IberoBench: A Benchmark for LLM Evaluation in Iberian Languages

Irene Baucells[1] **Javier Aula-Blasco**[1] **Iria de-Dios-Flores**[2] **Silvia Paniagua Suárez**[3]
Naiara Perez[4] **Anna Salles**[1] **Susana Sotelo Docio**[3] **Júlia Falcão**[1] **José Javier Saiz**[1]
Robiert Sepúlveda Torres[?] **Jeremy Barnes**[4] **Pablo Gamallo**[3]
Aitor Gonzalez-Agirre[1] **German Rigau**[4] **Marta Villegas**[1]
[1]Barcelona Supercomputing Center (BSC-CNS)
[2] Departament de Traducció i Ciències del Llenguatge, Universitat Pompeu Fabra (UPF)
[3] Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS-USC)
[4] HiTZ Center - IXA, University of the Basque Country (UPV/EHU)
{irene.baucells, javier.aulablasco}@bsc.es

Abstract

The current best practice to measure the performance of base Large Language Models is to establish a multi-task benchmark that covers a range of capabilities of interest. Currently, however, such benchmarks are only available in a few high-resource languages. To address this situation, we present IberoBench, a multilingual, multi-task benchmark for Iberian languages (i.e., Basque, Catalan, Galician, European Spanish and European Portuguese) built on the LM Evaluation Harness framework. The benchmark consists of 62 tasks divided into 179 subtasks. We evaluate 33 existing LLMs on IberoBench on 0- and 5-shot settings. We also explore the issues we encounter when working with the Harness and our approach to solving them to ensure high-quality evaluation.

1 Introduction

The rise of Large Language Models (LLMs) has led to large improvements in a range of typical NLP tasks, from question answering (Kamalloo et al., 2023) to mathematical reasoning (Hendrycks et al., 2021b). These models are capable of zero- and few-shot learning (Kojima et al., 2022; Wei et al., 2022a), which effectively makes them capable of performing well in a massive multi-task setup (Radford et al., 2019; Hendrycks et al., 2021a).

This situation, however, requires a change in the evaluation paradigm, as nowadays we are often more interested in how well a new model performs on a *range* of tasks, rather than a single task. To quantify this, several multi-task benchmarks have appeared (Lee et al., 2023; Gao et al., 2023), but these are largely restricted to English. Despite a small number of exceptions (Liu et al., 2024; Fenogenova et al., 2024), for the majority of other languages in the world, there is no comparable multi-task benchmark available to evaluate native or multilingual LLMs.

As a step towards improving this situation, we introduce IberoBench, a benchmark for automatic model evaluation focused on the official languages of the Iberian peninsula: European Spanish, Catalan, Basque, Galician and European Portuguese. We build upon the existing infrastructure of Eleuther AI's LM Evaluation Harness (Gao et al., 2023), which allows for new tasks to be implemented easily.[1]

We make the following contributions:

i. We share IberoBench, a benchmark for automatic model evaluation in European Spanish, Catalan, Basque, Galician and European Portuguese with a focus on quality and reproducibility, and make it publicly available through the LM Evaluation Harness.[2] This benchmark includes 22 new evaluation tasks in addition to 40 existing in the literature.

ii. We evaluate 33 small and medium-size base LLMs (monolingual and multilingual) and show that model performance in Iberian languages still is behind state-of-the-art results.

iii. We provide further analysis of errors found in evaluation datasets in LM Evaluation Harness and the limitations to consider when using this framework. We share the solutions and approaches we use to minimize the impact these issues have on evaluation quality.

2 Background

2.1 The state of LLM evaluation

We can broadly categorize LLM evaluation approaches into two main types: automatic and human evaluation.

[1] The Evaluation Harness is well-maintained, widely used in the literature, and, crucially for mid- and low-resource languages, open and free to use and implement.

[2] https://github.com/EleutherAI/lm-evaluation-harness

Proceedings of the 31st International Conference on Computational Linguistics, pages 10491–10519
January 19–24, 2025. ©2025 Association for Computational Linguistics

Automatic evaluation This type of evaluation allows for a fast, reproducible way of assessing a model's performance in downstream tasks. This makes it the most wide-spread type of LLM evaluation, as it is significantly more cost-effective and easy to implement than human evaluation.

For generative models, the prompting paradigm has become a standard automatic evaluation technique, consisting of providing the model with a natural language instruction (*input*) and evaluating the model's response (*output*). As popularized by Brown et al. (2020), prompting can be done in a 0-shot setting or in a few-shot setting. Few-shot prompting has been shown to positively impact model accuracy and is a useful strategy to unveil emergent abilities upon scaling (Wei et al., 2022b). Even though there is not a clear recommended number of examples to be used in few-shot prompting, Brown et al. (2020) show that substantial gains are achieved with 3-5 examples.

Recently, many evaluation datasets have appeared, covering a myriad of tasks from more traditional question answering (QA, Gordon et al., 2012; Clark et al., 2018; Sap et al., 2019) and paraphrasing (Zhang et al., 2019) to those exploring tasks believed to be beyond the capabilities of current models (Srivastava et al., 2023). These datasets evaluate a model's performance by comparing its answer against a predetermined correct answer (or *gold-standard*). This type of evaluation is particularly useful for *base* models (i.e., not fine-tuned for a specific task or domain), as their goal is just to generate, rather than engage in a particular task such as a conversation with a user.

As the number of evaluation dataset increases, efforts have been made to collect some of these datasets in evaluation benchmarks. Earlier efforts include GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) or more recently Stanford CRFM's Holistic Evaluation of Language Models (HELM) (Lee et al., 2023) and Eleuther AI's Evaluation Harness (Gao et al., 2023).

In an attempt to move away from reference answers, LLM Evaluators (or LLM-as-a-judge) make use of a state-of-the-art (SOTA) model to evaluate the outputs of another model (Kim et al., 2024; Ye et al., 2024; Üstün et al., 2024). Such LLM evaluators can return both scores for multiple highly-specific criteria based on any preferred scale, as well as the reasoning behind their choices. Chang et al. (2024) offer some suggestions for choosing the criteria LLM Evaluators can follow. Some studies have shown that this type of evaluation correlates in around 80% with human evaluation (Zheng et al., 2023), with its benefits in scalability and cost-effectiveness. However, a more recent study has highlighted that LLMs can distinguish themselves from other LLMs and humans, and that there is also a direct correlation between self-recognition capability and the strength of self-preference bias (Panickssery et al., 2024).

Human evaluation An alternative to any type of automatic evaluation is human evaluation. In spite of some natural, human biases (Wu and Aji, 2023; Hosking et al., 2024), this type of evaluation allows for more fine-grained feedback and better reflects real-world application scenarios. Some studies have shown that human evaluation is consistently more reliable than automatic evaluation metrics in natural language generation (NLG) tasks (Novikova et al., 2017; Sai et al., 2022). The main limitation for human evaluation is its high cost, both in time and money (Biderman et al., 2024).

Human evaluation can take various shapes, but the two most common ones are direct output assessment and A/B testing. The former implies evaluating a model's output in a similar manner to the LLM Evaluators described previously. The latter implies the comparison of the outputs of two different models by a human evaluator. Some studies have tried to use existing rating systems to adapt them to LLM A/B testing. An example of this is Elo, a method for calculating the relative skill levels of players in zero-sum games such as chess, which has become a popular method for examining NLG performance (Boubdir et al., 2023). While there are many ways to measure human preference, Chang et al. (2024) state that the key factors are diversity in demographic representation in tandem with relevant domain expertise, enabling the use of statistical significance testing.

2.2 A move towards multilinguality

As can be understood from the previous section, LLM evaluation lacks a reliable, comprehensive and reproducible methodology. The evaluation styles discussed all have advantages and disadvantages. Until the scientific epistemology in the field matures, the best option is to develop an evaluation suite that encompasses all evaluation styles. This is currently easy in English, as most of the work has been done in that language, making it impossible to measure the level of performance of LLMs in other languages. However, this English-centric approach does nothing but widen the gap between English and other languages with *moderate*, *fragmentary* or *weak* technological support (Rehm and Way, 2023).

To address this gap, several initiatives have emerged that provide comprehensive benchmark evaluations in non-English languages, similar to our own. Notable initiatives in this regard include IrokoBench (Adelani et al., 2024), a human-translated benchmark for 16 low-resource African languages covering NLI, mathematical reasoning, and QA tasks, and IndicGenBench (Singh et al., 2024), a benchmark focused on generation tasks across 29 Indic languages.

While some languages in the Iberian peninsula have a better status than those entailed in the mentioned efforts, Giagkou et al. (2023) still reports Spanish as a language with *moderate* technology support, Portuguese as *moderate* in tools and services but *fragmentary* in language resources, Catalan and Basque as *fragmentary* in general, and Galician as *weak* in language resources and *fragmentary* in tools and services. Thus, not only does IberoBench offer a quality-oriented multilingual evaluation benchmark, but it does so while working with a set of languages that, given their current situation, can greatly benefit from it.

Having a quality-oriented, standardized and reproducible evaluation benchmark such as IberoBench allows a fair comparison of LLMs, which in turn may foster the development of models for the languages covered by the benchmark, as capabilities can be systematically measured and reported. However, an evaluation benchmark lacks purpose without models to evaluate.

