La Leaderboard: A Large Language Model Leaderboard for Spanish Varieties and Languages of Spain and Latin America

Anonymous ACL submission

Abstract

Leaderboards showcase the current capabilities and limitations of Large Language Models (LLMs). To motivate the development of LLMs that represent the linguistic and cultural diversity of the Spanish-speaking community, we present La Leaderboard , the first open-source leaderboard to evaluate generative LLMs in languages and language varieties of Spain and Latin America. La Leaderboard is a community-driven project that aims to establish an evaluation standard for everyone interested in developing LLMs for the Spanish-speaking community. This initial version combines 66 datasets in Catalan, Basque, Galician, and different Spanish varieties, showcasing the evaluation results of 50 models. To encourage community-driven development of leaderboards in other languages, we explain our methodology, including guidance on selecting the most suitable evaluation setup for each downstream task. In particular, we provide a rationale for using fewer few-shot examples than typically found in the literature, aiming to reduce environmental impact and facilitate access to reproducible results for a broader research community.

1 Introduction

The evaluation of multilingual Large Language Models (LLMs) is challenging. LLMs are expected to perform a large variety of tasks, from problem-solving to text summarization, all in multiple languages (Guo et al., 2023). In this context, leaderboards have emerged, through transparent and standardized frameworks, as one of the standard approaches for evaluating and comparing LLMs. As we cannot improve what we cannot measure, it is important to develop leaderboards that enable a more comprehensive evaluation of LLMs across linguistic boundaries, contributing to the development of culturally aware AI systems that can serve diverse global linguistic communities.

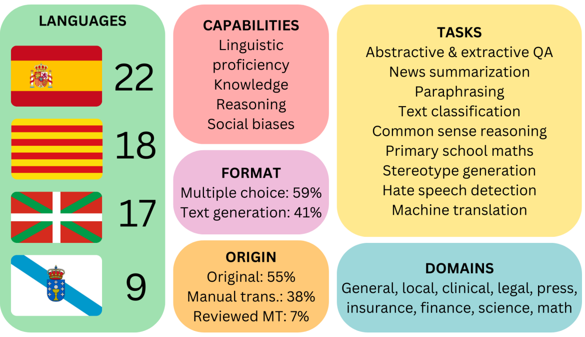

Figure 1: Summary of the evaluation datasets included in La Leaderboard . Disclaimer: A country does not represent a language; flags are used for simplicity.

Spanish is one of the most spoken languages worldwide, with more than 600 million speakers (Fernández and Mella, 2024). It is the predominant language in 21 countries, where it coexists with other languages. Many people use Spanish and the local language in their daily activities. Spain has four official languages: Spanish, Catalan, Basque, and Galician. While Catalan and Galician are Romance languages closely related to Spanish, Basque is one of the world's few language isolates (Campbell, 2010). In Latin America (LATAM), there are hundreds of indigenous languages, such as Guaraní and Náhuatl, which have influenced local Spanish varieties (Lustig, 1996). From a sociolinguistic point of view, this creates a unique scenario for multilingual LLM evaluation. Moreover, knowing which LLMs perform best in these languages can have deep implications for multilingual communication (Strassel and Tracey, 2016).

Existing leaderboards predominantly focus on English or a small set of high-resource languages (Fourrier et al., 2024; Mialon et al., 2023; Pal et al., 2024; Contributors, 2023). While Spanish is often included in multilingual leaderboards, evaluation datasets are typically limited and translated, either by machines (Barth et al., 2024), failing to capture the linguistic richness of the language (Plaza et al.,

2024) or by humans[1], still failing to represent the target culture (Singh et al., 2024). Moreover, despite the growing presence of LLMs in multilingual settings, no leaderboard currently evaluates a combination of languages spoken in Spain and Latin America. This lack of representation limits the development of models that can truly serve these communities (Mager et al., 2018).

To address this gap, we introduce LA LEADERBOARD ,[2] the first open-source leaderboard designed to evaluate generative LLMs based on the needs of the Spanish-speaking community. Beyond the initial set of languages that includes Spanish and the official languages of Spain (Catalan, Basque, and Galician), LA LEADERBOARD is designed to evolve, gradually expanding to encompass more languages and linguistic varieties, ensuring it reflects the rich diversity of the global community. This new leaderboard consists of a diverse set of evaluation tasks (see Figure 1) written in a way that reflects the nuances and actual usage of the target languages. It is a community-driven initiative aiming to foster the development of LLMs that better represent the linguistic and cultural diversity of the Spanish-speaking world. We share our approach to inspire other linguistic communities to create similar leaderboards.

The main contributions of this work are:

- We present the community-based methodology used to create the first open-source leaderboard for evaluating generative LLMs in Spanish and the official languages of Spain, with a scalable framework designed to include more languages and language varieties over time.

- We introduce a logical and resource-efficient approach to few-shot configurations, enabling accessible and reproducible evaluations for the wider community.

- We provide a comprehensive analysis of state-of-the-art (SOTA) LLMs, providing insights into their strengths and limitations in Spanish, Catalan, Basque, and Galician.

By aiming to address the linguistic and cultural diversity of Spain and LATAM, LA LEADERBOARD aims to set a new standard for multilingual LLM evaluation. Our goal is to encourage the development of models that are not only linguistically competent but also culturally aware, driving progress in Natural Language Processing (NLP) for the benefit of our whole community.

2 Related Work

Benchmarks Several benchmarks have been developed to evaluate the performance of LLMs in tasks like language understanding (Wang et al., 2019), general knowledge (Hendrycks et al., 2021a), reasoning (Sakaguchi et al., 2019), or mathematical problem solving (Hendrycks et al., 2021b). There are also efforts to develop holistic benchmarks or evaluation suites that provide a comprehensive evaluation of different capabilities of LLMs (Liang et al., 2023; Gao et al., 2021; Fourrier et al., 2023, 2024; Srivastava and et al, 2023). However, as LLM capabilities improve continuously, benchmarks soon become saturated (Kiela et al., 2021) or contaminated (Xu et al., 2024), and it is necessary to continuously develop new tasks to assess models' capabilities. For example, MMLU-Pro (Wang et al., 2024b) has been proposed to replace MMLU, and shortly after that, MMLU-Pro+ was proposed as an additional improvement (Taghanaki et al., 2024), with even harder tests being developed (Phan et al., 2025).

Multilingual and multicultural benchmarks LLMs are now trained in multiple high-resource languages at the same time (Ali et al., 2024; Martins et al., 2024; Qwen Team, 2024; Jiang et al., 2023), which means that the benchmarks must reflect this linguistic diversity. A common approach is machine translating English tests (Holtermann et al., 2024; OpenAI, 2023). However, translation errors may add noise to the results, making them less reliable (Plaza et al., 2024). Furthermore, each language has its nuances, preferred styles, and cultural background, which unrevised machine translation may fail to capture (Plaza-del-Arco et al., 2020; Singh et al., 2024). Ideally, specific test sets should be originally written in the target language or manually adapted (Nangia et al., 2020) to capture the richness and cultural and linguistic subtleties associated with it. This is what is slowly happening with language-specific (Mercorio et al., 2024; Quercia et al., 2024) and multilingual culture-aware (Romanou et al., 2024; Myung et al., 2025; Romero et al., 2024) benchmarks released recently.

[1] https://hf.co/datasets/openai/MMMLU

[2] Link removed for review.

Leaderboards Benchmarks are pieces of the LLM evaluation puzzle that provide valuable but fragmented information on their performance. Leaderboards and arenas use these evaluation sets to compare the performance of LLMs in a neutral, third-party manner through automatic evaluations (Mialon et al., 2023) or human judgments (Chiang et al., 2024). On some community-oriented leaderboards (Fourrier et al., 2024), anyone can submit their LLMs for evaluation, and the tools, tests, and results are open, allowing for reproducibility. This represents a good way to drive progress in LLM development by enabling people with limited compute to compare their models to the current SOTA.

Multilingual leaderboards Leaderboards exhibit the same shortcomings as benchmarks when evaluating languages other than English. To address this problem, specific leaderboards are being developed in different languages such as Italian (Mercorio et al., 2024), Korean (Kim et al., 2024), Chinese (Contributors, 2023), Arabic (Elfilali et al., 2024) or Polish (Jassem et al., 2025).

Spanish leaderboards Focusing on the Spanish language, the ODESIA leaderboard[3] by UNED NLP features 15 bilingual Spanish-English discriminative tasks. While submissions are open, the evaluation datasets are private, avoiding task contamination (Salido et al., 2025) but making it impossible to reproduce the results. Regarding text generation, Spanish is represented in the Chatbot Arena, which features a dedicated category, and in SCALE's private leaderboard.[4] However, both exclusively evaluate a fixed set of models. The only existing leaderboard including a language from Spain or Latin America other than Spanish is CLUB,[5] developed by the BSC as part of the AINA Project, which combines 8 Catalan datasets.

In this work, we present the methodology used to create a comprehensive, fully open-source leaderboard for languages and language varieties from Spain and Latin America that assesses different capabilities of generative models, including domain knowledge, information extraction, linguistic proficiency, and ethical aspects. La Leaderboard aims to serve as a reference for the Spanish-speaking scientific community, fostering the development of more robust and culturally adequate LLMs.

3 La Leaderboard

La Leaderboard is a community-driven initiative that brings together 66 datasets in Spanish, Catalan, Basque and Galician, covering diverse tasks and domains. Public since September 23, 2024, in four months La Leaderboard has received over 15,000 visits and currently showcases evaluation results from 50 models.

3.1 Data Collection

Most of the datasets in La Leaderboard were donated by 12 research groups. Initially, these contributions were received through a publicly shared Google Form (Appendix E) or direct outreach. In particular, 7 datasets were specifically created for La Leaderboard (AQuAS, ClinTreatES, ClinDiagnosES, HumorQA, SpaLawEx, TELEIA, and RAGQuAS). We also included widely used open-source benchmarks such as Belebele.

La Leaderboard keeps expanding with dataset contributions such as CONAN-EUS and VeritasQA. These new connections are bidirectional: we actively share this initiative in relevant conferences and reach out to research groups, while others contact us upon discovering La Leaderboard . Beyond collecting existing datasets, we are also fostering collaborations to enhance the representation of languages and linguistic varieties across Latin America. Specifically, we are going to launch a community hackathon to create a benchmark to evaluate cultural adequacy across Spanish-speaking countries. We have also established collaborations to generate benchmarks for indigenous languages (AmericasBench, TraduLATAM, and VocesOriginarias).

To thank research groups for their donations, we include in La Leaderboard 's interface the corresponding logo and dataset citation. Moreover, the dataset authors are acknowledged in this paper.

3.2 Task Construction

3.2.1 Datasets

Including diverse evaluation datasets is essential for building a comprehensive leaderboard. This section discusses the key axes that guided their selection. Table 1 enumerates the datasets organized by language and task type, while Table 2 shows the upcoming datasets that have been recently donated

[3] https://leaderboard.odesia.uned.es
[4] https://scale.com/leaderboard/spanish
[5] https://club.aina.bsc.es

and not yet evaluated. In Appendix A, we provide the citations and further details about the datasets, including origin and domain.

Languages La Leaderboard contains 22 evaluation datasets in Spanish, including the varieties of Spain, Mexico, Argentina, Chile, and Uruguay. It also gathers datasets in all the official languages of Spain, with 18 datasets in Catalan, 17 in Basque, and 9 in Galician.

Origin We aim to evaluate models with high-quality datasets that reflect the cultural and linguistic idiosyncrasies of each language. For this reason, we only include datasets that have been annotated or revised by at least one native speaker of the language. We prioritize the inclusion of datasets originally created in the language they evaluate, which constitute 55% of the leaderboard. When this is not possible and translation is required, we prioritize datasets translated by human professionals. Not only does this prevent the loss of linguistic nuances that happens with machine translation (Plaza et al., 2024), but it also allows translators to adapt the text to the target culture (Nangia et al., 2020) and to identify errors in the source datasets and ensure that no extra hints regarding the answer are given in the input prompt (Baucells et al., 2025). In La Leaderboard , 38% of the datasets have been manually translated from an existing English benchmark. We also acknowledge that, given the low-resource nature of some languages we cover, machine translation is more affordable than human translation. However, we only include such datasets if the automatic translation was comprehensively reviewed by a person proficient in the target language. Only 7% of the datasets in La Leaderboard are manual reviews of machine-translated datasets.

Format The multiple-choice question-answering (MCQA) format is widely used for automatic evaluations due to its simplicity. Thus, MCQA is the format of 59% of the tasks included in La Leaderboard . We acknowledge that the literature has identified some issues with MCQA tasks, such as models' sensitivity to answer order (Pezeshkpour and Hruschka, 2024; Mina et al., 2025) or lack of task understanding (Khatun and Brown, 2024). Moreover, some suggest that this type of task does not reflect the actual models' responses and capabilities (Li et al., 2024; Wang et al., 2024a). To address this issue, we also include text generation tasks, such as summarization, evaluated using NoticIA for Spanish, caBreu for Catalan, and Summarization-GL for Galician. We evaluate long-form question-answering (LFQA) in Spanish using the AQuAS and RagQuAS datasets. Finally, we assess counter-narrative generation with RefutES in Spanish and CONAN-EUS in Basque.

Domains La Leaderboard includes well-known generalist datasets aimed at evaluating a model's capability to understand and complete a task, such as Belebele, WNLI and XStoryCloze, among others. We also include evaluation datasets focused on truthfulness assessment, such as VeritasQA and the Galician translation of TruthfulQA. There are, in addition, several domains represented in La Leaderboard , such as the medical (e.g., ClinTreatES), legal (e.g., SpaLawEx), political (e.g., VaxxStance), and press (e.g., caBreu, NoticIA). We also include ethics-oriented datasets, evaluating stereotype generation in Spanish and Catalan with CrowsPairs and alignment with ecolinguistic values using H4rmonyEval.

Tasks The types of tasks chosen for our leaderboard extend those usually included in well-known leaderboards (e.g., reasoning, natural language inference, question answering or summarization) to other task types for which high-quality datasets exist in our target languages (e.g., counter-narrative generation or linguistic acceptability). For consistent performance comparisons across languages, we prioritize tasks available in multiple languages.

3.2.2 Metrics

The MCQA tasks are evaluated by measuring the logarithmic probabilities (LogProbs) of models' outputs among a restricted list of options. For text generation tasks, we compare the expected (*gold-standard*) and given responses using various metrics depending on the original authors' implementation, including BLEU (Papineni et al., 2002), ROUGE (Lin, 2004) and Semantic Answer Similarity (SAS, Risch et al., 2021). Furthermore, following the recent trend of evaluating text generation tasks using LLMs, we are adapting an automated Judge-LLM metric from Zubiaga et al. (2024). Since SAS and LLM-based metrics are not currently supported in the evaluation suite we use, the LM Evaluation Harness (Gao et al., 2021), we implement them in our open-source fork.[6]

[6] Link removed for review.

Task Type	**Spanish**	**Catalan**	**Basque**	**Galician**
Common-sense reasoning	copa_es xstorycloze_es	copa_ca xstorycloze_ca	xcopa_eu xstorycloze_eu	–
Linguistic acceptability	escola	catcola	–	galcola
Math	mgsm_direct_es	mgsm_direct_ca	mgsm_direct_eu	mgsm_direct_gl
NLI	wnli_es xnli_es	teca wnli_ca xnli_ca	qnli_eu wnli_eu xnli_eu	–
Paraphrasing	paws_es parafrases_sushi	parafraseja paws_ca	parafrases_gl paws_gl	
Question answering	aquas clindiagnoses clintreates spalawex teleia ragquas xquad_es	arc_ca catalanqa coqcat openbookqa_ca piqa_ca siqa_ca xquad_ca	bertaqa eus_exams eus_proficiency eus_trivia	openbookqa_gl
Reading comprehension	belebele_spa_Latn	belebele_cat_Latn	belebele_eus_Latn eus_reading	belebele_glg_Latn
Ethics	crows_pairs_es	crows_pairs_ca	–	–
Summarization	noticia xlsum_es	cabreu	–	summarization_gl
Text classification	humorqa fake_news_es offendes	catalonia_ independence	bec2016_eu	
Adaptation	phrases_es	phrases_ca	–	–

Table 1: Datasets of La Leaderboard as of February 2025 organized by task type and language.

Task Type	**Dataset**	**Languages**
Common-sense reasoning	*xstorycloze_gl*	Galician
Counter-narrative generation	*conan_eus/mt_es* *refutes*	Basque, Spanish Spanish
Question answering	*paes_cl* *voces_originarias* *medexpqa* *quales*	Spanish Aymara, Gurarani, Tehuelche, Náhuatl, Quechua Spanish Spanish
Natural language inference	*americasnlp_nli* *meta4xnli*	Aymara, Asháninka, Bribri, Guaraní, Náhuatl, Otomí, Quechua, Rarámuri, Shipibo-Konibo, Wixarika Spanish
Ethics	*h4rmony_eval*	Spanish
Text classification	*haha*	Spanish
Translation	*flores* *americasnlp_mt* *tradu_latam*	Spanish, Catalan, Basque, Galician Spanish, Aymara, Asháninka, Bribri, Guaraní, Náhuatl, Otomí, Quechua, Rarámuri, Shipibo-Konibo, Wixarika Spanish, Aymara, Gurarani, Tehuelche, Náhuatl, Quechua
Truthfulness	*truthfulqa* *veritasqa*	Spanish, Catalan, Basque, Galician Spanish, Catalan, Galician

Table 2: Datasets that have been recently donated to La Leaderboard and are not yet included in the evaluation results, including benchmarks involving American Indigenous languages.

