.95	45.89	**58.73**
bert-base-multil-cased	72.22	46.93	78.21	42.31	25.62	75.92	61.36	39.17	33.26	32.25	50.73
distilbert-base-multil-cased	72.22	46.69	75.07	40.36	22.22	68.68	58.51	38.23	28.74	22.07	47.28
roberta-base-bne	73.56	55.54	81.49	49.06	29.44	71.69	65.31	41.73	36.88	40.61	54.53
roberta-large-bne	72.41	56.68	81.77	51.73	38.94	67.57	66.71	42.37	37.98	46.40	56.26
bertin-roberta-base-sp	72.80	49.41	75.96	25.32	25.00	68.77	64.65	41.46	33.31	41.72	49.84
bert-base-sp-wwm-cased	71.46	53.70	79.16	48.74	29.31	74.78	63.26	41.82	37.38	41.18	54.08
distillbert-base-sp-uncased	72.03	51.18	77.08	41.98	17.82	65.31	61.28	41.60	33.24	24.84	48.64
ixambert-base-cased	67.43	48.75	76.66	37.96	05.43	75.80	61.17	38.90	34.12	35.70	48.19
Decoders											
Qwen2.5-7B	**78.16**	55.26	**82.82**	**55.72**	**45.28**	64.55	66.17	43.57	39.29	**54.99**	58.58
Llama-3.1-8B	76.63	**56.82**	81.69	53.85	44.19	63.19	**67.40**	**46.40**	**40.30**	54.12	58.46
Mistral-7B-v03	77.01	54.76	81.04	52.79	43.38	64.08	64.29	44.47	37.74	48.55	56.81

Table 5: Results on the Core Tasks. F1 metrics have been used for all tasks except for EXIST 2023, evaluated with ICM-soft normalized, and DIPROMATS (normalized ICM).