2.3 LLMs for Iberian languages

The landscape of LLMs focused on the languages covered by IberoBench varies significantly depending on the language. The levels of technology support of each language are a good representation of the historic and current availability of LLMs in those languages. From the first efforts to train encoder-type models (e.g., Berta for Catalan (Armengol-Estapé et al., 2021), BERTeus for Basque (Agerri et al., 2020) or BERT-gl for Galician (Garcia, 2021), current efforts and initiatives have been focused on developing competing decoder models for these languages. Nonetheless, with the exception of the MarIA family, which includes a Spanish GPT-2, and the recent release of Portuguese GlórIA, trained using a GPTNeo architecture (Black et al., 2021), it is worth emphasizing that all the available language-specific models have not been trained from scratch.

Instead, in order to surpass the major data constraint imposed when adopting the common approach of pre-training from scratch with randomly initialized weights, alternative strategies such as continual pre-training are being explored. Continual pre-training of an existing LLM is an efficient technique that allows to leverage the knowledge encapsulated within the model (Gupta et al., 2023). Such is the case of FLOR models for Catalan based on BLOOM (Da Dalt et al., 2024), Latxa models for Basque based on Llama 2 (Etxaniz et al., 2024), or Carballo models for Galician based on continual pre-trained versions of BLOOM and Cerebras models already tailored to Spanish, English, and Catalan (Gamallo et al., 2024), among others.

By offering an open-access benchmark with the qualities of IberoBench, the relevant parties will now have a reliable tool that allows for reproducible results to evaluate *toy* models during testing and model checkpoints during training. This tool should also offer these parties an open, shared leader-board to display the capabilities of their models, hopefully resulting in an increased interest in training LLMs aimed at performing their best in the Iberian languages.

3 Benchmark creation

3.1 General rationale

The design criteria of IberoBench stemmed from the need for a unified, high-quality, and localized evaluation benchmark for Iberian languages, offering comprehensive coverage of key LLM skills and evaluation categories. The task categories we adopt represent a balance between commonly recognized areas of evaluation[3] and the opportunity to leverage existing high-quality datasets. These categories are: Commonsense Reasoning, Linguistic Acceptability, Mathematics, Natural Language Inference (NLI), Paraphrasing, Question Answering (QA), Reading Comprehension, Summarization, Translation, and Truthfulness. We strive for broad task coverage to bring evaluation in Iberian languages

[3] To define our categories, we reviewed automatic evaluations addressed in LLM releases, identifying commonly used datasets such as MMLU, PIQA, ARC, Hellaswag, and GSM8K. These datasets, while widely adopted, are associated with varying categories across studies.

to a comparable level to English, and enable consistent performance assessment across languages. Consequently, we also targeted task parallelism across the languages covered, avoiding tasks that are exclusive to a single language.

To collect data for these domains in each of our target languages, we found that only in some cases did pre-existing datasets meet our criteria, which are outlined in detail below. For areas lacking suitable evaluation data, our strategies included human translation and the creation of datasets from scratch. It is important to note that IberoBench is designed as a living benchmark, welcoming further contributions to address existing gaps (see Table 1) and improve the quality and diversity of tasks.

Collection of pre-existing datasets. The benchmark aims to integrate high-quality pre-existing datasets, provided they meet well-defined criteria. For annotated datasets, we only included those with human-generated annotations (e.g., Belebele, EusProficiency) or datasets with automatic annotations that have undergone thorough human revision (e.g., ASSIN). For translated datasets, we required human translations (e.g., MGSM, xStoryCloze, XQuAD_ca) or, at least, translations reviewed by humans to ensure quality and accuracy.

Translations into Spanish and Portuguese were prioritized to adhere to the European variants of these languages, in line with the benchmark's regional focus. Some of the selected datasets, such as Belebele, already had existing task implementations within EleutherAI's evaluation framework and were simply incorporated into the benchmark. In contrast, others, like XQuAD and FLORES, required the implementation to be developed.

Human translation. Our strategy of professionally translating datasets was applied only to those we deemed suitable for this approach. As a result, we excluded certain promising datasets that were either better suited for replicating the creation methodology using target language data (e.g., LAMBADA –Paperno et al., 2016) or too closely tied to a specific source culture, such as MMLU (Hendrycks et al., 2021a), which includes U.S.-specific elements like laws and historical topics. For datasets selected for human translation, such as PIQA and OpenBookQA, both dataset-specific guidelines and unified general criteria were developed to ensure consistent standards, enhance benchmark homogeneity, and enable comparability across tasks available in multiple Iberian languages. These main general criteria are:

- Dates, metric systems, and currency should be adapted to the target-language context (e.g., "5 feet 11 inches, 170 pounds" could be translated as "1,80 metres, 77 quilos" in Catalan).

- Personal names should be translated to their local equivalents if such equivalents exist; otherwise, a common name within the target context should be selected. The chosen equivalents must be applied consistently throughout the dataset.

- The translated text should prioritize a rich lexical variety and include idiomatic expressions where possible, aiming to reflect the real usage and richness of the target language.

- It is essential to ensure that the internal logic of the datasets is preserved. For example, questions and answers (or questions and continuations, or other formats) must remain coherent and meaningful in the target language.

- Any errors in the original dataset, whether grammatical or content-related, should be corrected in the translation if they affect meaning or readability. This is not the case if the error is part of the task.

- Care must be taken not to alter the difficulty of the task, ensuring no lexical or syntactic patterns make the task easier to resolve.

- Where feasible, the length of the answers should closely match that of the original text.

While these general principles applied across all languages, specific adjustments were necessary for Basque due to its unique linguistic characteristics. These arose primarily from its ergative-absolutive alignment, agglutinative morphology, and head-final word order, which contrast with English's nominative-accusative alignment, synthetic morphology, and head-initial word order. These differences posed challenges, for instance, when translating evaluation instances framed as sentence completions, where direct translations into Basque frequently result in ungrammatical or unclear constructs. To address this, translators were instructed to restructure sentences minimally while preserving semantic integrity. Examples of these adaptations are presented in Appendix A.

Creating datasets from scratch. This approach aimed to leverage existing resources in the target languages to develop new evaluation datasets

Category	ca	es	eu	gl	pt
Commonsense Reasoning	copa_ca **xstorycloze_ca**	**copa_es** xstorycloze_es	**xcopa_eu** xstorycloze_eu		
Linguistic Acceptability	catcola	escola		galcola	
Math	**mgsm_direct_ca**	mgsm_direct_es	**mgsm_direct_eu**	**mgsm_direct_gl**	
NLI	xnli_ca wnli_ca teca	xnli_es wnli_es	xnli_eu **wnli_eu** qnli_eu		assin_entailment
Paraphrasing	parafraseja **paws_ca**	paws_es		**parafrases_gl** **paws_gl**	assin_paraphrase
QA	**arc_ca** catalanqa coqcat **openbookqa_ca** **piqa_ca** **siqa_ca** xquad_ca	xquad_es **openbookqa_es**	**piqa_eu** eus_exams eus_proficiency eus_trivia	**openbookqa_gl**	
Reading Comprehension	belebele_cat_Latn	belebele_spa_Latn	belebele_eus_Latn eus_reading	**belebele_glg_Latn**	belebele_por_Latn
Summarization	cabreu	xlsum_es		**summarization_gl**	
Translation / Adaptation	flores_ca **phrases_va**	flores_es **phrases_es**	flores_eu	flores_gl	flores_pt
Truthfulness	veritasqa_ca	veritasqa_es		veritasqa_gl **truthfulqa_gl**	

Table 1: Tasks included in IberoBench. Newly introduced datasets are marked in bold.

tailored to specific linguistic categories. Three datasets were created using this strategy:

- **Phrases**. A translation dataset focusing on Valencian –a linguistic variety closely related to Catalan– designed for the language pairs Catalan-Valencian and Valencian-Spanish. This dataset was constructed from 200,000 phrases extracted from the Common Voice tool (Ardila et al., 2020). The phrases were filtered for lexical and grammatical diversity, and the selected sentences were then translated from Spanish into Catalan and Valencian. The translation process was conducted by an expert in Catalan philology, ensuring both linguistic quality and cultural accuracy in the resulting datasets.

- **Parafrases_gl**. A paraphrase identification dataset in Galician comprising 2032 entries. Each entry consists of a pair of sentences annotated with one of three labels: full paraphrases, borderline paraphrases, or non-paraphrases. The examples were sourced from diverse materials, such as Wikipedia articles, novels, and parliamentary sessions. Labeled sentences were automatically generated using various strategies, such as term replacement with a BERT model to create lexical paraphrases and back-translation using a pivot language to generate syntactic paraphrases. The dataset was manually reviewed and annotated by two linguists.

- **Summarization_gl**. A summarization dataset in Galician featuring over 80,000 summaries paired with their corresponding full articles. The dataset was constructed from the news articles of three Galician newspapers: *Que pasa na costa*, *Nós diario*, and *Praza Pública*.

3.2 Tasks presented

IberoBench entails 62 tasks across 10 broad categories. The total number of subtasks is 179. These tasks cover five Iberian languages. We currently do not include other varieties of these languages, but we include two tasks (one in European Spanish and another in Catalan) that entail language adaptation to and from Valencian. Also, some types of tasks are not available yet for all the languages in the benchmark (particularly for European Portuguese). We hope to cover these linguistic and task-related gaps in the near future and welcome collaborators to help us achieve this goal. Table 1 provides an overview of the tasks and datasets included in IberoBench. Each of the five languages has its own corresponding benchmark, named after the language and containing all the tasks for that particular language (BasqueBench, CatalanBench, GalicianBench, PortugueseBench and SpanishBench).