3.3 Code Bases

3.3.1 Backend

We acknowledge the cost of running evaluations and want to ensure that any researcher or developer can compare their models to the state-of-the-art and follow their evolution. This is why submitting a model for evaluation is open to the whole community. Once a model has been added to the evaluation queue, the last commit of the model is stored for reproducibility and to enable future comparisons of different versions. The results from the LM Evaluation Harness (Gao et al., 2021) are normalized according to the following formula:

$$\text{normalized\_value} = \frac{\text{raw\_value} - \text{random\_baseline}}{\text{max\_value} - \text{random\_baseline}} \tag{1}$$

where $random\_baseline$ is 0 for generative tasks and $1/n$ for MCQA tasks with n choices.

3.3.2 Frontend

The implementation of La Leaderboard is based on the HuggingFace leaderboard template.[7] The frontend is developed using Gradio (Abid et al., 2019) and presents the evaluation results categorized by language. To ensure transparency and reproducibility, we share the evaluation command and normalization formula. To bring the tool closer to the community, the information and submission guidelines are available in English and Spanish.

3.3.3 License

Since we want to motivate other communities to create their own, La Leaderboard is published under the permissive Apache 2.0 license.[8]

3.4 Efficiency Considerations

3.4.1 Number of Few-Shot Examples

Recent literature reveals significant inconsistency in the number of examples (shots) used when evaluating large language models (LLMs). While early research demonstrated notable performance improvements with 3-5 in-context examples (Brown et al., 2020), current evaluation practices vary considerably across different models and benchmarks. For instance, the Open LLM Leaderboard employs 0-5 shots depending on the task, Mistral-7B generally follows this range with an exception of 8 shots for GSM8K (Cobbe et al., 2021), and Llama 3 and OLMo models focus primarily on zero-shot evaluation. In contrast, Gemini models use a broader range of 0-10 shots, including "variable-shot" configurations. This variation extends to language-specific models, with Salamandra[9] and Latxa (Etxaniz et al., 2024) families using different shot configurations in their evaluations, typically ranging from 0 to 5 shots.

Given this myriad of options, when choosing the number of shots to use in La Leaderboard , we take into consideration the following aspects:

A. Base vs. instruct models The number of shots should allow for a fair evaluation of the base models without helping instruct models too much. Also, the availability of structured datasets in specific evaluation formats—such as MCQA—is very low in mid- and low-resource languages. This means that models trained on English-heavy corpora are more likely to have encountered these structured formats in English than in other languages, potentially biasing their performance.

B. Cognitive bias Models suffer from cognitive bias depending on the order and options presented as few-shots (Zhao et al., 2021; Pezeshkpour and Hruschka, 2024; Mina et al., 2025). Thus, we ensure that, in MCQA tasks, all possible correct options are included in the in-context learning instances. For example, in an MCQA task with four possible answers, we evaluate on a 4-shot setting, with each shot showing one of the four options as correct, in random order. This is done unless it interferes with item A.

C. Context windows The context window limitations of language models vary significantly based on hardware constraints and architectural choices, affecting their ability to process long-form tasks such as summarization and reading comprehension. For example, while the Spanish government's 40B-parameter ALIA model[10] operates with a 4,096-token context window, Meta's Llama 3.2 1B model can handle up to 128K tokens[11]. To ensure fair evaluation across models with different context window capacities, few-shot examples are employed with a maximum limit of 2,048 tokens, following the methodology established in previous research on LLM analysis (Biderman et al., 2023).

D. Prompt format The evaluation methodology employed task-specific prompts from the LM

[7] https://hf.co/spaces/demo-leaderboard-backend/leaderboard

[8] https://www.apache.org/licenses/LICENSE-2.0

[9] https://hf.co/BSC-LT/salamandra-7b-instruct

[10] https://hf.co/BSC-LT/ALIA-40b

[11] https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/

Model	AVG	ES	CA	EU	GL
Qwen2.5-32B-IT-GPTQ-Int4	**55.65**	**64.06**	56.80	49.23	**52.52**
gemma-2-9b-it	54.90	61.69	57.30	**54.13**	46.49
gemma-2-9b	54.80	57.21	**59.60**	53.80	48.58
Qwen2.5-14B-IT-GPTQ-Int8	53.96	60.59	54.08	49.05	52.13
Meta-Llama-3.1-8B-IT	52.74	59.03	57.01	49.87	45.07
Qwen2.5-7B	51.35	58.79	57.28	42.51	46.82
Meta-Llama-3.1-8B	50.98	55.62	56.52	46.90	44.90
EuroLLM-9B	49.40	55.00	57.32	38.92	46.36
aya-expanse-8b	48.70	55.42	53.99	41.99	43.38
Yi-1.5-9B	48.37	54.51	54.17	40.36	44.44
occiglot-7b-eu5	48.27	55.02	53.71	38.73	45.62
EuroLLM-9B-IT	48.16	57.21	52.96	38.00	44.47
salamandra-7b-instruct	48.12	51.41	53.22	46.19	41.65
salamandra-7b	47.99	52.17	54.13	45.80	39.88
Qwen2.5-7B-IT	47.54	57.46	48.20	41.36	43.13

Table 3: Average results for the top 15 models, overall and per language. Full list available in Figure 3. Target language-optimized models are highlighted in bold.

Evaluation Harness, with new prompts created for previously unimplemented tasks following established formats and validated by dataset authors. The number of few-shots varied based on prompt complexity: convoluted prompts (e.g., paraphrasing with PAWS and reasoning with COPA) used 3 in-context examples to allow models to understand the task while complying with items A and C (Brown et al., 2020); straightforward question-answering tasks employed 2-shot evaluation, while tasks with explicit, naturally structured prompts (like ClinDiagnosES and NoticIA) and those evaluating sentence continuation probability (e.g., XStoryCloze) were conducted using 0-shot evaluation.

3.4.2 Measuring Model Efficiency

The evaluation was performed using two NVIDIA H100 GPUs with Hopper architecture and 64 GB of HBM memory in the MareNostrum 5 High-Performance Computer,[12] maintaining identical configurations across instances to ensure consistent measurements. Performance metrics included task execution time and energy consumption, tracked using the Energy Aware Runtime (EAR) package,[13] with all tasks running at a batch size of 1.

Task execution duration, which includes token prediction time, response length, and tokenizer efficiency, was measured to assess model speed for time-sensitive applications. The duration of task execution is influenced by multiple factors beyond token prediction time, including the response length generated and the language-specific tokenization efficiency (Conde et al., 2024).

Energy consumption was recorded in kWh and converted to CO2 equivalents using the European Commission's conversion ratio for Spain (0.158 kg CO2/kWh), as the evaluation for La Leaderboard was conducted in Barcelona (Lottick et al., 2019).

4 Evaluation Results

Table 3 shows average results for each model. Further visualizations can be found in Figures 3-10 in Appendix D. Raw results are publicly available.[14]

Models evaluated We focus on models accessible to the broader community. We evaluate 50 open-weights models from various families, primarily ranging from 1B to 9B parameters, while including larger quantized models. We assess both their base and instruction-tuned versions when available (Appendix C). Models can be categorized into two groups: *state-of-the-art family models* like Google-Gemma (Riviere et al., 2024) represent the leading edge of this field and *language-optimized models* such as EuroLLM (Martins et al., 2024) and Salamandra[15] have been designed specifically to process more efficiently the target languages and capture the cultural nuances of their speakers.

SOTA vs. Language-optimized models We observe that the first two-thirds of the top 15 models are SOTA models. This distribution suggests that technological advances in state-of-the-art models, coupled with access to greater resources by the companies involved in training them, play a more decisive role in the performance of language models than any specific solution based on pre-training, fine-tuning, or other mechanisms.

Performance per language In general, results are better for Spanish and Catalan and worse for Basque and Galician. This was expected for Basque, a language isolate very different from the rest, but not fully for Galician, as it shares Latin roots with Spanish and Catalan. However, the generalized lower scores in Galician could be a consequence of the reduced number of training and instruction datasets available for this language. Regarding specific models, Gemma2-9B is a cross-language high-performing pair of models. However, we find that some models excel for specific languages. For example, EuroLLM-9B for Catalan

[12] https://www.bsc.es/ca/marenostrum/marenostrum-5

[13] https://www.bsc.es/research-and-development/software-and-apps/software-list/ear-energy-management-framework-hpc

[14] Link removed for review.

[15] https://hf.co/collections/BSC-LT/salamandra-66fc171485944df79469043a

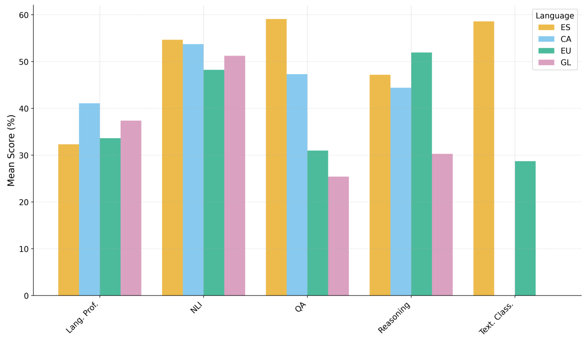

Figure 2: Results per type of task type and language.

and Galician and Salamandra-7B for Basque.

Performance per task As shown in Figure 2, the evaluation results are generally better for NLI tasks, including paraphrasing, and worse for language proficiency tests, with all four languages having similar performance on both tasks. Within the language proficiency tests, results are particularly lower for summarization tasks. In question answering and reasoning tasks, there is a larger inter-language difference, with Galician having significantly lower scores overall, while Basque has the best results for reasoning but second worst for question answering. While commonsense reasoning results are generally good, math reasoning is the category with the lowest results, which could be related to the metric used for validation being exact match. Further analysis is needed to understand whether these differences are due to the datasets used in each language or are indeed due to the models' performance. The poor results for language proficiency tests also deserve a more detailed exploration in future studies to understand their implications, as they may imply fundamental limitations of the models in their knowledge across languages.

Performance vs. size In general, our experiments show some correlation between performance and size, with models in the range of 1-2B parameters achieving better scores for their size. This is particularly true for Gemma2-2B and Qwen2.5-1.5B, both base and instructed models. Among the top 10 models, we find that all have between 8 and 9 billion parameters, except for the quantized versions from the Qwen family (Yang et al., 2024; Qwen Team, 2024).

Energy consumption The total computational resources amounted to 660.87 hours of processing time and 582.84 kWh of energy consumption, resulting in 92.09 kg of CO2 emissions. As expected, larger models consume more energy. The two largest models (Qwen2.5-32B-IT and Qwen2.5-14B-IT) are in the top three, while FLOR consumes less than models of the same size. Similarly, as anticipated, text generation tasks such as summarization require more energy for evaluation.

Energy consumption vs. performance Again, our experiments show a strong correlation between the energy consumed and the performance of a model. For one of the overall top models, Gemma2-9B, its instruction-tuned version excels with a third of the energy consumption compared to the base version.

5 Conclusions and Future Work

In this paper, we suggest a methodology to create community-driven leaderboards, including key points to gather diverse datasets, and the rationale behind a more efficient and accessible evaluation setup. In doing so, we hope to inspire the creation of more leaderboards that fulfil the needs of diverse linguistic communities.

In particular, we present LA LEADERBOARD , the first open-source leaderboard to evaluate LLMs in languages from Spain and Latin America. It is the result of a collaboration among 12 research groups. LA LEADERBOARD consists of 66 datasets in Spanish, Catalan, Basque and Galician and covers a wide range of task types and domains. Results of evaluating 50 LLMs show that performance is generally better in Spanish and Catalan. Models not optimized for the target languages (e.g., Gemma) achieve the highest scores, while fine-tuned or continually pre-trained models on these languages (e.g., EuroLLM) outperform foundation models designed with the same linguistic focus (e.g., Salamandra).

Our planned next steps include evaluating the recently donated datasets, with a special focus on indigenous languages. We will also add larger open models and proprietary models. Moreover, we are organizing a hackathon to create a benchmark to measure cultural adequacy in each Spanish-speaking country. Finally, we welcome any person or organization interested in joining our effort. In this way, we hope that LA LEADERBOARD will keep evolving to include more languages, language varieties, and use cases that motivate the development of LLMs that better serve our diverse community.

Limitations

Indigenous languages We acknowledge that indigenous languages from Latin America are not yet included among the evaluation results of LA LEADERBOARD . However, we have ongoing collaborations to include existing benchmarks and create new ones to keep extending LA LEADERBOARD to be as inclusive as possible and reflect the diversity of the Spanish-speaking community.

Spanish language varieties Currently, LA LEADERBOARD includes datasets in the Spanish varieties of Spain, Mexico, Argentina, Chile, and Uruguay. Although we don't know the exact origin of all the samples from some third-party datasets, we estimate that less than 25% of all the Spanish datasets in the leaderboard come from LATAM. We plan on increasing this percentage by collaborating with LATAM research groups in the creation of an open hackathon.

Large and proprietary models To improve the coverage of the state-of-the-art language models for the use cases included in LA LEADERBOARD , it would be interesting to evaluate larger language models as well as proprietary models.

Contamination Another pending task is to analyse potential contamination (Sainz et al., 2023) within our leaderboard. We have not addressed this yet because a high percentage of the datasets used are very recent and niche, making it unlikely that they have been incorporated into training data, unlike more established benchmarks such as MMLU (Hendrycks et al., 2021a; Wang et al., 2024b; Taghanaki et al., 2024) that serve as primary pillars in model evaluation in every model report. Nevertheless, we have started to evaluate contamination to ensure in the short-term future that we provide high-quality results.

For the datasets specifically created for LA LEADERBOARD , we advised the corresponding authors to release them gated to avoid being unintentionally included in training datasets by web scraping; AQuAS and RagQuAS are gated. The authors of TELEIA decided to release an adaptation of their dataset and keep the original private to be able to analyze contamination through time.

Ethical Considerations

Fair representation Since our objective is to establish an evaluation standard for Latin America and Spain, it is important to properly represent the linguistic and cultural diversity of the community in order to avoid the perpetuation, or even amplification, of stereotypes and inequalities.

Third-party datasets Some of the evaluation datasets included in LA LEADERBOARD were created by organizations other than our data contributors. As a result, we acknowledge the possibility that some of these datasets may have been developed using practices that could be considered unethical. These concerns range from potential legal violations to extractive data collection methods that may impact disadvantaged communities.

Environmental impact Evaluating 50 language models on 66 tasks required 660.87 hours of compute, translating to 92.09 kg of CO2. However, we hope that by publishing a comprehensive evaluation of the available models, LA LEADERBOARD will contribute to reducing the total environmental impact of individual private evaluations.

Misuse of La Leaderboard We welcome model submissions from everyone. This could potentially lead to overuse with people sending many different versions of the same model. We plan to mitigate this behaviour by following the spam mitigation strategies from the Open LLM Leaderboard (Fourrier et al., 2024).

References

Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. 2019. Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv preprint arXiv:1906.02569*.

Rodrigo Agerri, Roberto Centeno, María S. Espinosa, Joseba Fernandez de Landa, and Álvaro Rodrigo. 2021. VaxxStance@IberLEF 2021: Overview of the Task on Going Beyond Text in Cross-Lingual Stance Detection. *Procesamiento del Lenguaje Natural*, 67:173–181.

Mehdi Ali, Michael Fromm, Klaudia Thellmann, Jan Ebert, Alexander Arno Weber, Richard Rutmann, Charvi Jain, Max Lübbering, Daniel Steinigen, Johannes Leveling, Katrin Klug, Jasper Schulze Buschhoff, Lena Jurkschat, Hammam Abdelwahab, Benny Jörg Stein, Karl-Heinz Sylla, Pavel Denisov, Nicolo' Brandizzi, Qasid Saleem, Anirban

Bhowmick, Lennard Helmer, Chelsea John, Pedro Ortiz Suarez, Malte Ostendorff, Alex Jude, Lalith Manjunath, Samuel Weinbach, Carolin Penke, Oleg Filatov, Shima Asaadi, Fabio Barth, Rafet Sifa, Fabian Küch, Andreas Herten, René Jäkel, Georg Rehm, Stefan Kesselheim, Joachim Köhler, and Nicolas Flores-Herr. 2024. Teuken-7b-base & teuken-7b-instruct: Towards european llms. *Preprint*, arXiv:2410.03730.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, Colin Raffel, Leandro von Werra, and Thomas Wolf. 2025. Smollm2: When smol goes big – data-centric training of a small language model. *Preprint*, arXiv:2502.02737.

Iñigo Alonso, Maite Oronoz, and Rodrigo Agerri. 2024. MedExpQA: Multilingual Benchmarking of Large Language Models for Medical Question Answering. *Artificial Intelligence in Medicine*, 155:102938.