We present 22 new datasets and tasks in IberoBench, and make use of existing datasets to expand it with 40 tasks. The existing datasets we incorporate are two subsets from ASSIN (Fonseca et al., 2016), Belebele (Bandarkar et al., 2023),[4] caBREU, CatalanQA, COPA_ca, CoQCat, PAWS_ca, TE-ca, WNLI_ca and XNLI_ca (Gonzalez-Agirre et al., 2024), CatCoLA (Bel et al., 2024b), EsCoLA (Bel et al., 2024a), EusExams, EusReading, EusProficiency and EusTrivia (Etxaniz et al., 2024), FLORES-200 (Costa-jussà et al., 2022), GalCoLA (de Dios-Flores et al., 2023), MGSM (Shi et al., 2023), Parafraseja,[5] PAWS-X (Yang et al., 2019), QNLI_eu (Urbizu et al., 2022), VeritasQA (Aula-Blasco et al., 2025), XL-Sum (Hasan et al., 2021), XNLI (Conneau et al., 2018), XNLIeu (Heredia et al., 2024), XQuAD (Artetxe et al., 2020), XQuAD_ca (Armengol-Estapé et al., 2021), and XStoryCloze (Lin et al., 2022a).

4 Implementation

4.1 Overall Harness processes

IberoBench is offered through the LM Evaluation Harness (Gao et al., 2023), a framework that aims to provide a unified implementation for the automatic evaluation of LLMs. For each task added to the framework, the prompt, model generation parameters, data and output processing, metrics, and other variables, are specified in a configuration YAML, allowing anyone to replicate the results.

Tasks can be of two main types: multiple choice and open-ended generation. In multiple choice tasks, the model determines the best option from a set of choices. At the implementation level, the log-likelihoods of each option are compared to select the highest as the predicted label. A paradigmatic metric for such tasks is *accuracy*. In generation tasks, the model is given a specific text (e.g., a text to summarize or translate) and is requested to generate a text which is then compared with a reference text. BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) are examples of metrics used for this type of tasks.

4.2 Decisions in task implementation

We also focus on the quality of implementation. Some of these decisions had already been followed by the contributors to the official LM Evaluation Harness, such as task version control or the unification of task utilities (see Appendix B), but we also identify problems in the implementation of some tasks in IberoBench. These decisions align with the best practices from Biderman et al. (2024).

New types of tasks IberoBench includes summarization and machine translation tasks, two categories which are yet to be implemented in the official LM Evaluation Harness. This is the case of tasks such as CaBreu (Gonzalez-Agirre et al., 2024), XL-Sum (Hasan et al., 2021) and FLORES-200 (Costa-jussà et al., 2022). We also add new tasks to other categories already present, such as reading comprehension, paraphrasing, NLI, QA and truthfulness. For summarization tasks that require a specific type of summary (i.e., abstractive, extractive or extreme), we test if adding the type of summary required to the prompt made any difference to a generic prompt, but find that for small models (<2B parameters) there is no difference in performance. However, we keep the summary type in the prompt for future, more capable models.

Multilingual prompting To allow multilingual comparison, when a task was already implemented in English, the prompts for other languages are directly translated by humans into the target language. This happens in tasks that entailed datasets that have been translated from English in their creation phase, such as MGSM_eu or PAWS_gl. When the dataset for a task is not a result of a translation, we take prompts used in similar tasks. For instance, for EsCoLA, CatCoLA or GalCoLA, we translate the English CoLA prompts, as they share the task of measuring linguistic acceptability, even though they were created independently. When there were no parallel or similar tasks, we create comparable natural language prompts. For every task, the prompts are parallel between all IberoBench languages. Given our multilingual setup, we explored if translating the "Q:" and "A:" in the prompt (for instance, "Pregunta:" and "Resposta:" in Catalan, or alternatively "P:" and "R:") for QA tasks had an overall impact in performance, specifically with small models (<2B parameters). However, as we find no significant difference, and since we do not optimize prompts to increase the performance of any model following best practices suggested by Biderman et al. (2024), we keep the simplest and most natural translation (i.e., "P:" and "R:" in the example of Catalan). The only exception are the prompts for FLORES-200. This task is included in all five of our benchmarks and covers all combinations between a given language, other Iberian lan-

[4] In multilingual datasets, we only take the tasks that entail one of our languages.

[5] `https://huggingface.co/datasets/projecte-aina/Parafraseja`

guages, and four major European languages. Given all the potential combinations of languages within a single task, we made an exception by using only an English prompt across all FLORES-200 sub-tasks.

Metrics for new tasks Metrics used in new tasks are aligned with the metrics of parallel or related tasks already in the LM Evaluation Harness. In particular, summarization and machine translation tasks are evaluated with the BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) metrics. We opt for these metrics not only because they are preferred for other generation tasks in the official repository, but also because they are the best options given our setup and languages. For instance, metrics such as COMET (Rei et al., 2020), BERTScore (Zhang et al., 2020), METEOR (Banerjee and Lavie, 2005) and BLEURT (Sellam et al., 2020) can work better than BLEU and ROUGE for summarization (Deutsch and Roth, 2021; Wang et al., 2023; Zhang et al., 2024) and translation (Freitag et al., 2022), but these results are sometimes inconclusive and dependent on the subtask and evaluation criteria. More importantly, these metrics do not support all of our working languages (e.g., METEOR and BLEURT), or require loading another model to generate embeddings (e.g., COMET and BERTScore), which slows down the evaluation process, defeating one of the main aims of the LM Evaluation Harness. Thus, we use ROUGE for tasks in which the gold standard is not too variable (e.g., extractive summarization). ROUGE measures the recall of *n*-grams in the generated text against the reference text, and we want to ensure that all relevant instances from the gold standard are covered. For tasks in which the gold standard is just an instance of what a good answer could be (e.g., abstractive summarization), we use BLEU. This metric does the same as ROUGE but with a focus on precision instead of recall, meaning that it measures if all the generated instances are relevant.

4.3 Solving problems

The creation of IberoBench allows us to delve deep into the framework and the datasets that are currently covered. Working with human translators for some tasks implies that someone manually revised large reference English datasets that were automatically generated at first. Our focus on quality also means that we manually revise the inputs and outputs for each task for a variety of models. By doing this, we identify several issues with the way some instances of a dataset end that clash with the prompt format, with how some prompts and instances of a dataset interact to generate double or no spaces in the input for the model, and with the use of some non-UTF-8 characters in the prompt for the Belebele task, among others.

The original ARC dataset (Yadav et al., 2019), for example, contains data with grammatical errors, unfinished instances, or odd symbols and characters within the sentences. In this dataset we also observe mixed label types (using both numbers and letters). Thus, we ensure that these errors do not permeate to ARC_ca by manually revising and fixing each of them.

Another issue is that the existing LM Evaluation Harness implementation of XNLI (Conneau et al., 2018) uses a prompt that includes commas, while the original dataset includes instances that end in multiple ways such as commas, words, full stops, parentheses and question marks, among others. Due to the lack of pre-processing, models received inputs with, for example, two consecutive punctuation marks. Sclar et al. (2023) show that even the smallest differences in prompt formatting result in huge accuracy changes, so we ensure that pre-processing is comprehensive regarding punctuation for all our implementations of NLI tasks.

In the caBREU dataset (Gonzalez-Agirre et al., 2024), each text contains three summaries, meaning that one should be chosen as gold standard, discarding the other two, as LM Evaluation Harness does not support multiple reference texts. The original paper does not indicate any issues with any particular annotator, so we perform an internal test to assess which to keep. First, we tested small models (<2B parameters) in the task using each of the three annotators' summaries as gold standard. We observed an overall slight improved performance when using annotator #3 as gold standard. Thus, we undertook a small-scale human evaluation test, in which two native speakers (blind to the experimental setup) chose what they considered was the first summary among the three options. After testing 20 instances for each of the three types of summary covered by caBREU (60 evaluations overall), we observed no preference towards any annotator. For this reason, we use annotator #1 as gold standard, as this should also ensure that no model is benefited by using annotator #3 as gold standard.

5 Evaluation

5.1 Models selected

In order to put the benchmark into operation and check its effectiveness, we conducted an evaluation of 33 small and medium-size monolingual and multilingual LLMs. A full list referencing the models tested can be found in Appendix C. Although the list may appear heterogeneous at first, it responds to a clear objective. We wanted to provide an evaluation of 16 language-specific models recently published that have been tailored to the languages of the benchmark, even when these models are of different sizes or have been trained using different strategies. The group of language-specific models includes five models trained with Catalan data, six models trained with Spanish data, two models trained with Basque data, three models trained with Galician data, and five models trained with Portuguese data. Furthermore, in order to contrast the results obtained by language-specific models with existing multilingual and SOTA models or comparable sizes, we have included 17 models of different families (i.e., BLOOM, Falcon, Gemma, Llama 3, mGPT, Mistral, OLMo and XGLM). Testing the performance on the benchmark of these models, which have been mostly trained with English data, would also allow us to establish a comparison between their performance in English and their performance in the Iberian languages.

5.2 Normalization of metrics

Since each IberoBench task has a unique preferred metric, and each metric has specific minimum and maximum values, we normalize the results in order to perform a consistent comparison across different tasks. Following Srivastava et al. (2023), for a dataset i we adopt the Normalized Preferred Metric (NPM) as our primary evaluation measure:

$$NPM_i = 100 \times \frac{[\text{raw metric}]_i - [\text{random score}]_i}{[\text{max score}]_i - [\text{random score}]_i} \quad (1)$$

where $[\text{raw metric}]_i$ is the score obtained by the model on the i-th dataset, $[\text{random score}]_i$ is the score of performing random at the given task (e.g., 50% for a binary classification task) and $[\text{max score}]_i$ is the highest possible score on that dataset. Under this normalized preferred metric, tasks are calibrated so that a negative score or close to 0 corresponds to poor performance, and a score of 100 corresponds to perfect performance. On some tasks, model scores can be less than 0 if a model does worse than random.

5.3 Evaluation results

We evaluate the models using 0-shot and 5-shot prompting. In what follows, we report the results corresponding to 5-shot prompting. Normalized average scores per model and language are shown in Table 2. The results for the 0-shot prompting evaluation can be found in the Appendix D.1. Appendix D.2 presents 5-shot performance on parallel IberoBench tasks, including English. We provide figures on model performance per task category and language in Appendix D.3. Details on resources and reproducibility are found in Appendix E.

Considering Iberian-specific versus multilingual or English-centric SOTA models, we observe mixed results regarding absolute best performance. Models continually pre-trained in specific languages perform best in Spanish (Occiglot_eu5-7B) and Basque (Latxa-13B) for their respective languages. However, in all other cases, language-specific models are outperformed by English-centric or multilingual base models of comparable size. For Galician, the three language-specific models achieve similar or slightly worse results than the multilingual BLOOM-1.7B and mGPT-1.3B. For Catalan and Portuguese, the best language-specific models CataLlama-8B and Tower-13B, respectively, are outperformed by Mistral-7B, Llama 3-8B, and Gemma-7B.

We also observe mixed results regarding the impact of continual pre-training. The FLOR model family demonstrates small but consistent gains over its foundation BLOOM counterpart models for Catalan tasks. However, CataLlama-8B is surpassed by the model it was based on (i.e., Llama 3-8B), as is Carballo-B-1.3B in Galician tasks (based on FLOR-1.3B).

Basque seems to benefit the most from language-specific continual pre-training compared to open multilingual models like BLOOM, mGPT and XGLM. The margin for improvement in Basque is higher, considering the general tendency to include limited portions of Basque data in multilingual training. Knowledge transfer is also less likely to occur effectively for this language in a generic model due to its unique characteristics as an isolate language. In contrast, all Romance languages are likely to benefit from mixtures of training data.

Overall, Llama 3-8B and Gemma-7B seem to be good defaults for the studied languages, as they

Model	ca	es	eu	gl	pt
Language-specific					
Aitana-6.3B	**28.98**	22.41	2.28	2.20	21.47
CataLlama-8B	**34.35**	**24.06**	9.87	7.93	27.08
FLOR-760M	**19.46**	**15.16**	0.48	1.69	21.43
FLOR-1.3B	**21.34**	**17.87**	0.19	3.31	17.63
FLOR-6.3B	**29.58**	**23.33**	2.74	6.24	22.64
Occiglot_es/en-7B	35.48	**33.89**	6.10	14.41	38.45
Occiglot_eu5-7B	35.86	**35.27**	4.25	14.96	42.00
Latxa-7B	26.17	23.23	**18.11**	6.52	26.69
Latxa-13B	30.14	24.33	**24.96**	8.61	36.79
Carballo-B-1.3B	13.26	10.64	1.51	**2.70**	24.00
Carballo-C-1.3B	9.22	4.50	1.27	**1.90**	22.13
Carvalho-1.3B	9.74	6.21	0.69	**2.69**	**18.76**
GlorIA-1.3B	5.56	2.56	0.03	0.22	**16.88**
Sabiá-7B	25.67	22.52	1.00	5.50	**34.95**
Tower-7B	31.25	24.67	3.24	8.14	**37.27**
Tower-13B	34.99	31.81	2.88	11.93	**42.74**
Multilingual and SOTA					
BLOOM-1.1B	**17.52**	**14.21**	**3.21**	2.35	**20.83**
BLOOM-1.7B	**19.78**	**17.04**	**3.46**	3.89	**25.46**
BLOOM-3B	**23.30**	**18.98**	**6.11**	4.18	**23.05**
BLOOM-7.1B	**27.30**	**20.77**	**7.30**	5.92	**24.30**
Falcon-7B	17.92	19.42	1.02	5.31	30.45
Gemma-2B	22.11	21.74	4.83	8.37	36.50
Gemma-7B	36.07	31.26	16.57	18.00	49.00
Llama 3-8B	37.88	34.15	16.75	18.59	43.90
mGPT-1.3B	**7.99**	**8.09**	**3.83**	3.02	**23.39**
mGPT-13B	**10.97**	**11.66**	**1.87**	4.45	**18.73**
Mistral-7B	38.70	34.01	3.93	15.56	47.10
OLMo-7B	24.47	21.03	4.03	9.65	37.33
XGLM-564M	**5.35**	**3.25**	**2.64**	-0.78	**13.52**
XGLM-1.7B	**8.71**	**5.66**	**2.20**	0.11	**17.37**
XGLM-2.9B	**15.03**	**10.36**	**5.20**	0.49	**26.38**
XGLM-4.5B	**15.57**	**7.65**	**1.68**	3.02	**24.93**
XGLM-7.5B	**17.01**	**9.89**	**4.41**	0.65	**27.45**

Table 2: Normalized benchmark scores for models evaluated in a 5-shot setting for the five Iberian languages. Bold numbers show the languages each model officially supports. Double underlined numbers show the best-performing model per language.

provide the most balanced results across these languages. The strength and stability of these SOTA models are also evident from 0-shot results (Table 5 in Appendix D.1), where they obtain the best absolute performance for all languages except Basque.

6 Discussion and conclusions

IberoBench is a multi-task benchmark composed of 62 carefully curated tasks for the five official languages in Spain and Portugal, all created with a set of shared quality standards and made available via the LM Evaluation Harness framework. It represents a large step forward concerning the possibilities of assessing LLM performance in Iberian languages, while it is also a potentially valuable contribution to varieties of Spanish and Portuguese spoken outside the Iberian Peninsula (e.g., Brazilian Portuguese and the varieties of Spanish spoken in Latin America).

The evaluation of 33 LLMs clearly demonstrates that IberoBench offers a rich and challenging proving ground, showcased by the limited accuracy scores. This indicates that for Iberian languages, LLMs are still far away from being able to produce the desirable outputs on most tasks. Although the difference in the number of tasks available in each languages makes it difficult to compare results across languages, mathematical reasoning and linguistic acceptability appear to be the most challenging tasks in the benchmark.

IberoBench decreases the disparity of evaluation standards between Iberian languages and English, enabling comparable performance assessments. Importantly, the results obtained on the benchmark confirm the intuition that LLM performance is worse on Iberian languages than the results obtained for English (see Table 6 in Appendix D.2 for comparison across datasets). Altogether, this emphasizes the driving force behind the creation of such an evaluation suite: further efforts should be invested in the development of better multilingual LLMs, which is a fundamental enterprise to ensure that the millions of people who speak languages other than English have access to high-quality AI-driven tools and services. We hope this test bed contributes to fostering work on creating multilingual, general-purpose, machine-learning algorithms for language understanding that enable equitable access to information and technology.

Although a detailed analysis of the results obtained for each model is beyond the scope of this paper, IberoBench provides highly valuable insights for those engaged in the creation of LLMs with a focus on Iberian languages. We hope that the detailed evaluation presented in this work can help unveil the achievements and limitations of the strategies used to create language-specific models.

IberoBench is an ongoing effort and our goal is to keep expanding it to new tasks and datasets for the different target languages. In this respect, we invite and encourage contributions from researchers and practitioners who possess datasets that may be relevant for the benchmark. This participation will be invaluable in enhancing the comprehensiveness and robustness of our assessment and a better understanding of the capabilities of LLMs in our languages.

7 Limitations

Even though IberoBench tackles a diverse number of tasks with different degrees of complexity, some critical aspects are still missing from the benchmark. Importantly, not all languages are equally represented, and more efforts should be invested to fill these gaps for the task categories included at this point. Furthermore, soon we will also work on including tasks pertaining to bias and fairness evaluation, as well as the detection of harmful content.

One main limitation is the choice of metric for open-ended generative tasks such as summarization and machine translation. The most popular metric currently used in the LM Evaluation Harness is BLEU, which compares the generated text with the gold standard answer. For IberoBench, we use both BLEU and ROUGE depending on which one best fits each type of task (see §4.2). This means that, if the gold standard lacks details or the model generates a correct text that deviates much from the gold standard, scores will not reflect the actual performance of the model. Thus, results for these types of tasks should always be interpreted with caution, especially as research has shown that these metrics may not always align with human judgments (Wang et al., 2023; Zhang et al., 2024).

We acknowledge that gold-standard automatic evaluation has its limitations (see §2.1), and that some tasks and metrics in the LM Evaluation Harness also present some issues (see §4.3). However, we believe that IberoBench is a considerable step forward in the evaluation of multilingual LLMs encompassing languages that are spoken natively by over 510M speakers in the world (Eberhard et al., 2024).

Acknowledgments

This work is funded by the *Ministerio para la Transformación Digital y de la Función Pública* and *Plan de Recuperación, Transformación y Resiliencia* - Funded by EU – NextGenerationEU within the framework of the project ILENIA with references 2022/TL22/00215337, 2022/TL22/00215336, 2022/TL22/00215335 and 2022/TL22/00215334, and within the framework of the project *Desarrollo Modelos ALIA*.