Jordi Armengol-Estapé, Casimiro Pio Carrino, Carlos Rodriguez-Penagos, Ona de Gibert Bonet, Carme Armentano-Oller, Aitor Gonzalez-Agirre, Maite Melero, and Marta Villegas. 2021. Are multilingual models the best choice for moderately under-resourced languages? A comprehensive assessment for Catalan. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4933–4946, Online. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Javier Aula-Blasco, Júlia Falcão, Susana Sotelo, Silvia Paniagua, Aitor Gonzalez-Agirre, and Marta Villegas. 2025. VeritasQA: A truthfulness benchmark aimed at multilingual transferability. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5463–5474, Abu Dhabi, UAE. Association for Computational Linguistics.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Fabio Barth, Manuel Brack, Maurice Kraus, Pedro Ortiz Suarez, Malte Ostendorff, Patrick Schramowski, and Georg Rehm. 2024. Occiglot euro llm leaderboard.

Irene Baucells, Javier Aula-Blasco, Iria de Dios-Flores, Silvia Paniagua Suárez, Naiara Perez, Anna Salles, Susana Sotelo Docio, Júlia Falcão, Jose Javier Saiz, Robiert Sepulveda Torres, Jeremy Barnes, Pablo Gamallo, Aitor Gonzalez-Agirre, German Rigau, and Marta Villegas. 2025. IberoBench: A benchmark for LLM evaluation in Iberian languages. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10491–10519, Abu Dhabi, UAE. Association for Computational Linguistics.

Nuria Bel, Marta Punsola, and Valle Ruíz-Fernández. 2024a. EsCoLA: Spanish corpus of linguistic acceptability. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6268–6277, Torino, Italia. ELRA and ICCL.

Núria Bel, Marta Punsola, and Valle Ruiz-Fernández. 2024b. CatCoLA: Catalan corpus of linguistic acceptability. *Procesamiento del Lenguaje Natural*, 73.

Jaione Bengoetxea, Yi-Ling Chung, Marco Guerini, and Rodrigo Agerri. 2024. Basque and Spanish Counter Narrative Generation: Data Creation and Evaluation. In *Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. *Preprint*, arXiv:2304.01373.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Lyle Campbell. 2010. Language isolates and their history, or, what's weird, anyway? In *Proceedings of the Annual Meeting of the Berkeley Linguistics Society*, volume 36, pages 16–31. Linguistic Society of America.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*.

Luis Chiruzzo, Santiago Castro, Santiago Góngora, Aiala Rosá, JA Meaney, and Rada Mihalcea. 2021. Overview of haha at iberlef 2021: Detecting, rating and analyzing humor in spanish. *Procesamiento del Lenguaje Natural*, 67:257–268.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Javier Conde, Miguel González, Pedro Reviriego, Zhen Gao, Shanshan Liu, and Fabrizio Lombardi. 2024. Speed and conversational large language models: Not all is about tokens per second. *Computer*, 57(8):74–80.

OpenCompass Contributors. 2023. Opencompass: A universal evaluation platform for foundation models. `https://github.com/open-compass/opencompass`.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Severino Da Dalt, Joan Llop, Irene Baucells, Marc Pamies, Yishi Xu, Aitor Gonzalez-Agirre, and Marta Villegas. 2024. FLOR: On the effectiveness of language adaptation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7377–7388, Torino, Italia. ELRA and ICCL.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Iria de Dios-Flores, Juan Garcia Amboage, and Marcos Garcia. 2023. Dependency resolution at the syntax-semantics interface: psycholinguistic and computational insights on control dependencies. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 203–222, Toronto, Canada. Association for Computational Linguistics.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean

Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Meza-Ruiz, et al. 2021. Americasnli: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. *arXiv preprint arXiv:2104.08726*.

Ali Elfilali, Hamza Alobeidli, Clémentine Fourrier, Basma El Amel Boussaha, Ruxandra Cojocaru, Nathan Habib, and Hakim Hacid. 2024. Open arabic llm leaderboard. `https://huggingface.co/spaces/OALL/Open-Arabic-LLM-Leaderboard`.

Julen Etxaniz, Oscar Sainz, Naiara Perez, Itziar Aldabe, German Rigau, Eneko Agirre, Aitor Ormazabal, Mikel Artetxe, and Aitor Soroa. 2024. Latxa: An open language model and evaluation suite for basque. *Preprint*, arXiv:2403.20266.

Francisco Moreno Fernández and Héctor Álvarez Mella. 2024. Demografía del español en el mundo 2024. In *El español en el mundo. Anuario del Instituto Cervantes 2024*, pages 30–97. Instituto Cervantes.

Blanca Calvo Figueras, Eneko Sagarzazu, Julen Etxaniz, Jeremy Barnes, Pablo Gamallo, Iria De Dios Flores, and Rodrigo Agerri. 2025. Truth Knows No Language: Evaluating Truthfulness Beyond English. *Preprint*, arXiv:2502.09387.

Clémentine Fourrier, Nathan Habib, Hynek Kydlíček, Thomas Wolf, and Lewis Tunstall. 2023. Lighteval: A lightweight framework for llm evaluation.

Clémentine Fourrier, Nathan Habib, Alina Lozovskaya, Konrad Szafer, and Thomas Wolf. 2024. Open llm leaderboard v2. `https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard`.

Pablo Gamallo, Pablo Rodríguez, Iria de Dios-Flores, Susana Sotelo, Silvia Paniagua, Daniel Bardanca, José Ramom Pichel, and Marcos Garcia. 2024. Open generative large language models for galician. *Preprint*, arXiv:2406.13893.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation.

Iker García-Ferrero and Begoña Altuna. 2024. Noticia: A clickbait article summarization dataset in spanish. *Preprint*, arXiv:2404.07611.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Aitor Gonzalez-Agirre, Montserrat Marimon, Carlos Rodriguez-Penagos, Javier Aula-Blasco, Irene Baucells, Carme Armentano-Oller, Jorge Palomar-Giner, Baybars Kulebi, and Marta Villegas. 2024. Building a data infrastructure for a mid-resource language: The case of Catalan. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2556–2566, Torino, Italia. ELRA and ICCL.

Aaron Grattafiori et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, Deyi Xiong, et al. 2023. Evaluating large language models: A comprehensive survey. *arXiv preprint arXiv:2310.19736*.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the math dataset. *Preprint*, arXiv:2103.03874.

Maite Heredia, Julen Etxaniz, Muitze Zulaika, Xabier Saralegi, Jeremy Barnes, and Aitor Soroa. 2024. XNLIeu: a dataset for cross-lingual NLI in Basque. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4177–4188, Mexico City, Mexico. Association for Computational Linguistics.

Carolin Holtermann, Paul Röttger, Timm Dill, and Anne Lauscher. 2024. Evaluating the elementary multilingual capabilities of large language models with multiq. *arXiv preprint arXiv:2403.03814*.

Instituto de Ingeniería del Conocimiento. 2025. Rigochat-7b-v2.

Krzysztof Jassem, Michał Ciesiółka, Filip Graliński, Piotr Jabłoński, Jakub Pokrywka, Marek Kubis, Monika Jabłońska, and Ryszard Staruch. 2025. Llmzszł: a comprehensive llm benchmark for polish. *Preprint*, arXiv:2501.02266.

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.

Aisha Khatun and Daniel G. Brown. 2024. A study on large language models' limitations in multiple-choice question answering. *Preprint*, arXiv:2401.07955.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in nlp. *Preprint*, arXiv:2104.14337.

Hyeonwoo Kim, Dahyun Kim, Jihoo Kim, Sukyung Lee, Yungi Kim, and Chanjun Park. 2024. Open ko-llm leaderboard2: Bridging foundational and practical evaluation for korean llms. *arXiv preprint arXiv:2410.12445*.

Javier De la Rosa, Eduardo G. Ponferrada, Manu Romero, Paulo Villegas, Pablo González de Prado Salas, and María Grandury. 2022. Bertin: Efficient pre-training of a spanish language model using perplexity sampling. *Procesamiento del Lenguaje Natural*, 68(0):13–23.

Latam-GPT. 2025. paes (revision 2e7fda0).

Wangyue Li, Liangzhi Li, Tong Xiang, Xiao Liu, Wei Deng, and Noa Garcia. 2024. Can multiple-choice questions really be useful in detecting the abilities of llms? *Preprint*, arXiv:2403.17752.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks are all you need ii: **phi-1.5** technical report. *arXiv preprint arXiv:2309.05463*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*. Featured Certification, Expert Certification.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kadan Lottick, Silvia Susai, Sorelle A Friedler, and Jonathan P Wilson. 2019. Energy usage reports: Environmental awareness as part of algorithmic accountability. *arXiv preprint arXiv:1911.08354*.

Wolf Lustig. 1996. Mba'éichapa oiko la guarani? guaraní y jopara en el paraguay. *URL: www. staff. uni-mainz. de/lustig/guarani/art/yopara. pdf.–20 p*.

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza. 2018. Challenges of language technologies for the indigenous languages of the americas. *arXiv preprint arXiv:1806.04291*.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios Gonzales, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, et al. 2021. Findings of the americasnlp 2021 shared task on open machine translation for indigenous languages of the americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 202–217.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2024. Eurollm: Multilingual language models for europe. *Preprint*, arXiv:2409.16235.

Fabio Mercorio, Mario Mezzanzanica, Daniele Potertì, Antonio Serino, and Andrea Seveso. 2024. Disce aut deficere: Evaluating llms proficiency on the invalsi italian benchmark. *arXiv preprint arXiv:2406.17535*.

Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2023. Gaia: a benchmark for general ai assistants. *Preprint*, arXiv:2311.12983.

Mario Mina, Valle Ruiz-Fernández, Júlia Falcão, Luis Vasquez-Reina, and Aitor Gonzalez-Agirre. 2025. Cognitive biases, task complexity, and result intepretability in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1767–1784, Abu Dhabi, UAE. Association for Computational Linguistics.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez-García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, Nedjma Ousidhoum, Jose Camacho-Collados, and Alice Oh. 2025. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Preprint*, arXiv:2406.09948.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Online. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Ankit Pal, Pasquale Minervini, Andreas Geert Motzfeldt, and Beatrice Alex. 2024. openlifescienceai/open_medical_llm_leaderboard.

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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, Mexico City, Mexico. Association for Computational Linguistics.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra, et al. 2025. Humanity's last exam. *arXiv preprint arXiv:2501.14249*.

Irene Plaza, Nina Melero, Cristina del Pozo, Javier Conde, Pedro Reviriego, Marina Mayor-Rocher, and María Grandury. 2024. Spanish and llm benchmarks: is mmlu lost in translation? *arXiv preprint arXiv:2406.17789*.

Flor Miriam Plaza-del-Arco, Arturo Montejo-Ráez, L. Alfonso Ureña-López, and María-Teresa Martín-Valdivia. 2021. OffendES: A new corpus in Spanish for offensive language research. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1096–1108, Held Online.

Flor Miriam Plaza-del-Arco, Carlo Strapparava, L. Alfonso Urena Lopez, and Maite Martin. 2020. EmoEvent: A multilingual emotion corpus based on different events. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1492–1498, Marseille, France. European Language Resources Association.

Posadas-Durán, Helena Gómez-Adorno, Sidorov, and Escobar. 2019. Detection of fake news in a new corpus for the spanish language. *Journal of Intelligent & Fuzzy Systems*, 36(5):4869–4876.

Amandine Quercia, Jamil Zaghir, Christian Lovis, and Christophe Gaudet-Blavignac. 2024. Medfrenchmark, a small set for benchmarking generative llms in medical french. In *Digital Health and Informatics Innovations for Sustainable Health Care Systems*, pages 601–605. IOS Press.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Julian Risch, Timo Möller, Julian Gutsch, and Malte Pietsch. 2021. Semantic answer similarity for evaluating question answering models. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 149–157, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv e-prints*, pages arXiv–2408.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A Haggag, Alfonso Amayuelas, et al. 2024. Include: Evaluating multilingual language understanding with regional knowledge. *arXiv preprint arXiv:2411.19799*.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, Frederico Belcavello, Ganzorig Batnasan, Gisela Vallejo, Grainne Caulfield, Guido Ivetta, Haiyue Song, Henok Biadglign Ademtew, Hernán Maina, Holy Lovenia, Israel Abebe Azime, Jan Christian Blaise Cruz, Jay Gala, Jiahui Geng, Jesus-German Ortiz-Barajas, Jinheon Baek, Jocelyn Dunstan, Laura Alonso Alemany, Kumaranage Ravindu Yasas Nagasinghe, Luciana Benotti, Luis Fernando D'Haro, Marcelo Viridiano, Marcos Estecha-Garitagoitia, Maria Camila Buitrago Cabrera, Mario Rodríguez-Cantelar, Mélanie Jouitteau, Mihail Mihaylov, Mohamed Fazli Mohamed Imam, Muhammad Farid Adilazuarda, Munkhjargal

Gochoo, Munkh-Erdene Otgonbold, Naome Etori, Olivier Niyomugisha, Paula Mónica Silva, Pranjal Chitale, Raj Dabre, Rendi Chevi, Ruochen Zhang, Ryandito Diandaru, Samuel Cahyawijaya, Santiago Góngora, Soyeong Jeong, Sukannya Purkayastha, Tatsuki Kuribayashi, Teresa Clifford, Thanmay Jayakumar, Tiago Timponi Torrent, Toqeer Ehsan, Vladimir Araujo, Yova Kementchedjhieva, Zara Burzo, Zheng Wei Lim, Zheng Xin Yong, Oana Ignat, Joan Nwatu, Rada Mihalcea, Thamar Solorio, and Alham Fikri Aji. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *Preprint*, arXiv:2406.05967.

Aiala Rosá, Luis Chiruzzo, Lucía Bouza, Alina Dragonetti, Santiago Castro, Mathias Etcheverry, Santiago Góngora, Santiago Goycoechea, Juan Machado, Guillermo Moncecchi, et al. 2022. Overview of quales at iberlef 2022: Question answering learning from examples in spanish.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. Nlp evaluation in trouble: On the need to measure llm data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Winogrande: An adversarial winograd schema challenge at scale. *Preprint*, arXiv:1907.10641.

Eva Sánchez Salido, Roser Morante, Julio Gonzalo, Guillermo Marco, Jorge Carrillo de Albornoz, Laura Plaza, Enrique Amigó, Andrés Fernández, Alejandro Benito-Santos, Adrián Ghajari Espinosa, and Victor Fresno. 2025. Bilingual evaluation of language models on general knowledge in university entrance exams with minimal contamination. *Preprint*, arXiv:2409.12746.

Elisa Sanchez-Bayona and Rodrigo Agerri. 2024. Meta4XNLI: A Crosslingual Parallel Corpus for Metaphor Detection and Interpretation. *Preprint*, arXiv:2404.07053.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I. Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Madeline Smith, Antoine Bosselut, Alice Oh, Andre F. T. Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. 2024. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *Preprint*, arXiv:2412.03304.

Aarohi Srivastava and et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Stephanie Strassel and Jennifer Tracey. 2016. LORELEI language packs: Data, tools, and resources for technology development in low resource languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3273–3280, Portorož, Slovenia. European Language Resources Association (ELRA).

Saeid Asgari Taghanaki, Aliasgahr Khani, and Amir Khasahmadi. 2024. Mmlu-pro+: Evaluating higher-order reasoning and shortcut learning in llms. *arXiv preprint arXiv:2409.02257*.

Gorka Urbizu, Iñaki San Vicente, Xabier Saralegi, Rodrigo Agerri, and Aitor Soroa. 2022. BasqueGLUE: A Natural Language Understanding Benchmark for Basque. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1603–1612.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint 1905.00537*.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024a. "my answer is C": First-token probabilities do not match text answers in instruction-tuned language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7407–7416, Bangkok, Thailand. Association for Computational Linguistics.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. 2024b. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*.

Cheng Xu, Shuhao Guan, Derek Greene, M Kechadi, et al. 2024. Benchmark data contamination of large language models: A survey. *arXiv preprint arXiv:2406.04244*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang

Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. *Preprint*, arXiv:2102.09690.

Elena Zotova, Rodrigo Agerri, and German Rigau. 2021. Semi-automatic Generation of Multilingual Datasets for Stance Detection in Twitter. *Expert Systems with Applications*, 170:114547.

Irune Zubiaga, Aitor Soroa, and Rodrigo Agerri. 2024. A LLM-based Ranking Method for the Evaluation of Automatic Counter-Narrative Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9572–9585.

A Evaluation Datasets

The datasets are used only for evaluation, aligning with their intended uses.

Spanish datasets

The Spanish datasets are: AQuAS (Cite removed for review), Belebele (Bandarkar et al., 2024), EsCoLA (Bel et al., 2024a), Fake News ES (Posadas-Durán et al., 2019), FLORES-200 (Costa-jussà et al., 2022), ClinTreatES and ClinDiagnosES (Cite removed for review), HumorQA (Cite removed for review), MGSM (Shi et al., 2023), MultiLingualCrowsPairs (Nangia et al., 2020), NoticIA (García-Ferrero and Altuna, 2024), OffendES (Plaza-del-Arco et al., 2021), RagQuAS (Cite removed for review), SpaLawEx (Cite removed for review), TELEIA (Cite removed for review), WNLI (Gonzalez-Agirre et al., 2024; Baucells et al., 2025),[16] XL-Sum (Hasan et al., 2021), XStoryCloze (Lin et al., 2022; Baucells et al., 2025), and XQuAD (Artetxe et al., 2020).