This work has been promoted and financed by the *Generalitat de Catalunya* through the Aina project.

This work has been funded by the *Ministerio de Ciencia, Innovación y Universidades*, grants with references PID2021-128811OA-I00, PLEC2021-007662, and TED2021-130295B-C33, the latter also funded by the European Union Next Generation EU/PRTR

This work has been funded by the grant ED431G-2023/04 by the Galician Ministry of Education, University and Professional Training, and the European Regional Development Fund (ERDF/FEDER program).

This work has been funded by the VIVES: "Pla de Tecnologies de la Llengua per al valencià" project (2022/TL22/00215334) from the Projecte Estratègic per a la Recuperació i Transformació Econòmica (PERTE).

This work is part of Project JDC2022-049433-I, financed by the MCIN/AEI/10.13039/501100011033 and the European Union "NextGenerationEU"/PRTR".

References

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba O. Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, Chiamaka Chukwuneke, Happy Buzaaba, Blessing Sibanda, Godson Kalipe, Jonathan Mukiibi, Salomon Kabongo, Foutse Yuehgoh, Mmasibidi Setaka, Lolwethu Ndolela, Nkiruka Odu, Rooweither Mabuya, Shamsuddeen Hassan Muhammad, Salomey Osei, Sokhar Samb, Tadesse Kebede Guge, and Pontus Stenetorp. 2024. Irokobench: A new benchmark for african languages in the age of large language models. *Preprint*, arXiv:2406.03368.

Rodrigo Agerri, Iñaki San Vicente, Jon Ander Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa, and Eneko Agirre. 2020. Give your text representation models some love: the case for Basque. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4781–4788, Marseille, France. European Language Resources Association.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The Falcon series of open language models. *Preprint*, arXiv:2311.16867.

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. *Preprint*, arXiv:2402.17733.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben

Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

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.

Jordi Armengol-Estapé, Ona de Gibert Bonet, and Maite Melero. 2022. On the multilingual capabilities of very large-scale English language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3056–3068, Marseille, France. European Language Resources Association.

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 Docio, Silvia Paniagua Suárez, 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*, Abu Dhabi, United Arab Emirates. International Committee on Computational Linguistics.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2023. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. *Preprint*, arXiv:2308.16884.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Jeanine Banks. 2024. Gemma: Introducing new state-of-the-art open models.

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.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, Anthony DiPofi, Julen Etxaniz, Benjamin Fattori, Jessica Zosa Forde, Charles Foster, Jeffrey Hsu, Mimansa Jaiswal, Wilson Y. Lee, Haonan Li, Charles Lovering, Niklas Muennighoff, Ellie Pavlick, Jason Phang, Aviya Skowron, Samson Tan, Xiangru Tang, Kevin A. Wang, Genta Indra Winata, François Yvon, and Andy Zou. 2024. Lessons from the trenches on reproducible evaluation of language models. *Preprint*, arXiv:2405.14782.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. If you use this software, please cite it using these metadata.

Meriem Boubdir, Edward Kim, Beyza Ermis, Sara Hooker, and Marzieh Fadaee. 2023. Elo uncovered: Robustness and best practices in language model evaluation. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 339–352, Singapore. Association for Computational Linguistics.

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.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3).

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? Try ARC, the AI2 Reasoning Challenge. *Preprint*, arXiv:1803.05457.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and

Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

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.

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.

Iria de Dios-Flores, Carmen Magariños, Adina Ioana Vladu, John E. Ortega, José Ramom Pichel, Marcos García, Pablo Gamallo, Elisa Fernández Rei, Alberto Bugarín-Diz, Manuel González González, Senén Barro, and Xosé Luis Regueira. 2022. The nós project: Opening routes for the Galician language in the field of language technologies. In *Proceedings of the Workshop Towards Digital Language Equality within the 13th Language Resources and Evaluation Conference*, pages 52–61, Marseille, France. European Language Resources Association.

Daniel Deutsch and Dan Roth. 2021. Understanding the extent to which content quality metrics measure the information quality of summaries. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 300–309, Online. Association for Computational Linguistics.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2024. *Ethnologue: Languages of the World*. SIL International.

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.

Alena Fenogenova, Artem Chervyakov, Nikita Martynov, Anastasia Kozlova, Maria Tikhonova, Albina Akhmetgareeva, Anton Emelyanov, Denis Shevelev, Pavel Lebedev, Leonid Sinev, Ulyana Isaeva, Katerina Kolomeytseva, Daniil Moskovskiy, Elizaveta Goncharova, Nikita Savushkin, Polina Mikhailova, Denis Dimitrov, Alexander Panchenko, and Sergei Markov. 2024. MERA: A comprehensive LLM evaluation in Russian. *arXiv*, arXiv:2401.04531.

Erick R. Fonseca, Leandro B. dos Santos, Marcelo Criscuolo, and Sandra M. Aluisio. 2016. Assin: Avaliação de similaridade semântica e inferência textual. In *Computational Processing of the Portuguese Language-12th International Conference, Tomar, Portugal*, pages 13–15.

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

Pablo Gamallo, Pablo Rodríguez, Daniel Santos, Susana Sotelo, Nuno Miquelina, Silvia Paniagua, Daniela Schmidt, de Dios Flores, Iria, Paulo Quaresma, Daniel Bardanca, José Ramom Pichel, Vítor Nogueira, and Senén Barro. A Galician-Portuguese Generative Model. In *Progress in Artificial Intelligence*, pages 292–304. Springer Nature Switzerland.

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. *Procesamiento del Lenguaje Natural*, 73:259–270.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800GB dataset of diverse text for language modeling. *Preprint*, arXiv:2101.00027.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Marcos Garcia. 2021. Exploring the representation of word meanings in context: A case study on homonymy and synonymy. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3625–3640. Association for Computational Linguistics.

Maria Giagkou, Teresa Lynn, Jane Dunne, Stelios Piperidis, and Georg Rehm. 2023. *European Language Technology in 2022/2023*, pages 75–94. Springer International Publishing, Cham.

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.

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 394–398, Montréal, Canada. Association for Computational Linguistics.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. 2024. OLMo: Accelerating the science of language models. *Preprint*, arXiv:2402.00838.

Kshitij Gupta, Benjamin Thérien, Adam Ibrahim, Mats L. Richter, Quentin Anthony, Eugene Belilovsky, Irina Rish, and Timothée Lesort. 2023. Continual pre-training of large language models: How to (re)warm your model? *Preprint*, arXiv:2308.04014.

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. *Proceedings of the International Conference on Learning Representations (ICLR)*.

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. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Maite Heredia, Julen Etxaniz, Muitze Zulaika, Xabier Saralegi, Jeremy Barnes, and Aitor Soroa. 2024. XNLIeu: a dataset for cross-lingual NLI in Basque. *Preprint*, arXiv:2404.06996.

Tom Hosking, Phil Blunsom, and Max Bartolo. 2024. Human feedback is not gold standard. In *The Twelfth International Conference on Learning Representations*.

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.

Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. 2023. Evaluating open-domain question answering in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5591–5606, Toronto, Canada. Association for Computational Linguistics.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. *Preprint*, arXiv:2405.01535.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *NeurIPS*.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, Jörg Frohberg, Mario Šaško, Quentin Lhoest, Angelina McMillan-Major, Gerard Dupont, Stella Biderman, Anna Rogers, Loubna Ben allal, Francesco De Toni, Giada Pistilli, Olivier Nguyen, Somaieh Nikpoor, Maraim Masoud, Pierre Colombo, Javier de la Rosa, Paulo Villegas, Tristan Thrush, Shayne Longpre, Sebastian Nagel, Leon Weber, Manuel Muñoz, Jian

Zhu, Daniel Van Strien, Zaid Alyafeai, Khalid Almubarak, Minh Chien Vu, Itziar Gonzalez-Dios, Aitor Soroa, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Adelani, Long Phan, Hieu Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilic, Margaret Mitchell, Sasha Alexandra Luccioni, and Yacine Jernite. 2023. The BigScience ROOTS corpus: A 1.6TB composite multilingual dataset. *Preprint*, arXiv:2303.03915.

Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Benita Teufel, Marco Bellagente, Minguk Kang, Taesung Park, Jure Leskovec, Jun-Yan Zhu, Li Fei-Fei, Jiajun Wu, Stefano Ermon, and Percy Liang. 2023. Holistic evaluation of text-to-image models. *Preprint*, arXiv:2311.04287.

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. 2022a. 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.

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. 2022b. Few-shot learning with multilingual language models. *Preprint*, arXiv:2112.10668.

Chuang Liu, Linhao Yu, Jiaxuan Li, Renren Jin, Yufei Huang, Ling Shi, Junhui Zhang, Xinmeng Ji, Tingting Cui, Tao Liu, Jinwang Song, Hongying Zan, Sun Li, and Deyi Xiong. 2024. OpenEval: Benchmarking chinese LLMs across capability, alignment and safety. *Preprint*, arXiv:2403.12316.

Ricardo Lopes, Joao Magalhaes, and David Semedo. 2024. GlórIA: A generative and open large language model for Portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 441–453, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. LLM evaluators recognize and favor their own generations. *Preprint*, arXiv:2404.13076.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The lambada dataset: Word prediction requiring a broad discourse context. *Preprint*, arXiv:1606.06031.

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

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for Falcon LLM: Outperforming curated corpora with web data, and web data only. *Preprint*, arXiv:2306.01116.

Ramon Pires, Hugo Abonizio, Thales Sales Almeida, and Rodrigo Nogueira. 2023. *Sabiá: Portuguese Large Language Models*, page 226–240. Springer Nature Switzerland.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report.

Georg Rehm and Andy Way. 2023. *European Language Equality: A Strategic Agenda for Digital Language Equality*. Springer Cham.

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

Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2022. A survey of evaluation metrics used for NLG systems. *ACM Comput. Surv.*, 55(2).

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. 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 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Teven Le Scao et al. 2023. BLOOM: A 176B-parameter open-access multilingual language model. *Preprint*, arXiv:2211.05100.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. *Preprint*, arXiv:2310.11324.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

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*.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. 2023. mGPT: Few-shot learners go multilingual. *Preprint*, arXiv:2204.07580.

Harman Singh, Nitish Gupta, Shikhar Bharadwaj, Dinesh Tewari, and Partha Talukdar. 2024. IndicGenBench: A multilingual benchmark to evaluate generation capabilities of LLMs on Indic languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11047–11073, Bangkok, Thailand. Association for Computational Linguistics.

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

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, Marseille, France. European Language Resources Association.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Yiming Wang, Zhuosheng Zhang, and Rui Wang. 2023. Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8640–8665, Toronto, Canada. Association for Computational Linguistics.

Jason Wei, Maarten Paul Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew Mingbo Dai, and Quoc V. Le. 2022a. Finetuned language models are zero-shot learners. In *Proceedings of the Tenth International Conference on Learning Representations*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022b. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Minghao Wu and Alham Fikri Aji. 2023. Style over substance: Evaluation biases for large language models. *Preprint*, arXiv:2307.03025.

Vikas Yadav, Steven Bethard, and Mihai Surdeanu. 2019. Quick and (not so) dirty: Unsupervised selection of justification sentences for multi-hop question answering. 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 2578–2589, Hong Kong, China. Association for Computational Linguistics.

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.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2024. FLASK: Fine-grained language model evaluation based on alignment skill sets. In *The Twelfth International Conference on Learning Representations*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore:

Evaluating text generation with BERT. *Preprint*, arXiv:1904.09675.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking Large Language Models for News Summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. *Preprint*, arXiv:2402.07827.

A Basque-specific translation adjustments: examples

Some examples of the adjustments needed for the dataset translations into Basque are presented in Table 3 and Table 4.

Original (EN)	Question: "Volume and mass are properties of" A: "friction." B: "light." C: "matter."
Translated (EU)	Question: "Bolumenta eta masa honen propietateak dira:" A: "marruskadura." B: "argia." C: "materia." *Question: "Volume and mass are properties of the following:"* *A: "friction."* *B: "light."* *C: "matter."*
Explanation	The sentence “Volume and mass are properties of **friction**” (with one possible answer in bold) is translated literally as “Bolumena eta masa **marruskadura**ren propietateak dira". Due to word order differences, the literal Basque translation requires the answer to be positioned mid-sentence instead of at the end. This structure is unsuitable for the sentence completion format used in English.

Table 3: Basque-specific translation adjustments: example 1

Original (EN)	Question: "The compass" A: "knows orientation." B: "tracks people." C: "was invented in 1905."
Translated (EU)	Question: "Iparrorratza" A: "orientazioa ezagutzeko erabiltzen da." B: "jendeari jarraitzeko erabiltzen da." C: "1905ean asmatu zen." *Question: "The compass"* *A: "is used to know orientation."* *B: "is used to track people."* *C: "was invented in 1905."*
Explanation	“The compass tracks people” requires an ergative subject in Basque: (“**Iparroratzak** jendea jarraitzen du”, while “The compass was invented in 1905” requires an absolutive subject (“**Iparrorratza** 1905ean asmatu zen”). However, the question must be the same for all the possible answers.

Table 4: Basque-specific translation adjustments: example 2

B Additional details on task implementation

B.1 Unify utilities

Some existing tasks in the LM Evaluation Harness use pre-processing functions to adapt the data and/or the prompt, the target or the choice of the model. For each language Bench in IberoBench, we combine the functions for each task into a single set for simplicity and ease of use.

B.2 Versioning

Each task has its corresponding version (1.0 at the moment of release), and will be incremented each time a task is modified in a way that affects its scoring. In this way, if task implementations must be updated, the relevant parties can still reference the version of the task used, to ensure future research can always reproduce the reported results.

C Models evaluated

Language-specific models:

- **FLOR** models (Da Dalt et al., 2024). For Catalan and Spanish, FLOR 760M and 1.3B are continually pre-trained models adapted from BLOOM-1.1B and BLOOM-1.7B respectively. They have been trained on 26B tokens of Catalan (42.1%), Spanish (41.3%), and English (16.6%) data, after vocabulary adaptation of the source model to a new tokenizer in the target languages. FLOR 6.3B was trained using the same vocabulary adaptation technique by continually pre-training BLOOM-7.1B on 140B tokens on Catalan (CATalog 1.0 dataset),[6] Spanish, and English data with equal proportion of each language.

- **Aitana-6.3B**[7] is a continually pre-trained model based on FLOR-6.3B and trained for 2 epochs on 1.304 million tokens per epoch on Valencian data.

- **CataLlama-8B v0.1**[8] is a continually pre-trained model based on Llama-3 from Meta and trained for one epoch on 331M tokens of Catalan data, sourced from the Catalan General Crawling dataset[9], as part of the Catalan Textual Corpus (Armengol-Estapé et al., 2022).

- **Occiglot_eu5-7B**[10] and **Occiglot_es/en-7B**[11] are continually pre-trained models from Mistral-7B-v0.1, the first trained on 293B tokens on code and the top 5 European languages, i.e. English, Spanish, French, German, and Italian, and the second trained on 112B tokens of a subset of the previous that included the code, Spanish and English data.

- **Latxa v1.1** (7B and 13B; Etxaniz et al., 2024) are continually pre-trained models based on Meta's LLaMA 2 7B and 13B, respectively, trained on a mix of Basque corpora of 4.3M documents, and an additional set of 500K documents of English data taken from The Pile (Gao et al., 2020).

- **Carballo-B-1.3B** and **Carballo-C-1.3B** (Gamallo et al., 2024) are continually pre-trained from FLOR-1.3B and a Cerebras-GPT-1.3B model adapted to Catalan, Spanish and English, respectively. They both adapt the source model's tokenizer and embeddings to their target language, Galician, before continually pre-training the model on the 2.1B-word Galician corpus called CorpusNÓS (de Dios-Flores et al., 2022).

- **Carvalho-1.3B** (Gamallo et al.) is a continual pre-training of Cerebras-GPT-1.3B model adapted to Catalan, Spanish and English. It adapts the source model's tokenizer and embeddings to its two target languages, Galician and Portuguese, before continually pre-training the model on two corpora: the Galician CorpusNÓS and a 3B-word European Portuguese corpus.

- **GlorIA-1.3B** (Lopes et al., 2024) is a decoder model with a GPTNeo architecture trained from scratch on 35B tokens of a European Portuguese corpus.

- **Sabiá-7B** (Pires et al., 2023) is a continually pre-trained model based on LLaMA-1-7B and trained on 7.3B tokens in Portuguese.

- **Tower v0.1** (7B and 14B; Alves et al., 2024) are continually pre-trained models based on Llama 2 and trained on a mix of 20 billion tokens in English, Portuguese, Spanish, French, German, Dutch, Italian, Korean, Chinese, Russian, and parallel data.

Multilingual and SOTA models:

[6] https://hf.co/datasets/projecte-aina/CATalog
[7] https://hf.co/gplsi/Aitana-6.3B
[8] https://hf.co/catallama/CataLlama-v0.1-Base
[9] https://hf.co/datasets/projecte-aina/catalan_general_crawling
[10] https://hf.co/occiglot/occiglot-7b-eu5
[11] https://hf.co/occiglot/occiglot-7b-es-en

- **BLOOM** (1.1B, 1.7B, 3B, 7.1B; Scao et al., 2023) models are the result of a collaborative effort to train open-access multilingual LLMs of up to 176 billion parameters. They have been trained on the ROOTS corpus (Laurençon et al., 2023), which contains 1.5TB of data in 46 natural languages and 13 programming languages.

- **Falcon-7B** (Almazrouei et al., 2023) has been trained on 1,500B tokens from the RefinedWeb (Penedo et al., 2023) English web dataset.

- **Gemma** (2B and 7B; Banks, 2024) is a family of open models based on Google's Gemini models. They have been trained on 6 trillion tokens whose main components are reported to be web documents, code, and mathematics. They are intended for use in English.

- **Llama 3-8B**[12] is the latest version to date of the Llama family of models. It uses groped query attention and has been pre-trained on 15 trillion tokens from publicly available sources. It is intended for use in English.

- **mGPT** (1.3B and 13B; Shliazhko et al., 2023) are models with GPT-3 architecture trained on 60 languages from 25 language families obtained from Wikipedia and the Colossal Clean Crawled Corpus[13].

- **Mistral-7B v0.1** (Jiang et al., 2023) is a model using two key architectural choices: grouped query attention and sliding window attention. The sources and size of the training data are unknown. It is intended for use on English.

- **OLMo-7B** (Groeneveld et al., 2024) is an open LLM trained on a subset of 1.715 trillion tokens from the Dolma dataset, consisting of 3 trillion tokens from English sources.