Catalan datasets

The Catalan datasets are: caBREU, CatalanQA, COPA-ca, CoQCat, PAWS-ca, TE-ca, WNLI-ca and XNLI-ca (Gonzalez-Agirre et al., 2024), IberoBench (Baucells et al., 2025), CatCoLA (Bel et al., 2024b), FLORES-200 (Costa-jussà et al., 2022), MGSM (Shi et al., 2023), XStoryCloze (Lin et al., 2022), XQuAD-ca (Armengol-Estapé et al., 2021), XStoryCloze (Lin et al., 2022; Baucells et al., 2025), Parafraseja,[17] PAWS-X (Yang et al., 2019), and VeritasQA (Aula-Blasco et al., 2025).

Basque datasets

The Basque datasets are: EusExams, EusReading, EusProficiency and EusTrivia from Etxaniz et al. (2024); BEC2016eu, BHTCv2, EpecKorrefBin, QNLIeu, WiCeu from BasqueGlue (Urbizu et al., 2022); QNLI-eu (Urbizu et al., 2022), VaxxStance (Agerri et al., 2021), XNLIeu (Heredia et al., 2024), FLORES-200 (Costa-jussà et al., 2022), MGSM (Shi et al., 2023), and XStoryCloze (Lin et al., 2022; Baucells et al., 2025).

Galician datasets

The Galician datasets are: FLORES-200 (Costa-jussà et al., 2022), GalCoLA (de Dios-Flores et al., 2023), TruthfulQA-GL,[18] and XStoryCloze (Lin et al., 2022; Baucells et al., 2025).[19]

Datasets created for La Leaderboard

The 7 datasets specifically created for La Leaderboard are AQuAS, ClinDiagES, ClinTreatES, HumorQA, RagQuAS, SpaLawEx, and TELEIA. Their corresponding datasheets are included in Appendix F.

Newly donated datasets

The new datasets donated will be evaluated shortly. These include CONAN-EUS (Bengoetxea et al., 2024), RefutES,[20] TruthfulQA in Basque, Catalan, Galician and Spanish (Figueras et al., 2025), VeritasQA (Aula-Blasco et al., 2025), PAES Chile (Latam-GPT, 2025), meta4xnli (Sanchez-Bayona and Agerri, 2024), MedExpQA (Alonso et al., 2024), Catalonia Independence Corpus (CIC) in Catalan and Spanish (Zotova et al., 2021), HAHA humor detection and analysis in Spanish (Chiruzzo et al., 2021), QuALES for question-answering in Spanish in the COVID-19 domain (Rosá et al., 2022), AmericasNLP-MT (Mager et al., 2021), AmericasNLI (Ebrahimi et al., 2021), TraduLATAM, and VocesOriginarias evaluating indigenous languages.

Evaluation dataset details

The Tables 4 (Spanish), 5 (Catalan), 6 (Basque), and 7 (Galician) list these datasets, providing additional information about their task type, domain, and origin.

We run the evaluations using our fork of the LM Evaluation Harness,[21] synced with the main repository on commit 6ccd520f3fb2b5d74c6f14c05f9d189521424719. The tables mentioned also include details about the evaluation configuration providing the Harness task ID, metric, and number of shots.

B Frontend Detailed Description

The implementation of La Leaderboard is based on the HuggingFace leaderboard template. [22] The frontend is developed using Gradio (Abid

[16] For Spanish, see https://hf.co/datasets/PlanTL-GOB-ES/wnli-es.

[17] https://hf.co/datasets/projecte-aina/Parafraseja

[18] https://hf.co/datasets/proxectonos/truthfulqa_gl

[19] For Galician, see https://hf.co/datasets/proxectonos/xstorycloze_gl.

[20] https://huggingface.co/datasets/SINAI/RefutES

[21] Link removed for review.

[22] https://hf.co/spaces/demo-leaderboard-backend/leaderboard

Dataset	Task	Metric	Domain	Origin	#Examples	#Shots
AQuAS	Abstractive QA, Long Form QA	sas_encoder	Miscellaneous	Original	87	1
Belebele Spa	Reading Comprehension	acc	Miscellaneous	Human translation	900	2
ClinDiagnosES	Long Form QA	sas_encoder	Clinical	Original	62	0
ClinTreatES	Long Form QA	sas_encoder	Clinical	Original	62	0
COPA_es	Commonsense Reasoning	acc	Lang. prof., Misc.	Human translation	500	3
Crows Pairs Spanish	Stereotype Detection	pct_stereotype	Ethics, Hate speech	Original	1509	0
EsCoLA	Linguistic Acceptability	mcc	Language proficiency	Original	1060	2
Fake News ES	Fake News Detection	acc	Press	Original	572	2
HumorQA	Humor Classification	acc	Language proficiency	Original	51	0
MGSM_es	Math Reasoning	exact_match	Math	Human translation	250	2
NoticIA	Summarization	rouge1	Language proficiency, Press	Original	100	0
OffendES	Hate Speech Detection	acc	Hate speech	Original	13600	2
OpenBookQA_es	Multiple Choice QA	acc	General knowledge	Human translation	500	0
PAWS-X_es	Paraphrasing	acc	Lang. prof., Misc.	Original	2000	3
RagQuAS	Abstractive QA, Long Form QA	sas_encoder	Miscellaneous	Original	201	1
SpaLawEx	Multiple Choice QA	acc	Legal	Original	119	0
TELEIA	Multiple Choice QA	acc	General knowledge, Lang. prof.	Original	100	2
WNLI ES	Natural Language Inference	acc	Lang. prof., Misc.	Human translation	146	2
XL-Sum_es	Summarization	bleu	Press	Original	4763	1
XNLI_es	Natural Language Inference	acc	Miscellaneous	Original	5010	3
XQuAD_es	Extractive QA	f1	Miscellaneous	Original	1190	2
xStoryCloze_es	Commonsense Reasoning	acc	Miscellaneous	Human translation	1510	0

Table 4: Details of the evaluation datasets in Spanish (ES).

Dataset	Task	Metric	Domain	Origin	#Examples	#Shots
ARC_ca	Multiple Choice QA	acc	Science	Human translation	869	2
Belebele Cat	Reading Comprehension	acc	Miscellaneous	Human translation	900	2
caBREU	Summarization	bleu	Press	Original	301	1
CatalanQA	Extractive QA	f1	Miscellaneous	Original	2135	2
CatCoLA	Linguistic Acceptability	mcc	Language proficiency	Original	1020	2
COPA_ca	Commonsense Reasoning	acc	Lang. prof., Misc.	Human translation	500	3
CoQCat	Extractive QA	f1	Miscellaneous	Original	8986	1
MGSM_ca	Math Reasoning	exact_match	Math	Human translation	250	2
OpenBookQA_ca	Multiple Choice QA	acc	General knowledge	Human translation	500	0
Parafraseja	Paraphrasing	acc	Language proficiency	Original	21984	3
PAWS_ca	Paraphrasing	acc	Lang. prof., Misc.	Human translation	2000	3
PIQA_ca	Multiple Choice QA	acc	General knowledge	Human translation	1838	2
SIQA_ca	Multiple Choice QA	acc	General knowledge	Human translation	1954	2
TE-ca	Natural Language Inference	acc	Lang. prof., Misc.	Original	2117	3
WNLI_ca	Natural Language Inference	acc	Lang. prof., Misc.	Human translation	146	2
XNLI_ca	Natural Language Inference	acc	Lang. prof., Misc.	Human translation	5010	3
XQuAD_ca	Extractive QA	f1	Miscellaneous	Human translation	1190	2
xStoryCloze_ca	Commonsense Reasoning	acc	Miscellaneous	Human translation	1510	0

Table 5: Details of the evaluation datasets in Catalan (CA).

Dataset	Task	Metric	Domain	Origin	#Examples	#Shots
BEC2016eu	Sentiment Analysis	f1	Politics, Twitter	Original	1302	3
Belebele Eus	Reading Comprehension	acc	Miscellaneous	Human translation	900	2
BertaQA	Multiple Choice QA	acc	Cultural Knowledge	Original	4760	3
BHTCv2	Topic Classification	f1	Press	Original	1854	2
EpecKorrefBin	Natural Language Inference	acc	Press	Original	587	3
EusExams	Multiple Choice QA	acc	Miscellaneous	Original	16000	4
EusProficiency	Multiple Choice QA	acc	Language proficiency	Original	5169	4
EusReading	Reading Comprehension	acc	Miscellaneous	Original	352	1
EusTrivia	Multiple Choice QA	acc	General knowledge	Original	1715	4
MGSM_eu	Math Reasoning	exact_match	Math	Human translation	250	2
QNLIeu	Natural Language Inference	acc	Miscellaneous	Original	238	2
VaxxStance	Stance Detection	f1	Politics, Twitter	Original	312	3
WiCeu	Natural Language Inference	acc	Language proficiency	Original	1400	2
WNLI_eu	Natural Language Inference	acc	Lang. prof., Misc.	Human translation	146	2
XCOPA_eu	Commonsense Reasoning	acc	Lang. prof., Misc.	Human translation	500	3
XNLI_eu	Natural Language Inference	acc	Lang. prof., Misc.	Reviewed MT	5010	3
xStoryCloze_eu	Commonsense Reasoning	acc	Miscellaneous	Human translation	1510	0

Table 6: Details for evaluation datasets in Basque (EU).

Dataset	Task	Metric	Domain	Origin	#Examples	#Shots
Belebele Glg	Reading Comprehension	acc	Miscellaneous	Reviewed MT	900	2
GalCoLA	Linguistic Acceptability	mcc	Language proficiency	Original	1710	2
MGSM_gl	Math Reasoning	exact_match	Math	Reviewed MT	250	2
OpenBookQA_gl	Multiple Choice QA	acc	General knowledge	Reviewed MT	500	0
ParafrasesGL	Paraphrasing	acc	Language proficiency	Original	294	3
PAWS_gl	Paraphrasing	acc	Lang. prof., Misc.	Reviewed MT	2000	3
SummarizationGL	Summarization	bleu	Press	Original	8080	1
XNLI_gl	Natural Language Inference	acc	Lang. prof., Misc.	Reviewed MT	5010	3
xStoryCloze_gl	Commonsense Reasoning	acc	Miscellaneous	Human translation	1510	0

Table 7: Details for evaluation datasets in Galician (GL).

et al., 2019) and divided into four tabs:

- The landing tab, named "La Leaderboard", is divided into five sub-tabs, each containing tables with all the evaluated models and their corresponding average results. These sub-tabs include overall and language-specific results for Spanish, Catalan, Basque, and Galician. The results are aggregated by averaging the scores across all tasks for each language.

- For transparency and reproducibility purposes, the second tab, "Info", includes the command we use to evaluate the models and also the normalization formula. In the acknowledgements section, we list the institutions and every person who contributed to the project.

- The next tab describes all the "Tasks" included in LA LEADERBOARD .

- Finally, there is a tab where everyone can submit their model for evaluation.

The text of the information and submission tabs is available both in English and Spanish to bring the tool closer to the community.

In the footer, we can find the citation information for the software, all the included datasets, and the evaluation suite. Below are the fourteen logos from all the collaborating institutions. The entities in the acknowledgements are ordered chronologically by the date they joined the project to thank early adopters, whereas the logos in the footer are ordered by the number of datasets donated.

C Models Evaluated

Table 8 details the 50 models evaluated, including the following families: Aitana[23], BERTIN (la Rosa et al., 2022), Carballo (Gamallo et al., 2024), FLOR (Da Dalt et al., 2024), LeniaChat,[24] RigoChat (Instituto de Ingeniería del Conocimiento, 2025), Salamandra,[25] Occiglot[26], EuroLLM (Martins et al., 2024), Aya (Dang et al., 2024), DeepSeek (DeepSeek-AI et al., 2025), Gemma (Riviere et al., 2024), Llama (Grattafiori et al., 2024), Mistral (Jiang et al., 2023), Phi (Li et al., 2023), SmolLM (Allal et al., 2025), and Qwen (Qwen Team, 2024).

[23] https://huggingface.co/gplsi/Aitana-6.3B

[24] https://huggingface.co/LenguajeNaturalAI/leniachat-gemma-2b-v0

[25] https://hf.co/collections/BSC-LT/salamandra-66fc171485944df79469043a

[26] https://huggingface.co/collections/occiglot/occiglot-eu5-7b-v01-65dbed502a6348b052695e01

Family	Model ID	Model Type	Size (B)
Aitana	gplsi/Aitana-6.3B	pretrained	6.25
BERTIN	bertin-project/bertin-gpt-j-6B	pretrained	6.06
BERTIN	bertin-project/Gromenauer-7B	pretrained	7.24
BERTIN	bertin-project/Gromenauer-7B-Instruct	instruction-tuned	7.24
Carballo	proxectonos/Carballo-bloom-1.3B	pretrained	1.31
Flor	projecte-aina/FLOR-6.3B	pretrained	6.25
Flor	projecte-aina/FLOR-6.3B-Instructed	instruction-tuned	6.25
Latxa	HiTZ/latxa-7b-v1.2	pretrained	7.00
LeniaChat	LenguajeNaturalAI/leniachat-gemma-2b-v0	instruction-tuned	2.51
LeniaChat	LenguajeNaturalAI/leniachat-qwen2-1.5B-v0	instruction-tuned	1.54
RigoChat	IIC/RigoChat-7b-v2	instruction-tuned	7.62
Salamandra	BSC-LT/salamandra-2b	pretrained	2.25
Salamandra	BSC-LT/salamandra-2b-instruct	instruction-tuned	2.25
Salamandra	BSC-LT/salamandra-7b	pretrained	7.77
Salamandra	BSC-LT/salamandra-7b-instruct	instruction-tuned	7.77
EuroLLM	utter-project/EuroLLM-1.7B	pretrained	1.70
EuroLLM	utter-project/EuroLLM-1.7B-Instruct	instruction-tuned	1.70
EuroLLM	utter-project/EuroLLM-9B	pretrained	9.15
EuroLLM	utter-project/EuroLLM-9B-Instruct	instruction-tuned	9.15
Occiglot	occiglot/occiglot-7b-es-en	pretrained	7.24
Occiglot	occiglot/occiglot-7b-es-en-instruct	instruction-tuned	7.24
Occiglot	occiglot/occiglot-7b-eu5	pretrained	7.24
Occiglot	occiglot/occiglot-7b-eu5-instruct	instruction-tuned	7.24
Aya	CohereForAI/aya-expanse-8b	pretrained	8.03
DeepSeek	deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B	instruction-tuned	1.78
DeepSeek	deepseek-ai/DeepSeek-R1-Distill-Qwen-7B	instruction-tuned	7.62
DeepSeek	unsloth/DeepSeek-R1-Distill-Qwen-14B-bnb-4bit	instruction-tuned	8.37
Gemma	google/gemma-2-2b	pretrained	2.61
Gemma	google/gemma-2-2b-it	instruction-tuned	2.61
Gemma	google/gemma-2-9b	pretrained	9.24
Gemma	google/gemma-2-9b-it	instruction-tuned	9.24
Llama	meta-llama/Llama-3.2-1B	pretrained	1.24
Llama	meta-llama/Llama-3.2-1B-Instruct	instruction-tuned	1.24
Llama	meta-llama/Meta-Llama-3.1-8B	pretrained	8.03
Llama	meta-llama/Meta-Llama-3.1-8B-Instruct	instruction-tuned	8.03
Mistral	mistralai/Mistral-7B-Instruct-v0.3	instruction-tuned	7.25
Mistral	mistralai/Mistral-7B-v0.3	pretrained	7.25
Phi	microsoft/phi-1_5	pretrained	1.42
SmolLM	HuggingFaceTB/SmolLM2-1.7B	pretrained	1.71
SmolLM	HuggingFaceTB/SmolLM2-1.7B-Instruct	instruction-tuned	1.71
Qwen	Qwen/Qwen2.5-1.5B	pretrained	1.54
Qwen	Qwen/Qwen2.5-1.5B-Instruct	instruction-tuned	1.54
Qwen	Qwen/Qwen2.5-7B	pretrained	7.62
Qwen	Qwen/Qwen2.5-7B-Instruct	instruction-tuned	7.62
Qwen	Qwen/Qwen2.5-14B-Instruct-GPTQ-Int8	instruction-tuned	4.99
Qwen	Qwen/Qwen2.5-32B-Instruct-GPTQ-Int4	instruction-tuned	5.74

Table 8: Models evaluated in LA LEADERBOARD as of February 2025. The size is specified in billions of parameters, as appears in the corresponding Hugging Face model page. The table is divided into sections starting at the top with the models trained for the languages of Spain, then the ones from European projects, and finally the international ones.

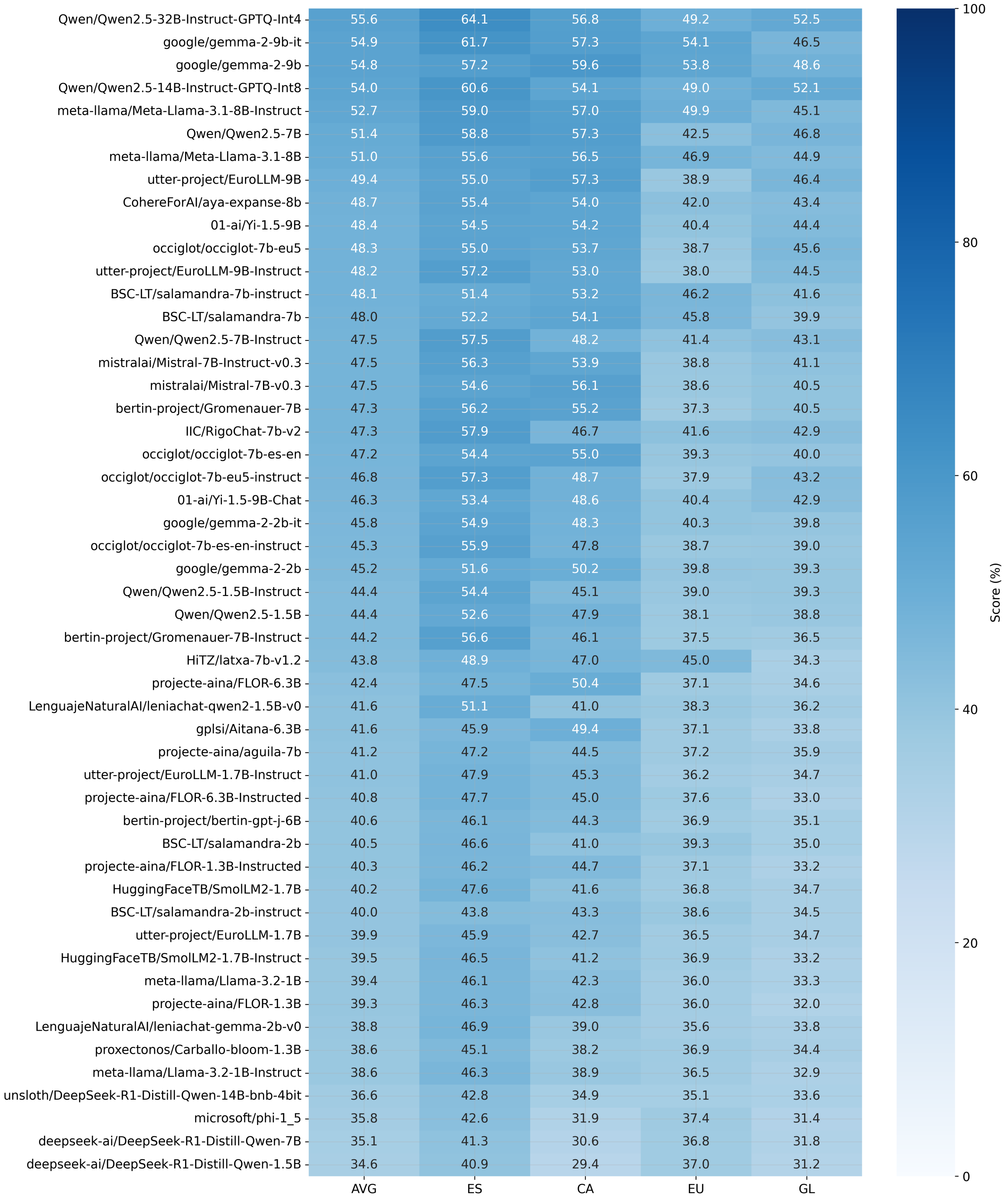

Figure 3: Results of the first set of models evaluated on LA LEADERBOARD .

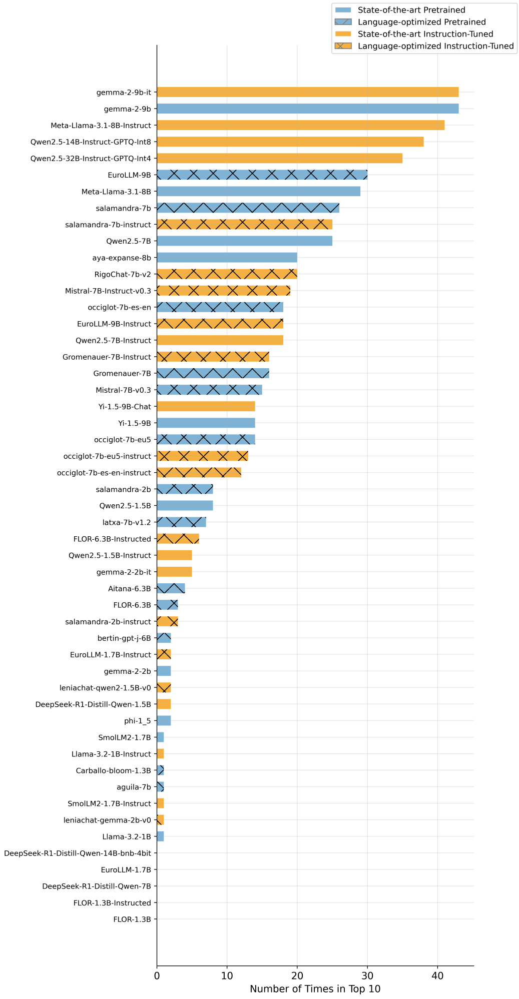

Figure 4: Number of tasks in which a model is among the top 10 models.

D Evaluation Results

This section briefly presents and discusses the evaluation results, comparing models, languages, and tasks considering metrics such as performance and energy efficiency. Each pair model-task was only evaluated once. All individual task results are publicly available in our Hugging Face dataset.[27]

Overall performance

The average evaluation results are summarized in Figure 3, overall and for each language separately. In general, the models that achieve better results are Qwen2.5-32B-IT, Gemma-2 9B-IT, Gemma-2 9B, Qwen2.5-14B-IT and Llama-3.1-8B-IT. The best results in Spanish and Galician are from Qwen2.5-32B-IT, while the best for Catalan is Gemma-2 9B and for Basque its instructed version. Interestingly, the sixth model in the classification is Qwen2.5-7B, which has a size closer to that of the Gemma and LLama models. Therefore, the performance of Qwen2.5-32B-IT and Qwen2.5-14B-IT is probably due to these models having more parameters than the rest.

A very bad or good score in a few tasks can lower or raise the average score for a model and distort the comparison. Therefore, we show the results in terms of the number of tasks for which a model is in the top 10 in Figure 4. This provides an alternative view of the results, focusing on the number of tasks for which the performance of the model is good. It can be seen that the top 5 models are the same as before, but the order changes. Now the Gemma models are in the first two positions, Llama in the third, and the two Qwen models in the last two positions. The results per language are presented in Figure 5. It can be seen that Gemma models are the best in all four languages, but the top 5 changes significantly, and some language-optimized models are in top positions. For example, EuroLLM-9B is the second in Catalan and Galician, and Salamandra-7B is the fourth in Basque.

Performance per language

In general, results are better for Spanish and Catalan and worse for Basque and Galician. This was expected for Basque, a language isolate very different from the rest, but not fully for Galician, as it shares Latin roots with Spanish and Catalan. However, the generalized lower scores in Galician could be a consequence of the reduced number of training and instruction datasets available for this language.

SOTA vs. Language-optimized models

The comparison between these two groups of models is of particular importance to study different training strategies. By analysing the top 15 models, we observe that two-thirds belong to SOTA models, while only one-third correspond to optimized models, which are found at the end of this list. It is relevant to note that all the models have roughly the same number of parameters except for the Qwen family as discussed before (Yang et al., 2024; Qwen Team, 2024).

These results suggest that, as of today, models developed by large companies still have the best overall performance across languages despite the efforts to implement language-specific models. Whether this is due to actual language proficiency or a mirage caused by good task format understand-

[27] Link removed for review.

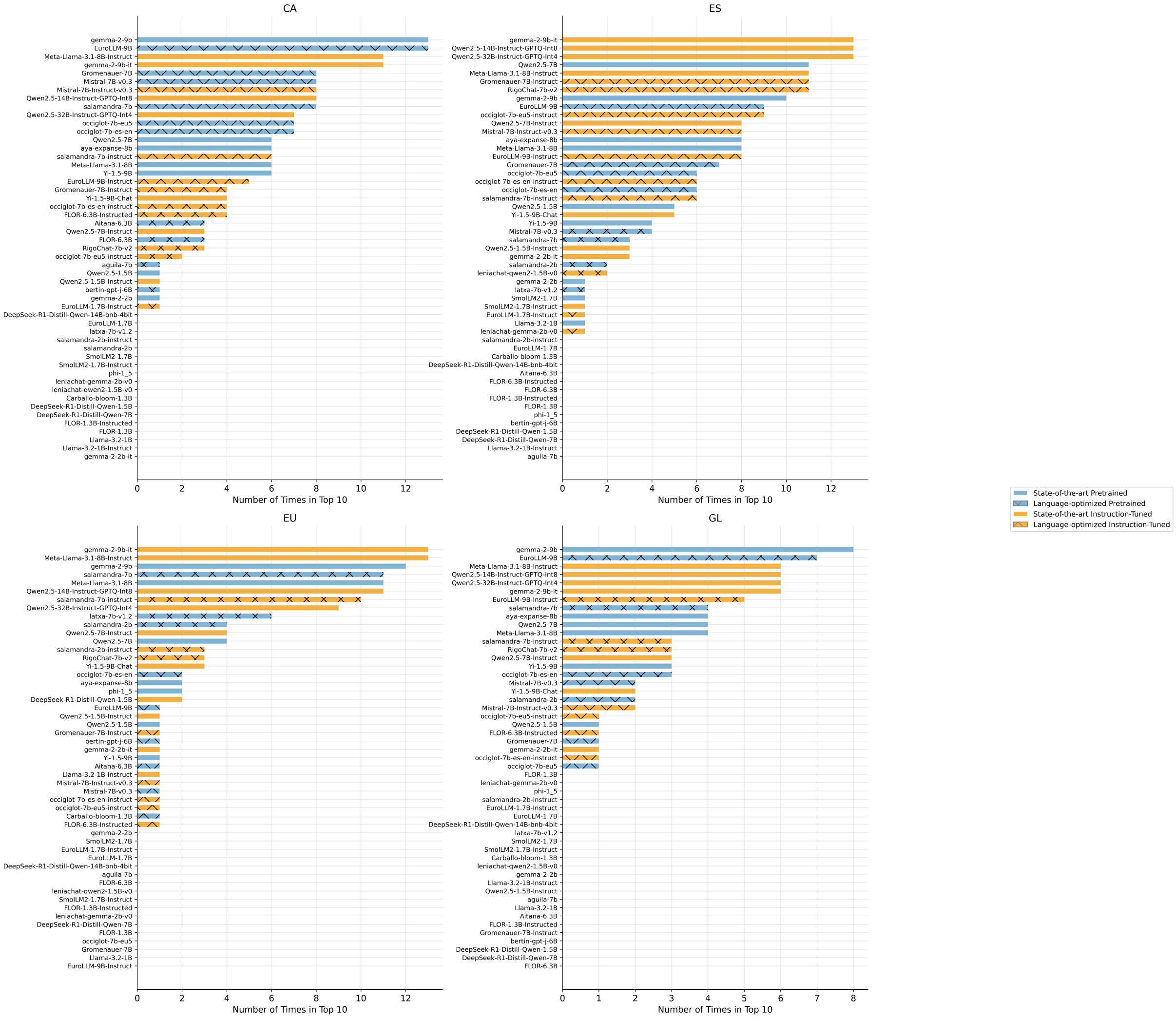

Figure 5: Number of tasks in which a model is among the top 10 models, by language.

ing and inter-language generalization is something we consider researching in future work.

Performance per task

Figure 6 shows the results per type of task for each language. As can be seen, results are generally better for natural language inference tasks and worse for language proficiency tests, with all four languages having similar performance on both tasks. In question answering and reasoning tasks, there is a larger inter-language difference, with Galician having significantly lower scores overall, while Basque has the best results for reasoning but second worst for question answering. Further analysis is needed to understand whether these differences are due to the datasets used in each language or are indeed due to the models' performance. The poor results for language proficiency tests also deserve a more detailed exploration in future studies to understand their implications, as they may imply fundamental limitations of the models in their knowledge across languages.

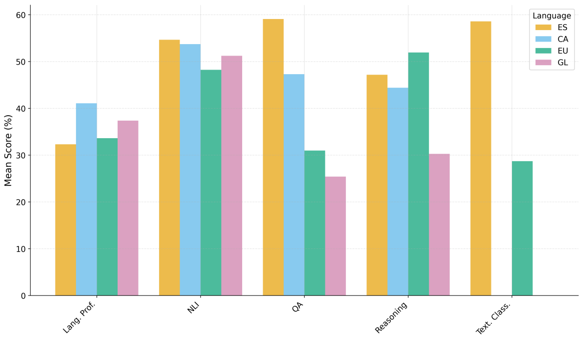

Figure 6: Results per type of task type, where "Language Proficiency" includes reading comprehension, linguistic acceptability and summarization, "NLI" includes textual entailment and paraphrasing, and "Reasoning" includes commonsense and mathematical reasoning.

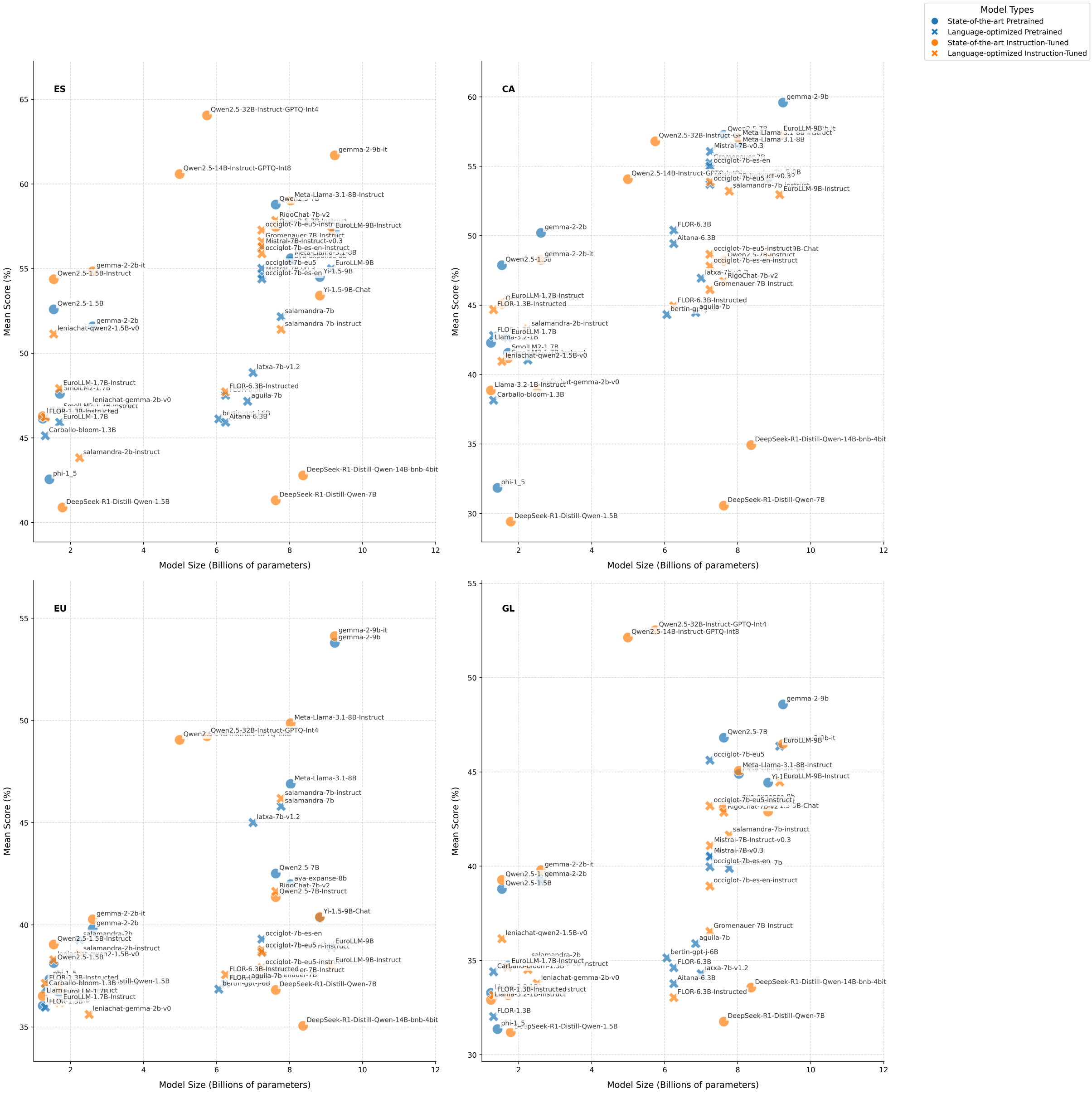

Figure 7: Results of the first set of models evaluated on LA LEADERBOARD organized by language, model family, size, and model type.