- **XGLM** (544M, 1.7B, 2.9B, 4.5B and 7.5B; Lin et al., 2022b) is a family of multilingual models trained on a 500 billion token corpus of 30 languages from 16 different language families in an approximately balanced distribution.

[12] https://hf.co/meta-llama/Meta-Llama-3-8B

[13] https://hf.co/datasets/oscar-corpus/colossal-oscar-1.0

D Detailed results

D.1 Results per language.

Table 5 shows the average 0-shot results per language in IberoBench. The results are analogous to those reported in Table 2 for the 5-shot setting.

Model	ca	es	eu	gl	pt
Language-specific					
Aitana-6.3B	**23.82**	14.62	2.37	1.05	7.41
CataLlama-8B	**24.70**	**13.20**	4.96	2.12	9.93
FLOR-760M	**14.83**	**10.07**	2.04	-0.73	17.56
FLOR-1.3B	**15.02**	**8.08**	0.70	1.76	10.66
FLOR-6.3B	**22.53**	**16.94**	2.98	5.81	11.21
Occiglot_es/en-7B	26.48	**20.16**	3.16	6.73	20.47
Occiglot_eu5-7B	26.17	**20.74**	3.43	7.60	18.11
Latxa-7B	16.62	13.70	**12.46**	2.10	13.40
Latxa-13B	19.14	15.79	**16.45**	1.84	17.35
Carballo-B-1.3B	11.02	4.50	2.36	**1.07**	7.47
Carballo-C-1.3B	7.55	3.73	-0.12	**0.79**	9.39
Carvalho-1.3B	6.04	3.40	-0.15	**1.52**	**2.59**
GlorIA-1.3B	4.82	0.78	1.37	-2.15	**2.61**
Sabiá-7B	20.03	17.31	2.17	0.71	**16.47**
Tower-7B	23.03	18.55	1.19	3.22	**16.84**
Tower-13B	25.81	20.97	3.07	4.14	**20.03**
Multilingual and SOTA					
BLOOM-1.1B	**10.83**	**5.84**	**4.55**	-0.79	**9.51**
BLOOM-1.7B	**13.50**	**6.79**	**3.77**	1.51	**2.85**
BLOOM-3B	**17.87**	**11.07**	**5.81**	0.87	**7.09**
BLOOM-7.1B	**18.55**	**12.34**	**5.62**	1.70	**6.08**
Falcon-7B	13.18	13.07	1.34	2.14	11.84
Gemma-2B	20.09	15.28	3.78	4.48	17.91
Gemma-7B	28.48	24.11	10.80	10.00	24.54
Llama 3-8B	27.18	21.79	9.33	9.83	25.60
mGPT-1.3B	**7.64**	**5.53**	**6.66**	0.28	**7.81**
mGPT-13B	**10.71**	**7.38**	**4.02**	3.67	**6.48**
Mistral-7B	25.29	20.34	2.49	7.55	17.76
OLMo-7B	20.05	15.93	3.15	4.01	14.79
XGLM-564M	**6.83**	**3.50**	**2.31**	-2.68	**3.60**
XGLM-1.7B	**11.52**	**5.73**	**4.18**	-1.24	**3.86**
XGLM-2.9B	**16.28**	**11.32**	**5.58**	-2.03	**10.30**
XGLM-4.5B	**15.39**	**9.63**	**2.93**	2.70	**8.76**
XGLM-7.5B	**18.35**	**10.64**	**5.17**	-0.22	**13.25**

Table 5: Normalized benchmark scores across task categories for the models evaluated for the five Iberian languages in 0-shot prompting. Bold numbers show the languages each model officially supports. Double underlined numbers show the best-performing model per language.

D.2 Parallel tasks

Table 6 shows the 5-shot results of IberoBench tasks that have parallel tasks in English. We report the best language-specific model for each task in that language, the best non-specific (multilingual or SOTA) model for that task, and the results on three SOTA models: Gemma-7B, Llama 3-8B, and Mistral-7B.

Task	Lang	Best language-specific		Gemma-7B	Llama 3-8B	Mistral-7B	Best non-specific	
arc_easy	en		-	78.56	**78.91**	78.68		(Llama 3-8B)
	ca	(CataLlama-8B)	53.53	**60.60**	56.40	58.87		(Gemma-7B)
arc_challenge	en		-	40.84	38.45	**41.97**		(Mistral-7B)
	ca	(CataLlama-8B)	21.16	**26.51**	21.84	23.67		(Gemma-7B)
copa	en		-	**88.00**	86.00	86.00		(Gemma-7B)
	es	(Occiglot_eu5-7B)	**70.4**	58.4	64.8	60.4		(Llama 3-8B)
	ca	(CataLlama-8B)	60.00	45.20	**58.00**	49.60		(Llama 3-8B)
	eu	(Latxa-13B)	**52.00**	16.80	18.80	1.20		(Llama 3-8B)
xstorycloze	en		-	67.70	62.28	66.78	**70.22**	(OLMo-7B)
	es	(Occiglot_eu5-7B)	**51.82**	47.98	46.52	42.28		(Gemma-7B)
	ca	(CataLlama-8B)	**45.86**	43.34	43.88	42.42		(Llama 3-8B)
	eu	(Latxa-13B)	**38.46**	19.52	13.56	0.46		(Gemma-7B)
xnli	en		-	27.90	24.22	**28.26**		(Mistral-7B)
	es	(Occiglot_eu5-7B)	**23.38**	21.03	21.33	22.23		(Mistral-7B)
	ca	(FLOR-6.3B)	25.36	21.99	25.42	25.42	**25.78**	(BLOOM-3B)
	eu	(Latxa-7B)	**21.24**	12.34	9.85	2.97	14.76	(BLOOM-7.1B)
openbookqa	en		-	14.93	15.47	**15.73**		(Mistral-7B)
	es	(Occiglot_es/en-7B)	**20.00**	13.60	17.87	16.27		(Llama 3-8B)
	ca	(CataLlama-8B)	13.60	11.73	13.07	**13.87**		(Mistral-7B)
	gl	(Carballo-B-1.3B)	2.13	6.67	**10.40**	8.00		(Llama 3-8B)
piqa	en		-	61.70	60.94	**62.78**		(Mistral-7B)
	ca	(CataLlama-8B)	**35.80**	34.06	30.68	31.22		(Gemma-7B)
	eu	(Latxa-13B)	**33.76**	11.66	11.44	7.74	16.12	(XGLM-7.5B)
social_iqa	en		-	-1.71	-1.71	**1.74**		(Mistral-7B)
	ca	(CataLlama-8B)	**22.54**	19.09	19.86	19.63		(Llama 3-8B)
wnli	en		-	**40.84**	18.30	32.40		(Gemma-7B)
	es	(Occiglot_eu5-7B)	23.94	-12.68	7.04	**26.76**		(Mistral-7B)
	ca	(Aitana-6.3B)	15.50	**35.22**	9.86	26.76		(Gemma-7B)
	eu	(Latxa-13B)	21.12	1.40	9.86	-4.22	**26.76**	(mGPT-1.3B)
truthfulqa	en		-	**20.37**	19.48	19.67		(Gemma-7B)
	gl	(Carvalho-1.3B)	6.30	10.62	**12.11**	8.32		(Llama 3-8B)
paws	en		-	31.40	31.10	40.10		(Mistral-7B)
	es	(Occiglot_eu5-7B)	**38.70**	26.70	29.00	35.20		(Mistral-7B)
	ca	(CataLlama-8B)	32.40	25.40	37.50	**43.30**		(Mistral-7B)
	gl	(Carballo-B-1.3B)	7.90	**36.00**	32.30	35.60		(Gemma-7B)
xquad	en		-	81.81	**83.40**	81.99		(Llama 3-8B)
	es	(Occiglot_eu5-7B)	75.56	75.17	76.88	76.71	**77.48**	(OLMo-7B)
	ca	(FLOR-6.3B)	59.74	75.36	75.76	**76.24**		(Mistral-7B)
mgsm_direct	en		-	**15.00**	13.00	13.00		(Gemma-7B)
	es	(Tower-13B)	9.00	11.00	**12.00**	10.00		(Llama 3-8B)
	ca	(Aitana-6.3B)	3.00	**38.00**	9.00	8.00		(Gemma-7B)
	eu	(Latxa-13B)	4.00	**10.00**	9.00	4.00		(Gemma-7B)
	gl		0.00	**12.00**	9.00	6.00		(Gemma-7B)
belebele	en		-	81.48	**81.92**	79.41		(Llama 3-8B)
	es	(Occiglot_es/en-7B)	50.96	72.00	**75.41**	66.81		(Llama 3-8B)
	ca	(CataLlama-8B)	69.92	71.11	**72.29**	65.33		(Llama 3-8B)
	eu	(Latxa-13B)	35.41	**53.19**	44.75	16.29		(Gemma-7B)
	gl	(Carballo-B-1.3B)	0.19	68.85	**72.56**	54.91		(Llama 3-8B)
	pt	(Tower-13B)	45.63	72.75	**76.00**	68.44		(Llama 3-8B)

Table 6: Five-shot performance on IberoBench tasks parallel in English and at least one language in the benchmark.

D.3 Results per category and language

Figures 1, 2, 3, 4, and 5 report the 5-shot average performances per category in the CatalanBench, SpanishBench, BasqueBench, GalicianBench, and PortugueseBench, respectively. Figures 6, 7, 8, 9, and 10 show the values for 0-shot.