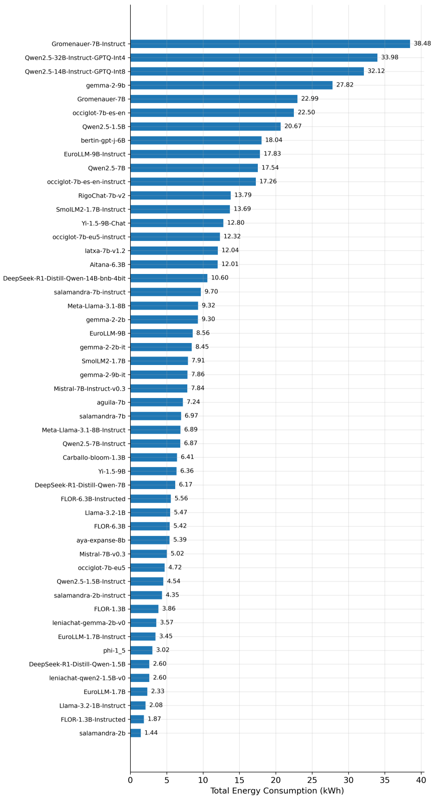

Figure 8: Distribution of results of models evaluated on LA LEADERBOARD organized by energy consumption.

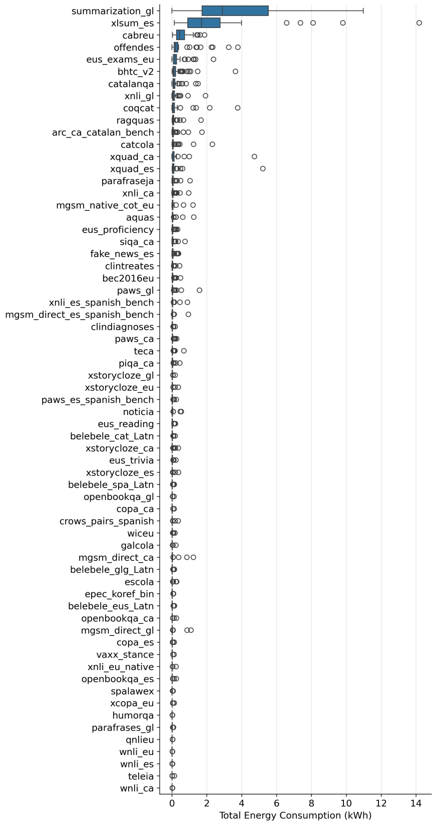

Figure 9: Energy consumption for the tasks evaluated on LA LEADERBOARD .

Performance vs. size

Figure 7 shows the average performance across all tasks for each language versus the model size. It can be seen that there is some correlation between size and performance but with large variations among models.

Model efficiency

Figure 8 represents the total energy consumed by each model. On average, each model consumed 9.25 kWh (median = 6.88, SD = 8.42), showing a wide variety in energy usage. The models that consumed the most energy were Grommeanuer-7B-Instruct, Qwen2.5-32B-Instruct-GPTQ-Int4, and Qwen2.5-14B-Instruct-GPTQ-Int8, each exceeding 30 kWh. On the other hand, Salamandra-2b, FLOR-1.3B-Instructed, and LLama-3.2-1B-Instruct were the most energy-efficient, consuming less than 2.1 kWh each. In this case, a strong correlation between model size and energy dissipation is observed as the number of arithmetic operations required to predict a token is related to the number of parameters of the model.

Regarding the tasks that required the most energy, those focused on text summarization (xlsum_es, summarization_gl, and cabreu) stood out (Figure 9). LLMs generate text token by token and their prediction speed remains constant (assuming the same hardware and stable conditions). Therefore, the most expected energy-intensive tasks are those that require the generation of larger amounts of text.

Figure 10 presents a comparison between model size and energy consumption. As expected, the general trend indicates that larger models consume more energy, with consumption increasing approximately threefold between the smallest models (1–2 billion parameters) and the largest ones (6–9 billion). However, some outliers are observed, such as Qwen, which consumes significantly more energy across all its sizes compared to other models. Conversely, models like FLOR exhibit considerably lower energy consumption across their different sizes relative to other models of similar scale.

Finally, Figure 11 shows the relation between performance and energy consumption. It can be seen that again there is a strong correlation but with large variations across models. For example, Gemma-2-9B-IT achieves one of the best scores with a low energy consumption.

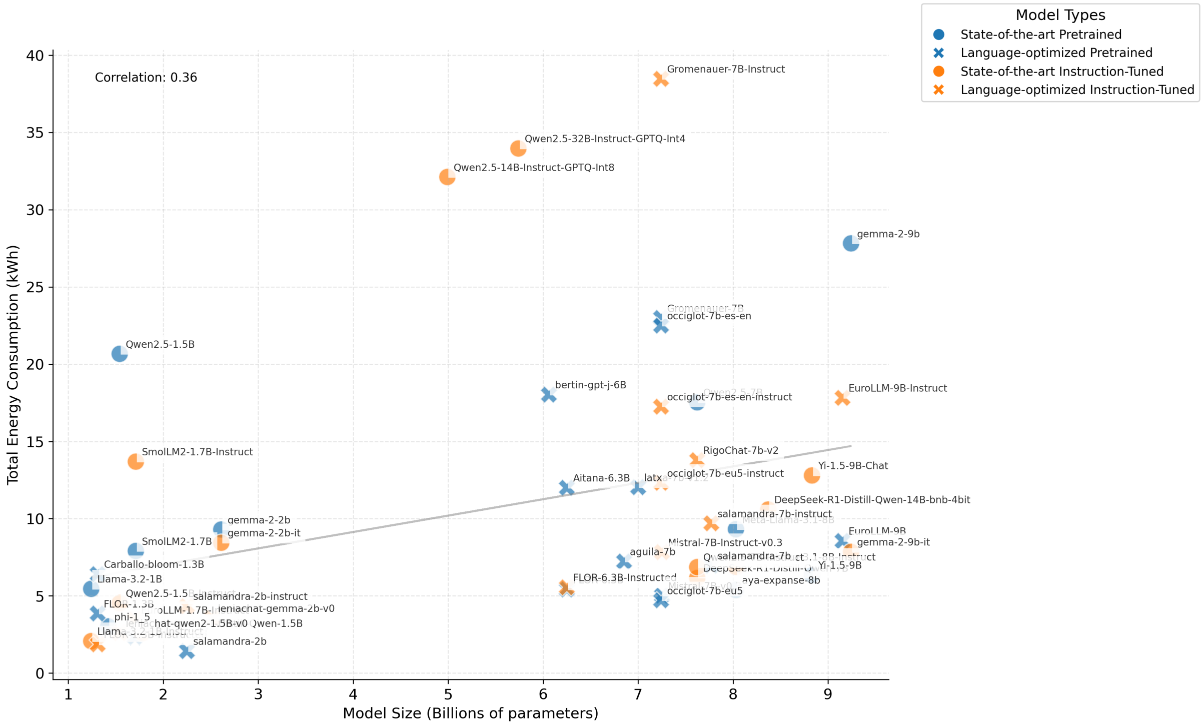

Figure 10: Distribution of results of models evaluated on LA LEADERBOARD energy consumption versus size.

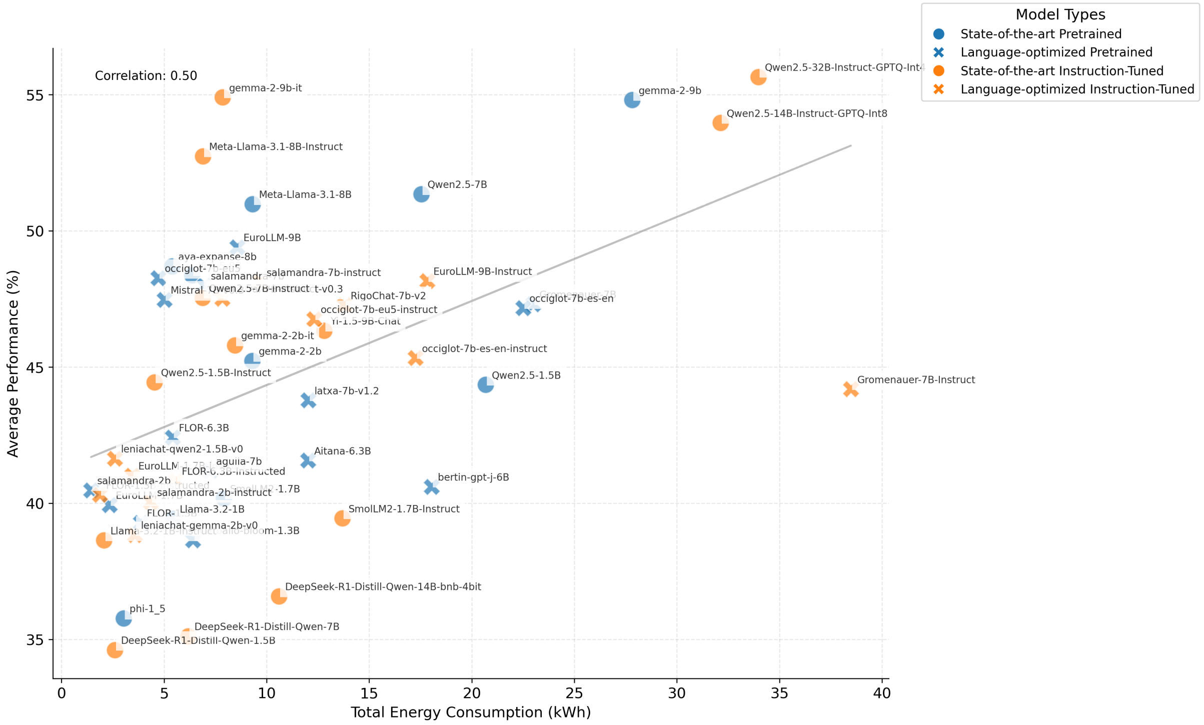

Figure 11: Distribution of results of models evaluated on LA LEADERBOARD energy consumption versus performance.

E Data Collection Campaign

Below are the questions, translated into English, corresponding to the Google Form used in the open data collection campaign.

1. Email *
2. Data source (Select one option) *
   (a) The dataset is public
   (b) Instructions to recreate it are available
   (c) The dataset is private but access can be requested on a website
   (d) The dataset is currently private, but we want to open it as a donation
   (e) The dataset is private, but you should try contacting the organization that created it
3. Dataset link * This can be the dataset link, the instructions to recreate it, or the corresponding organization's website if private.
4. If your dataset is not uploaded to Hugging Face, would you like us to take care of uploading it? (Select one option)
   (a) Yes, upload it to the SomosNLP organization
   (b) Yes, help me create my own organization and upload it
   (c) No, I prefer to create my own organization and upload it myself
5. Modality * (Select one option)
   (a) Text
   (b) Audio
   (c) Image (e.g., images with descriptions)
6. Language(s) * (Select all that apply)
   (a) Spanish
   (b) Other: _____
7. Country(ies) * Country(ies) of origin of the data and/or the people who annotated it. A region can also be specified if known. The more information, the better.
8. Tasks * (Select all that apply)
   (a) Language modeling (unannotated)
   (b) Question answering (QA)
   (c) Classification
   (d) Token classification (e.g., NER, PoS)
   (e) Translation
   (f) Summarization
   (g) Semantic similarity
   (h) Multimodal (e.g., text-to-image, audio-to-text)
9. Subtask For example, subtasks of "text classification" could be "sentiment analysis" or "hate speech detection."
10. Domain * (Select all that apply)
    (a) Legal
    (b) Clinical or biomedical
    (c) Academic or technical
    (d) Literature or music
    (e) Social media or forums
    (f) News or articles
    (g) Dialogues
    (h) General
11. Number of examples Enter the exact number of examples if known, otherwise provide a range.
12. License type *
    (a) Commercial
    (b) Non-commercial
13. License link
14. Link to the dataset documentation or any other relevant information: description, annotation and cleaning process, ethical considerations... *
15. Link to the script/repository on GitHub to download or process the dataset
16. Thank you very much for your contribution! To publicly acknowledge your contribution, you may share your name and/or affiliation to be displayed on the website. If this is a donation, we will contact you soon—thank you!
17. Name
18. Affiliation
19. How could we improve this campaign? Who would you recommend we contact? Anything else you'd like to tell us?

F Datasheets

We present the datasheets (Gebru et al., 2021) corresponding to each of the datasets specifically created for La Leaderboard : AQuAS, ClinDiagnosES, ClinTreatES, HumorQA, RagQuAS, SpaLawEx, TELEIA. Moreover, we propose an adaptation for leaderboards and fill it for La Leaderboard .

La Leaderboard

Motivation for Leaderboard Creation

Why was the leaderboard created? (e.g., were there specific tasks in mind, or a specific gap that needed to be filled?)

La Leaderboard is the first open-source leaderboard to evaluate generative LLMs in languages of Spain and Latin America. By aiming to address the linguistic and cultural diversity of the Spanish-speaking community, La Leaderboard aims to set a new standard for multilingual LLM evaluation. Our goal is to encourage the development of models that are not only linguistically competent but also culturally aware, ultimately driving progress in Natural Language Processing (NLP) for the benefit of our whole community.

Who funded the creation of the leaderboard? If there is an associated grant, provide the grant number.

La Leaderboard is an initiative launched by an international open-source community and was promoted by volunteers. The funding of each of the individual datasets donated to La Leaderboard will be disclosed after review.

Leaderboard Composition

What are the instances? (that is, examples; e.g., documents, images, people, countries) Are there multiple types of instances? (e.g., movies, users, ratings; people, interactions between them; nodes, edges)

La Leaderboard consists of 66 evaluation datasets. All the evaluation datasets in the leaderboard consist solely of text instances.

Are relationships between instances made explicit in the data (e.g., social network links, user/movie ratings, etc.)?

There are no known relationships between instances.

How many instances of each type are there? Summing all the instances of the 66 evaluation datasets, the leaderboard consists of 149,782 examples.

Is everything included or does the data rely on external resources? (e.g., websites, tweets, datasets) If external resources, a) are there guarantees that they will exist, and remain constant, over time; b) is there an official archival version. Are there licenses, fees or rights associated with any of the data?

Everything is included in the datasets.

Are there recommended data splits or evaluation measures? (e.g., training, development, testing; accuracy/AUC)

The splits used in La Leaderboard are the corresponding test splits of each dataset.

Data Collection Process

How was the data collected? (e.g., hardware apparatus/sensor, manual human curation, software program, software interface/API; how were these constructs/measures/methods validated?)

The datasets were collected through an open data collection campaign.

Who was involved in the data collection process? (e.g., students, crowdworkers) How were they compensated? (e.g., how much were crowdworkers paid?)

Professional researchers from academia and industry. The logo and names of the donators are included in the user interface, and the creators of the datasets are acknowledged in the paper.

Over what time-frame was the data collected? Does the collection time-frame match the creation time-frame?

During 2024.

Does the dataset contain all possible instances? Or is it, for instance, a sample (not necessarily random) from a larger set of instances?

The evaluations are launched including all the available test instances for each donated dataset.

If the dataset is a sample, then what is the population? What was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? Is the sample representative of the larger set (e.g., geographic coverage)? If not, why not (e.g., to cover a more diverse range of instances)? How does this affect possible uses?

Not applicable.

Is there information missing from the dataset and why? (this does not include intentionally dropped instances; it might include, e.g., redacted text, withheld documents) Is this data missing because it was unavailable?

Are there any known errors, sources of noise, or redundancies in the data? No.

Leaderboard Distribution

How is the leaderboard distributed? (e.g., website, API, etc.; does the data have a DOI; is it archived redundantly?)

The leaderboard is available in the HuggingFace hub: URL.

When will the leaderboard be released/first distributed? (Is there a canonical paper/reference for this dataset?)

The leaderboard was released in September 2024.

What license (if any) is it distributed under? Are there any copyrights on the data?

The leaderboard is licensed under "Apache 2.0".

Are there any fees or access/export restrictions?

There are no fees or restrictions.

Leaderboard Maintenance

Who is supporting/hosting/maintaining the leaderboard? How does one contact the owner/curator/manager of the leaderboard (e.g. email address, or other contact info)?

The leaderboard is hosted at HuggingFace[28] and the community can be contacted through the "Discussions" tab in the interface or via email.[29]

Will the leaderboard be updated? How often and by whom? How will updates/revisions be documented and communicated (e.g., mailing list, GitHub)? Is there an erratum?

Yes, every time there is a new donation the maintainer will update the leaderboard and communicate the update on the usual communication channels of the open-source community.

Is there a repository to link to any/all papers/systems that use this leaderboard? Yes, all the datasets and tools used by LA LEADERBOARD are referenced in the "Citation" section of the interface.[30]

Legal & Ethical Considerations

If the dataset relates to people (e.g., their attributes) or was generated by people, were they informed about the data collection? (e.g., datasets that collect writing, photos, interactions, transactions, etc.)

Not applicable.

If it relates to other ethically protected subjects, have appropriate obligations been met? (e.g., medical data might include information collected from animals)

Not applicable.

If it relates to people, were there any ethical review applications/reviews/approvals? (e.g. Institutional Review Board applications)

Not applicable.

If it relates to people, were they told what the dataset would be used for and did they consent?

What community norms exist for data collected from human communications? If consent was obtained, how? Were the people provided with any mechanism to revoke their consent in the future or for certain uses?

Not applicable.

If it relates to people, could this dataset expose people to harm or legal action? (e.g., financial social or otherwise) What was done to mitigate or reduce the potential for harm?

Not applicable.

If it relates to people, does it unfairly advantage or disadvantage a particular social group? In what ways? How was this mitigated?

Not applicable.

If it relates to people, were they provided with privacy guarantees? If so, what guarantees and how are these ensured?

Not applicable.

Does the dataset comply with the EU General Data Protection Regulation (GDPR)? Does it comply with any other standards, such as the US Equal Employment Opportunity Act?

Yes, it complies with GDPR.

[28] Link removed for review.

[29] Email removed for review.

[30] Link removed for review.

Does the dataset contain information that might be considered sensitive or confidential? (e.g., personally identifying information)

No.

Does the dataset contain information that might be considered inappropriate or offensive? No.

AQuAS

The Abstractive Question-Answering in Spanish (AQuAS) dataset (Cite removed for review) developed by a research institute,[31] is a monolingual Spanish dataset designed for abstractive question-answering. It contains 107 examples covering a diverse range of topics, including finance, insurance, healthcare, music, and law. Each example consists of a context passage, a related question, and a human-crafted answer. The dataset is aimed at evaluating the ability of large language models (LLMs) to generate well-formed, coherent, and informative responses.

Motivation for Dataset Creation

Why was the dataset created? (e.g., were there specific tasks in mind, or a specific gap that needed to be filled?) AQuAS was created to provide high-quality examples of pairs of questions and answers with a related context that can be used to evaluate the ability of large language models (LLMs) to generate well-formed, coherent, and informative responses (abstractive question answering).

What (other) tasks could the dataset be used for? Are there obvious tasks for which it should not be used? There are no recommended uses for this dataset other than evaluation.

Who funded the creation of the dataset? If there is an associated grant, provide the grant number. The dataset was created and funded by the research institute.

Dataset Composition

What are the instances? (that is, examples; e.g., documents, images, people, countries) Are there multiple types of instances? (e.g., movies, users, ratings; people, interactions between them; nodes, edges) Each instance is a pair of a question and an answer accompanied by the related context on which the answer has been based and the corresponding topic.

Are relationships between instances made explicit in the data (e.g., social network links, user/movie ratings, etc.)? There are no known relationships between instances.

How many instances of each type are there? The dataset consists of 107 examples.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images)? Features/attributes? Is there a label/target associated with instances? If the instances are related to people, are subpopulations identified (e.g., by age, gender, etc.) and what is their distribution? The instances consist of text data and are labelled with the corresponding topic.

Is everything included or does the data rely on external resources? (e.g., websites, tweets, datasets) If external resources, a) are there guarantees that they will exist, and remain constant, over time; b) is there an official archival version. Are there licenses, fees or rights associated with any of the data? Everything is included in the dataset.

Are there recommended data splits or evaluation measures? (e.g., training, development, testing; accuracy/AUC) Since the dataset is intended for testing, there is no recommended split.

Data Collection Process

How was the data collected? (e.g., hardware apparatus/sensor, manual human curation, software program, software interface/API; how were these constructs/measures/methods validated?) The data for the contexts was gathered from different sources on the web using software to crawl those sites. The rest of the dataset (question-answer pairs) was curated and created manually.

Who was involved in the data collection process? (e.g., students, crowdworkers) How were they compensated? (e.g., how much were crowdworkers paid?) The data was collected by computational linguists and data scientists from a research institute.

Over what time-frame was the data collected? Does the collection time-frame match the creation time-frame? The data was collected during 2023, when the dataset was created.

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part of speech tags; model-based guesses for age or language)? If the latter two, were they validated/verified and if so how? The question-answer pairs were created and revised by computational linguists.

[31] Name removed for review.

Does the dataset contain all possible instances? Or is it, for instance, a sample (not necessarily random) from a larger set of instances? The dataset is composed of selected instances of different datasets created by a research institute.

If the dataset is a sample, then what is the population? What was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? Is the sample representative of the larger set (e.g., geographic coverage)? If not, why not (e.g., to cover a more diverse range of instances)? How does this affect possible uses? This dataset is a 24,5% sample of the original complete datasets. The instances were randomly selected from the original datasets.

Is there information missing from the dataset and why? (this does not include intentionally dropped instances; it might include, e.g., redacted text, withheld documents) Is this data missing because it was unavailable? There is no data missing.

Are there any known errors, sources of noise, or redundancies in the data? There are no known errors because the revision process ensured the data is as clean and error free as possible.

Data Preprocessing

What preprocessing/cleaning was done? (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values, etc.) The text contained in context part of each instance in the dataset has not undergone any preprocessing or changes. There was no need to apply any cleaning to the question-answer pairs because they were created manually by computational linguists following a rigorous methodology and were subjected to revision afterwards.

Was the "raw" data saved in addition to the preprocessed/clean data? (e.g., to support unanticipated future uses) No, the text in the dataset is the raw data.

Is the preprocessing software available? No preprocessing software was used.

Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet? Yes, the collection procedure ensures the dataset is sufficiently varied so it can be used to evaluate a model on a wide range of topics. However, there are some potential limitations in the dataset which might slightly bias the data towards particular topics, because not all topics included have the exact same representation in the dataset, and obviously it was not possible to cover all topics in existence.

Dataset Distribution

How is the dataset distributed? (e.g., website, API, etc.; does the data have a DOI; is it archived redundantly?) The dataset is available in HuggingFace.[32]

When will the dataset be released/first distributed? (Is there a canonical paper/reference for this dataset?) The dataset was released in 2024.

What license (if any) is it distributed under? Are there any copyrights on the data? The dataset is licensed under CC BY-NC-SA 4.0.

Are there any fees or access/export restrictions? There are no fees or restrictions.

Dataset Maintenance

Who is supporting/hosting/maintaining the dataset? How does one contact the owner/curator/manager of the dataset (e.g. email address, or other contact info)? The dataset is hosted at HuggingFace and the research institute can be contacted through email.[33]

Will the dataset be updated? How often and by whom? How will updates/revisions be documented and communicated (e.g., mailing list, GitHub)? Is there an erratum? It is not planned to update the dataset at the moment.

Is there a repository to link to any/all papers/systems that use this dataset? No.

Legal & Ethical Considerations

If the dataset relates to people (e.g., their attributes) or was generated by people, were they informed about the data collection? (e.g., datasets that collect writing, photos, interactions, transactions, etc.) Not applicable. The data was collected from public web sources, and does not contain sensitive personal information.

[32] Link removed for review.

[33] Email removed for review.

If it relates to other ethically protected subjects, have appropriate obligations been met? (e.g., medical data might include information collected from animals) Not applicable.

If it relates to people, were there any ethical review applications/reviews/approvals? (e.g. Institutional Review Board applications) Not applicable.

If it relates to people, were they told what the dataset would be used for and did they consent? What community norms exist for data collected from human communications? If consent was obtained, how? Were the people provided with any mechanism to revoke their consent in the future or for certain uses? Not applicable.

If it relates to people, could this dataset expose people to harm or legal action? (e.g., financial social or otherwise) What was done to mitigate or reduce the potential for harm? Not applicable.

If it relates to people, does it unfairly advantage or disadvantage a particular social group? In what ways? How was this mitigated? Not applicable.

If it relates to people, were they provided with privacy guarantees? If so, what guarantees and how are these ensured? Not applicable.

Does the dataset comply with the EU General Data Protection Regulation (GDPR)? Does it comply with any other standards, such as the US Equal Employment Opportunity Act? The dataset complies with GDPR.

Does the dataset contain information that might be considered sensitive or confidential? (e.g., personally identifying information) No.

Does the dataset contain information that might be considered inappropriate or offensive? No.

ClinTreatES

The ClinTreatES (Cite removed for review) dataset consists of clinical cases collected directly from doctors in various medical specialties (cardiology, traumatology, emergency, psychiatry, neurology, dermatology, ENT-laryngology, and anaesthesia) across European medical centers. It was developed through a joint collaboration between an NLP startup[34] and healthcare professionals. The dataset is intended for evaluating the ability of large language models (LLMs) to generate effective treatment plans based on provided clinical cases and diagnoses.

Motivation for Dataset Creation

Why was the dataset created? ClinTreatES was created to evaluate LLMs' capability to design appropriate treatments from real clinical cases and their corresponding diagnoses.

What (other) tasks could the dataset be used for? In addition to treatment planning, the dataset may be used to study medical reasoning and decision-making; however, it is not recommended for diagnostic tasks.

Who funded the creation of the dataset? The dataset was developed through a collaboration between an NLP startup and healthcare professionals.

Dataset Composition

What are the instances? Each instance comprises a clinical case description and its associated diagnosis.

Are relationships between instances made explicit in the data? No, there are no explicit relationships between instances.

How many instances of each type are there? The dataset contains 62 examples.

What data does each instance consist of? Each instance includes text data: a clinical case and its corresponding diagnosis, which serves as the basis for generating a treatment plan.

Is everything included or does the data rely on external resources? The dataset is self-contained with no reliance on external resources.

Are there recommended data splits or evaluation measures? No specific splits are recommended; the dataset is intended primarily for evaluation purposes.

Data Collection Process

How was the data collected? Data was collected directly from healthcare professionals across various specialties in European medical centers.

Who was involved in the data collection process? Medical professionals from cardiology, traumatology, emergency medicine, psychiatry, neurology, dermatology, ENT-laryngology, and anesthesia contributed to the dataset.

Over what time-frame was the data collected? The data was collected in 2024.

How was the data associated with each instance acquired? Clinical cases and their corresponding diagnoses were directly provided by the contributing healthcare professionals.

Does the dataset contain all possible instances? It is a curated collection and does not cover every possible clinical case.

If the dataset is a sample, then what is the population? The dataset represents a curated sample of clinical cases from European medical centers.

Is there information missing from the dataset and why? No, all relevant information is included.

Are there any known errors, sources of noise, or redundancies in the data? The data has been carefully curated and reviewed to minimize errors and noise.

Data Preprocessing

What preprocessing/cleaning was done? The clinical texts were formatted according to a standardized template; only minimal preprocessing was applied.

Was the "raw" data saved in addition to the preprocessed/clean data? Yes, the dataset contains the original clinical texts as provided by the contributors.

Is the preprocessing software available? No specific preprocessing software was used.

[34] Name removed for review.

Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet? Yes, the collection and curation process ensures the dataset is suitable for evaluating treatment design tasks by LLMs.

Dataset Distribution

How is the dataset distributed? The dataset is available on HuggingFace.[35]

When will the dataset be released/first distributed? The dataset was released in March 2024.

What license (if any) is it distributed under? It is distributed under the CC BY-NC-SA 4.0 license.

Are there any fees or access/export restrictions? There are no fees or restrictions.

Dataset Maintenance

Who is supporting/hosting/maintaining the dataset? The dataset is hosted on HuggingFace and maintained by the NLP startup.

Will the dataset be updated? How often and by whom? No updates are planned at this time.

Is there a repository to link to any/all papers/systems that use this dataset? The dataset is available on HuggingFace.[36]

Legal & Ethical Considerations

If the dataset relates to people, were they informed about the data collection? The clinical cases were provided by healthcare professionals; any personal details have been removed to ensure anonymity. They were anonymized by the healthcare professionals themselves, before transferring the data to the NLP startup.

If it relates to other ethically protected subjects, have appropriate obligations been met? Yes, all obligations have been met and ensured in the data collection process.

If it relates to people, were there any ethical review applications/reviews/approvals? Yes, healthcare professionals ensured the ethical review was complete.

If it relates to people, were they told what the dataset would be used for and did they consent? Yes, patients were told in advance about the objective of data collection and they provided their consent for this use.

If it relates to people, could this dataset expose people to harm or legal action? No, as the data is anonymized by the healthcare professionals.

If it relates to people, does it unfairly advantage or disadvantage a particular social group? No.

If it relates to people, were they provided with privacy guarantees? Yes, all personal information has been removed by the healthcare professionals.

Does the dataset comply with the EU General Data Protection Regulation (GDPR)? Yes, it complies with GDPR.

Does the dataset contain information that might be considered sensitive or confidential? No, all potentially sensitive or confidential information has been removed.

Does the dataset contain information that might be considered inappropriate or offensive? No.

[35] Link removed for review.

[36] Link removed for review.

ClinDiagnosES

The ClinDiagnosES (Cite removed for review) dataset comprises clinical cases accompanied by corresponding diagnoses, collected directly from healthcare professionals across multiple specialties in Europe. It is intended for evaluating LLMs' diagnostic reasoning abilities.

Motivation for Dataset Creation

Why was the dataset created? ClinDiagnosES was created to assess the ability of LLMs to generate accurate diagnoses based on clinical case descriptions.

What (other) tasks could the dataset be used for? Besides diagnostic evaluation, it can be used to study medical reasoning; however, it is not suitable for treatment planning tasks.

Who funded the creation of the dataset? The dataset was developed through a collaboration between an NLP startup[37] and healthcare professionals.

Dataset Composition

What are the instances? Each instance consists of a clinical case description along with its corresponding diagnosis.

Are relationships between instances made explicit in the data? No, there are no explicit relationships between instances.

How many instances of each type are there? The dataset contains 62 examples.

What data does each instance consist of? Each instance includes text data representing a clinical case and its associated diagnosis.

Is everything included or does the data rely on external resources? The dataset is self-contained.

Are there recommended data splits or evaluation measures? No specific splits are recommended; it is intended for evaluation purposes.

Data Collection Process

How was the data collected? Data was collected directly from healthcare professionals across various medical specialties.

Who was involved in the data collection process? Healthcare professionals from fields such as cardiology, traumatology, emergency medicine, psychiatry, neurology, dermatology, ENT-laryngology, and anesthesia contributed.

Over what time-frame was the data collected? The data was collected in 2024.

How was the data associated with each instance acquired? Each clinical case was accompanied by a diagnosis provided by a medical expert.

Does the dataset contain all possible instances? It is a curated collection and does not encompass every possible clinical case.

If the dataset is a sample, then what is the population? The dataset represents a curated sample of clinical cases from European medical centers.

Is there information missing from the dataset and why? No, all necessary information is included.

Are there any known errors, sources of noise, or redundancies in the data? The dataset has been reviewed to minimize errors and inconsistencies.

Data Preprocessing

What preprocessing/cleaning was done? The clinical cases and diagnoses were formatted using a standardized template with minimal cleaning.

Was the "raw" data saved in addition to the preprocessed/clean data? Yes, the raw clinical texts and diagnoses are preserved.

Is the preprocessing software available? No specific preprocessing software was utilized.

Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet? Yes, the procedure ensures the dataset is suitable for evaluating diagnostic reasoning in LLMs.

Dataset Distribution

How is the dataset distributed? The dataset is available on HuggingFace.[38]

When will the dataset be released/first distributed? It was released in March 2024.

[37] Name removed for review.

[38] Link removed for review.

What license (if any) is it distributed under? It is distributed under the CC BY-NC-SA 4.0 license.

Are there any fees or access/export restrictions? There are no fees or restrictions.

Dataset Maintenance

Who is supporting/hosting/maintaining the dataset? The dataset is hosted on HuggingFace and maintained by the NLP startup.

Will the dataset be updated? How often and by whom? No updates are planned at this time.

Is there a repository to link to any/all papers/systems that use this dataset? The dataset is available on HuggingFace.[39]

Legal & Ethical Considerations

If the dataset relates to people, were they informed about the data collection? The clinical cases were provided by healthcare professionals; any personal details have been removed to ensure anonymity. They were anonymized by the healthcare professionals themselves, before transferring the data to the NLP startup.

If it relates to other ethically protected subjects, have appropriate obligations been met? Yes, all obligations have been met and ensured in the data collection process.

If it relates to people, were there any ethical review applications/reviews/approvals? Yes, healthcare professionals ensured the ethical review was complete.

If it relates to people, were they told what the dataset would be used for and did they consent? Yes, patients were told in advance about the objective of data collection and they provided their consent for this use.

If it relates to people, could this dataset expose people to harm or legal action? No, as the data is anonymized by the healthcare professionals.

If it relates to people, does it unfairly advantage or disadvantage a particular social group? No.

If it relates to people, were they provided with privacy guarantees? Yes, all personal information has been removed by the healthcare professionals.

Does the dataset comply with the EU General Data Protection Regulation (GDPR)? Yes, it complies with GDPR.

Does the dataset contain information that might be considered sensitive or confidential? No, all potentially sensitive or confidential information has been removed.

Does the dataset contain information that might be considered inappropriate or offensive? No.

[39] Link removed for review.

HumorQA

The HumorQA dataset (Cite removed for review), developed collaboratively by an NLP startup and a psychology consulting firm,[40] focuses on humor classification. It consists of jokes paired with labels indicating the joke type: C/E (Comparison/Exaggeration), JP (Wordplay), R3 (Rule of Three) and AI (Animating the Inanimate). The data set is based on a study involving 94 executives and is intended to evaluate the ability of LLMs to understand and classify humor.

Motivation for Dataset Creation

Why was the dataset created? HumorQA was created to assess the ability of LLMs to recognize and classify different types of humor.

What (other) tasks could the dataset be used for? It can also be used for research on sentiment analysis and humor recognition, although its primary purpose is humor classification.

Who funded the creation of the dataset? The dataset was developed through a collaboration between an NLP startup and a psychology consulting firm.

Dataset Composition

What are the instances? Each instance comprises a joke and its corresponding humor-type label.

Are relationships between instances made explicit in the data? No, there are no explicit relationships between instances.

How many instances of each type are there? The dataset contains 51 examples.

What data does each instance consist of? Each instance includes text data representing a joke and a label indicating its humor category.

Is everything included or does the data rely on external resources? The dataset is self-contained.

Are there recommended data splits or evaluation measures? No specific splits are recommended; it is intended for evaluation purposes.

Data Collection Process

How was the data collected? Jokes were collected and curated as part of a research study involving humor workshops and interviews with 94 executives.

Who was involved in the data collection process? The data collection involved humor experts at Human Profit Consulting along with participating executives.

Over what time-frame was the data collected? The data was collected in 2024.

How was the data associated with each instance acquired? Jokes were labeled according to a predefined categorization based on the study's methodology.

Does the dataset contain all possible instances? It is a curated sample representing various humor styles.

If the dataset is a sample, then what is the population? The sample represents humorous content identified in a study with executives from diverse sectors.

Is there information missing from the dataset and why? No, all relevant information is included.

Are there any known errors, sources of noise, or redundancies in the data? The dataset has been thoroughly reviewed; no significant errors or redundancies have been identified.

Data Preprocessing

What preprocessing/cleaning was done? The jokes and labels were formatted into a standardized template with minimal preprocessing.

Was the "raw" data saved in addition to the preprocessed/clean data? Yes, the original joke texts are preserved.

Is the preprocessing software available? No specific preprocessing software was used.

Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet? Yes, the curation process supports the evaluation of humor classification by LLMs.

[40] Names removed for review.

Dataset Distribution

How is the dataset distributed? The dataset is available on HuggingFace.[41]

When will the dataset be released/first distributed? It was released in March 2024.

What license (if any) is it distributed under? It is distributed under the CC BY-NC-SA 4.0 license.

Are there any fees or access/export restrictions? There are no fees or restrictions.

Dataset Maintenance

Who is supporting/hosting/maintaining the dataset? The dataset is hosted on HuggingFace by the NLP startup.

Will the dataset be updated? How often and by whom? No updates are planned at this time.

Is there a repository to link to any/all papers/systems that use this dataset? The dataset is available on HuggingFace.[42]

Legal & Ethical Considerations

If the dataset relates to people, were they informed about the data collection? The dataset is based on humorous content and research; it does not involve personal data.

If it relates to other ethically protected subjects, have appropriate obligations been met? Not applicable.

If it relates to people, were there any ethical review applications/reviews/approvals? Not applicable.

If it relates to people, were they told what the dataset would be used for and did they consent? Not applicable.

If it relates to people, could this dataset expose people to harm or legal action? No.

If it relates to people, does it unfairly advantage or disadvantage a particular social group? No.

If it relates to people, were they provided with privacy guarantees? Not applicable.

Does the dataset comply with the EU General Data Protection Regulation (GDPR)? Yes, it complies with GDPR.

Does the dataset contain information that might be considered sensitive or confidential? No.

Does the dataset contain information that might be considered inappropriate or offensive? No.

[41] Link removed for review.

[42] Link removed for review.

RagQuAS

The Retrieval-Augmented-Generation and Question-Answering in Spanish (RagQuAS) dataset (Cite removed for review), created by a research institute,[43] is a high-quality monolingual Spanish dataset designed to evaluate models in retrieval-augmented generation (RAG) and question-answering tasks. It consists of 201 examples covering a wide range of knowledge domains, such as hobbies, linguistics, health, astronomy, and customer service. Each example includes a question, multiple context passages extracted from different documents, and a gold-standard answer. This dataset is particularly useful for assessing a model's ability to retrieve relevant information from multiple sources and generate accurate, contextually appropriate responses.

Motivation for Dataset Creation

Why was the dataset created? (e.g., were there specific tasks in mind, or a specific gap that needed to be filled?) RagQuAS was created to provide high-quality examples of questions and answers with related contexts that can be used to evaluate models in retrieval-augmented generation (RAG) and question-answering tasks.

What (other) tasks could the dataset be used for? Are there obvious tasks for which it should not be used? There are no recommended uses for this dataset other than evaluation.

Who funded the creation of the dataset? If there is an associated grant, provide the grant number. The dataset was created and funded by Instituto de Ingeniería de Conocimiento.

Dataset Composition

What are the instances? (that is, examples; e.g., documents, images, people, countries) Are there multiple types of instances? (e.g., movies, users, ratings; people, interactions between them; nodes, edges) Each instance consists of several categories of text: the topic, a question, an indicator of the variant of the question (this represents questions with linguistic differences but pertaining to the same contexts than other questions), an answer, one to five contexts, one to five complete documents from where the contexts were extracted and the links to these documents.

Are relationships between instances made explicit in the data (e.g., social network links, user/movie ratings, etc.)? There are no known relationships between instances.

How many instances of each type are there? The dataset consists of 201 examples.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images)? Features/attributes? Is there a label/target associated with instances? If the instances are related to people, are subpopulations identified (e.g., by age, gender, etc.) and what is their distribution? The instances consist of text data and are labeled with the corresponding topic.

Is everything included or does the data rely on external resources? (e.g., websites, tweets, datasets) If external resources, a) are there guarantees that they will exist, and remain constant, over time; b) is there an official archival version. Are there licenses, fees or rights associated with any of the data? Everything is included in the dataset.

Are there recommended data splits or evaluation measures? (e.g., training, development, testing; accuracy/AUC) Since the dataset is intended for testing, there is no recommended split.

Data Collection Process

How was the data collected? (e.g., hardware apparatus/sensor, manual human curation, software program, software interface/API; how were these constructs/measures/methods validated?) The data for the contexts was gathered from different sources manually with the help of generative models (to suggest web searches and results). The rest of the dataset was curated and created manually.

Who was involved in the data collection process? (e.g., students, crowdworkers) How were they compensated? (e.g., how much were crowdworkers paid?) The data was collected by computational linguists and data scientists from the research institute.

Over what time-frame was the data collected? Does the collection time-frame match the creation time-frame? The data was collected during 2023, when the dataset was created.

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g.,

[43] Name removed for review.

survey responses), or indirectly inferred/derived from other data (e.g., part of speech tags; model-based guesses for age or language)? If the latter two, were they validated/verified and if so how? The question-answer pairs were created and revised by computational linguists.

Does the dataset contain all possible instances? Or is it, for instance, a sample (not necessarily random) from a larger set of instances? The dataset is composed of selected instances of a dataset created by the research institute.

If the dataset is a sample, then what is the population? What was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? Is the sample representative of the larger set (e.g., geographic coverage)? If not, why not (e.g., to cover a more diverse range of instances)? How does this affect possible uses? This dataset is a 24% sample of the original complete datasets. The instances were randomly selected from the original dataset.

Is there information missing from the dataset and why? (this does not include intentionally dropped instances; it might include, e.g., redacted text, withheld documents) Is this data missing because it was unavailable? There is no data missing.

Are there any known errors, sources of noise, or redundancies in the data? There are no known errors because the revision process ensured the data is as clean and error free as possible.

Data Preprocessing

What preprocessing/cleaning was done? (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values, etc.) The text contained in context and document part of each instance in the dataset has not undergone any preprocessing or changes. The questions were created manually by computational linguists following a rigorous methodology and were subjected to revision afterwards. The answers were carefully curated and revised by linguists from generated texts.

Was the "raw" data saved in addition to the preprocessed/clean data? (e.g., to support unanticipated future uses) No, the text in the dataset is the raw data.

Is the preprocessing software available? No preprocessing software was used.

Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet? Yes, the methodology used when creating the dataset ensures it is sufficiently varied so it can be used to evaluate a model on a wide range of topics. However, there are some potential limitations in the dataset which might slightly bias the data towards particular topics, because not all topics included have the exact same representation in the dataset, and obviously it was not possible to cover all topics in existence.

Dataset Distribution

How is the dataset distributed? (e.g., website, API, etc.; does the data have a DOI; is it archived redundantly?) The dataset is available in HuggingFace.[44]

When will the dataset be released/first distributed? (Is there a canonical paper/reference for this dataset?) The dataset was released in 2024.

What license (if any) is it distributed under? Are there any copyrights on the data? The dataset is licensed under CC BY-NC-SA 4.0.

Are there any fees or access/export restrictions? There are no fees or restrictions.

Dataset Maintenance

Who is supporting/hosting/maintaining the dataset? How does one contact the owner/curator/manager of the dataset (e.g. email address, or other contact info)? The dataset is hosted at HuggingFace and the research institute can be contacted through email.[45]

Will the dataset be updated? How often and by whom? How will updates/revisions be documented and communicated (e.g., mailing list, GitHub)? Is there an erratum? It is not planned to update the dataset at the moment.

Is there a repository to link to any/all papers/systems that use this dataset? No.

[44] Link removed for review.

[45] Email removed for review.

Legal & Ethical Considerations

If the dataset relates to people (e.g., their attributes) or was generated by people, were they informed about the data collection? (e.g., datasets that collect writing, photos, interactions, transactions, etc.) Not applicable. The data was collected from public web sources, and does not contain sensitive personal information.

If it relates to other ethically protected subjects, have appropriate obligations been met? (e.g., medical data might include information collected from animals) Not applicable.

If it relates to people, were there any ethical review applications/reviews/approvals? (e.g. Institutional Review Board applications) Not applicable.

If it relates to people, were they told what the dataset would be used for and did they consent? What community norms exist for data collected from human communications? If consent was obtained, how? Were the people provided with any mechanism to revoke their consent in the future or for certain uses? Not applicable.

If it relates to people, could this dataset expose people to harm or legal action? (e.g., financial social or otherwise) What was done to mitigate or reduce the potential for harm? Not applicable.

If it relates to people, does it unfairly advantage or disadvantage a particular social group? In what ways? How was this mitigated? Not applicable.

If it relates to people, were they provided with privacy guarantees? If so, what guarantees and how are these ensured? Not applicable.

Does the dataset comply with the EU General Data Protection Regulation (GDPR)? Does it comply with any other standards, such as the US Equal Employment Opportunity Act? The dataset complies with GDPR.

Does the dataset contain information that might be considered sensitive or confidential? (e.g., personally identifying information) No.

Does the dataset contain information that might be considered inappropriate or offensive? No.

SpaLawEx

The SpaLawEx dataset (Cite removed for review) consists of multiple-choice legal questions extracted from Spanish Bar Examination papers of 2022 and 2023. Each instance includes a legal question along with four answer options (A, B, C, and D).

Motivation for Dataset Creation

Why was the dataset created? SpaLawEx was created to evaluate the legal reasoning and knowledge of LLMs within the context of Spanish law using multiple-choice questions.

What (other) tasks could the dataset be used for? In addition to benchmarking legal question answering systems, it may be used for legal education; it is not intended for non-legal tasks.

Who funded the creation of the dataset? The dataset was developed by an NLP startup, with contributions from legal experts.

Dataset Composition

What are the instances? Each instance is a multiple-choice legal question accompanied by four answer options.

Are relationships between instances made explicit in the data? No, there are no explicit relationships between instances.

How many instances of each type are there? The dataset contains 119 examples.

What data does each instance consist of? Each instance comprises text data including a legal question and its four answer options (A, B, C, and D).

Is everything included or does the data rely on external resources? The dataset is self-contained, extracted from publicly available examination papers.

Are there recommended data splits or evaluation measures? No specific splits are recommended; the dataset is intended for evaluation purposes.

Data Collection Process

How was the data collected? Data were extracted from official Spanish Bar Examination papers from 2022 and 2023.

Who was involved in the data collection process? The extraction was performed by the developers at an NLP startup, with input from legal experts.

Over what time-frame was the data collected? The data was collected in 2024.

How was the data associated with each instance acquired? Questions and answer options were directly extracted from exam documents.

Does the dataset contain all possible instances? It is a comprehensive collection of questions from the specified examination periods. However, it is not exhaustive and it does not contain all possible instances.

If the dataset is a sample, then what is the population? It represents the pool of questions from the Spanish Bar Examinations of 2022 and 2023.

Is there information missing from the dataset and why? No, all relevant information is included.

Are there any known errors, sources of noise, or redundancies in the data? The dataset has been checked for accuracy; any minor extraction errors are not known to be significant.

Data Preprocessing

What preprocessing/cleaning was done? The exam questions and answer options were formatted into a standardized template with minimal cleaning.

Was the "raw" data saved in addition to the preprocessed/clean data? Yes, the original extracted text is preserved.

Is the preprocessing software available? No specific preprocessing software was used.

Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet? Yes, the process ensures the dataset is suitable for evaluating legal reasoning in LLMs.

Dataset Distribution

How is the dataset distributed? The dataset is available on HuggingFace.[46]

When will the dataset be released/first distributed? It was released in March 2024.

[46] Link removed for review.

What license (if any) is it distributed under? It is distributed under the CC BY-NC-SA 4.0 license.

Are there any fees or access/export restrictions? There are no fees or restrictions.

Dataset Maintenance

Who is supporting/hosting/maintaining the dataset? The dataset is hosted on HuggingFace by the NLP startup.

Will the dataset be updated? How often and by whom? No updates are planned at this time.

Is there a repository to link to any/all papers/systems that use this dataset? No repository has been provided.

Legal & Ethical Considerations

If the dataset relates to people, were they informed about the data collection? The dataset is derived from public examination materials and does not involve personal data.

If it relates to other ethically protected subjects, have appropriate obligations been met? Not applicable.

If it relates to people, were there any ethical review applications/reviews/approvals? Not applicable.

If it relates to people, were they told what the dataset would be used for and did they consent? Not applicable.

If it relates to people, could this dataset expose people to harm or legal action? No.

If it relates to people, does it unfairly advantage or disadvantage a particular social group? No.

If it relates to people, were they provided with privacy guarantees? Not applicable.

Does the dataset comply with the EU General Data Protection Regulation (GDPR)? Yes, it complies with GDPR.

Does the dataset contain information that might be considered sensitive or confidential? No.

Does the dataset contain information that might be considered inappropriate or offensive? No.

TELEIA

The TELEIA (Cite removed for review) dataset is intended for the evaluation of Spanish language knowledge focusing on reading comprehension and grammatical competence. The dataset is designed as a set of multiple-choice questions that have the same format and level as those used in several Spanish evaluation tests for humans. The questions are divided into three blocks which resemble existing tests of Spanish for foreign learners and University access. In total, one hundred questions are included that have been prepared and revised by experts on Spanish language, and that have been validated by comparing the results with the original exams.

Motivation for Dataset Creation

Why was the dataset created? The main motivation was to have a simple test to evaluate the competence of LLMs in Spanish similar to tests used with humans.

What (other) tasks could the dataset be used for? The test also checks reading comprehension and thus can be used to evaluate natural language understanding.

Who funded the creation of the dataset? Funding details removed for review.

Dataset Composition

What are the instances? The test is made of multiple choice questions.

Are relationships between instances made explicit in the data No.

How many instances of each type are there? The dataset consists of 100 questions.

What data does each instance consist of? Each question has a text presenting the question and four answer options of which only one is correct.

Is everything included or does the data rely on external resources? Everything is included in the dataset.

Are there recommended data splits or evaluation measures? No.

Data Collection Process

How was the data collected? Questions were formulated and peer reviewed by experts in Spanish.

Who was involved in the data collection process? Experts in Spanish that participated as researchers in our group.

Over what time-frame was the data collected? The questions were created during the spring of 2024.

How was the data associated with each instance acquired? Data was created by experts.

Does the dataset contain all possible instances? Questions are examples, and many other similar questions can be formulated.

If the dataset is a sample, then what is the population? Not applicable.

Is there information missing from the dataset and why? No.

Are there any known errors, sources of noise, or redundancies in the data? No.

Data Preprocessing

What preprocessing/cleaning was done? None.

Was the "raw" data saved in addition to the preprocessed/clean data? Not applicable.

Is the preprocessing software available? Not applicable.

Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet? Yes.

Dataset Distribution

How is the dataset distributed? Websites.

When will the dataset be released/first distributed? Data is available since July 2024.

What license (if any) is it distributed under? No license or restrictions are applicable.

Are there any fees or access/export restrictions? No.

Dataset Maintenance

Who is supporting/hosting/maintaining the dataset? The dataset is hosted at Zenodo[47] providing contact details for all the authors.

Will the dataset be updated? No updates are expected but the repository supports versioning.

[47] Link removed for review.

Is there a repository to link to any/all papers/systems that use this dataset? No.

Legal & Ethical Considerations

If the dataset relates to people (e.g., their attributes) or was generated by people, were they informed about the data collection? Not applicable.

If it relates to other ethically protected subjects, have appropriate obligations been met? Not applicable.

If it relates to people, were there any ethical review applications/reviews/approvals? Not applicable.

If it relates to people, were they told what the dataset would be used for and did they consent? Not applicable.

If it relates to people, could this dataset expose people to harm or legal action? Not applicable.

If it relates to people, does it unfairly advantage or disadvantage a particular social group? Not applicable.

If it relates to people, were they provided with privacy guarantees? Not applicable.

Does the dataset comply with the EU General Data Protection Regulation (GDPR)? Yes.

Does the dataset contain information that might be considered sensitive or confidential? No.

Does the dataset contain information that might be considered inappropriate or offensive? No.