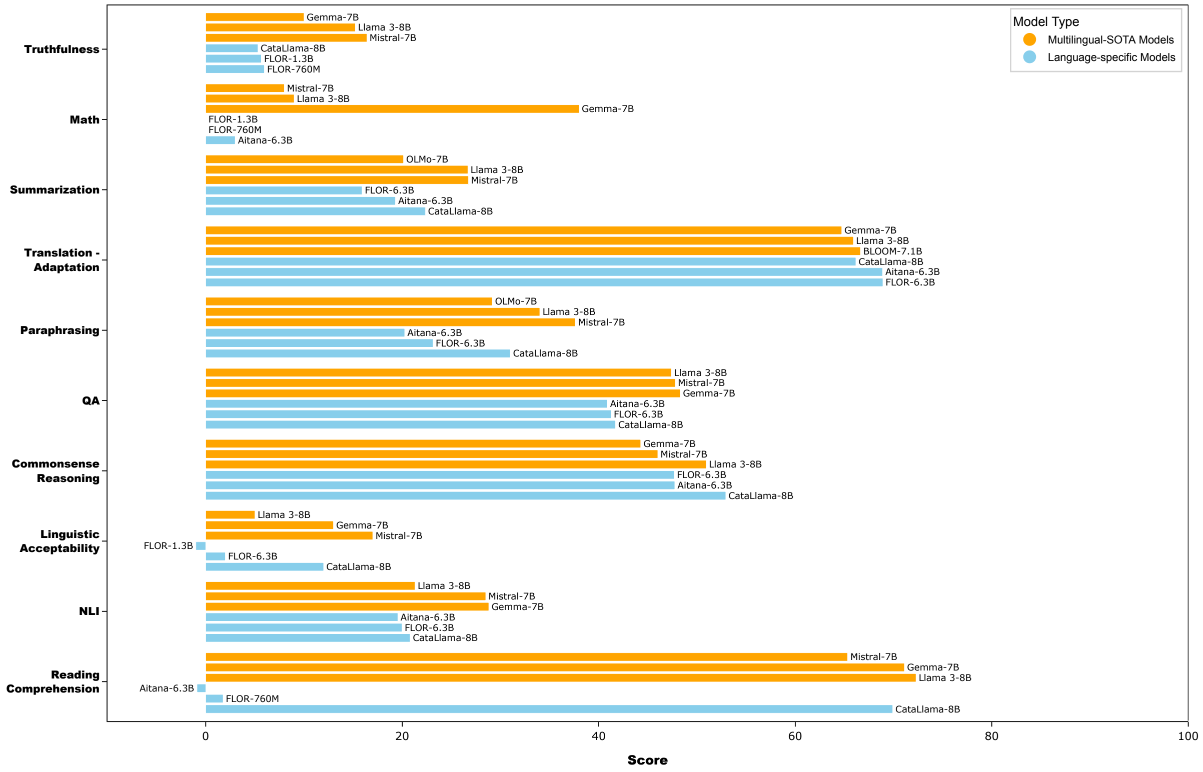

Figure 1: Five-shot performance using NPM scores on CatalanBench tasks. For each category, we report the three best performing monolingual Catalan-specific models (in blue) and the three best performing multilingual-SOTA models (in orange).

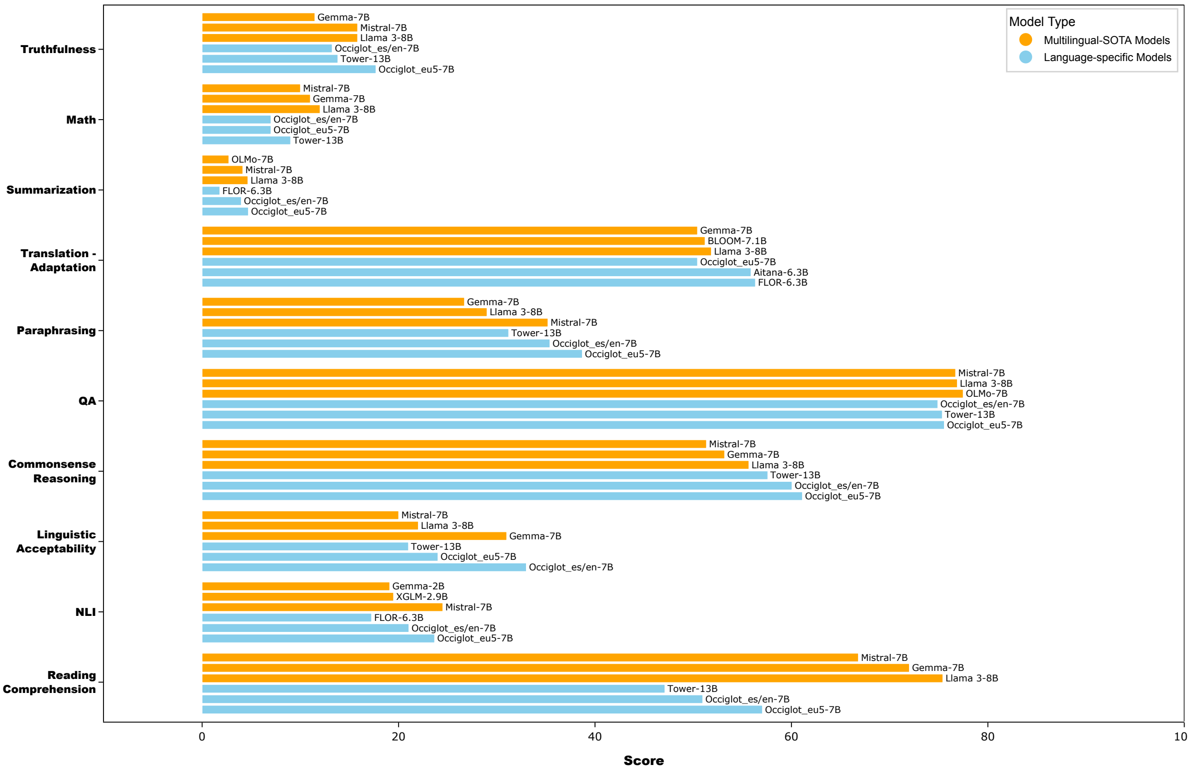

Figure 2: Five-shot performance using NPM scores on SpanishBench tasks. For each category, we report the three best performing monolingual Spanish-specific models (in blue) and the three best performing multilingual-SOTA models (in orange).

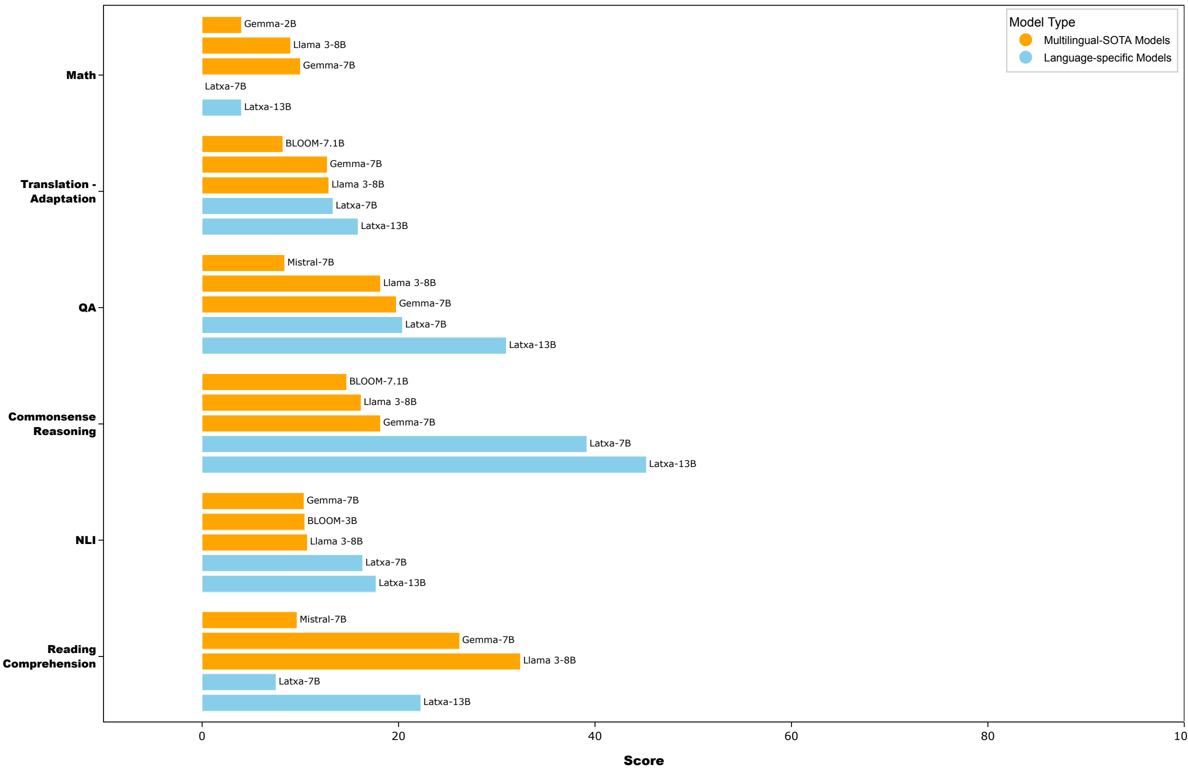

Figure 3: Five-shot performance using NPM scores on BasqueBench tasks. For each category, we report the three best performing monolingual Basque-specific models (in blue) and the three best performing multilingual-SOTA models (in orange).

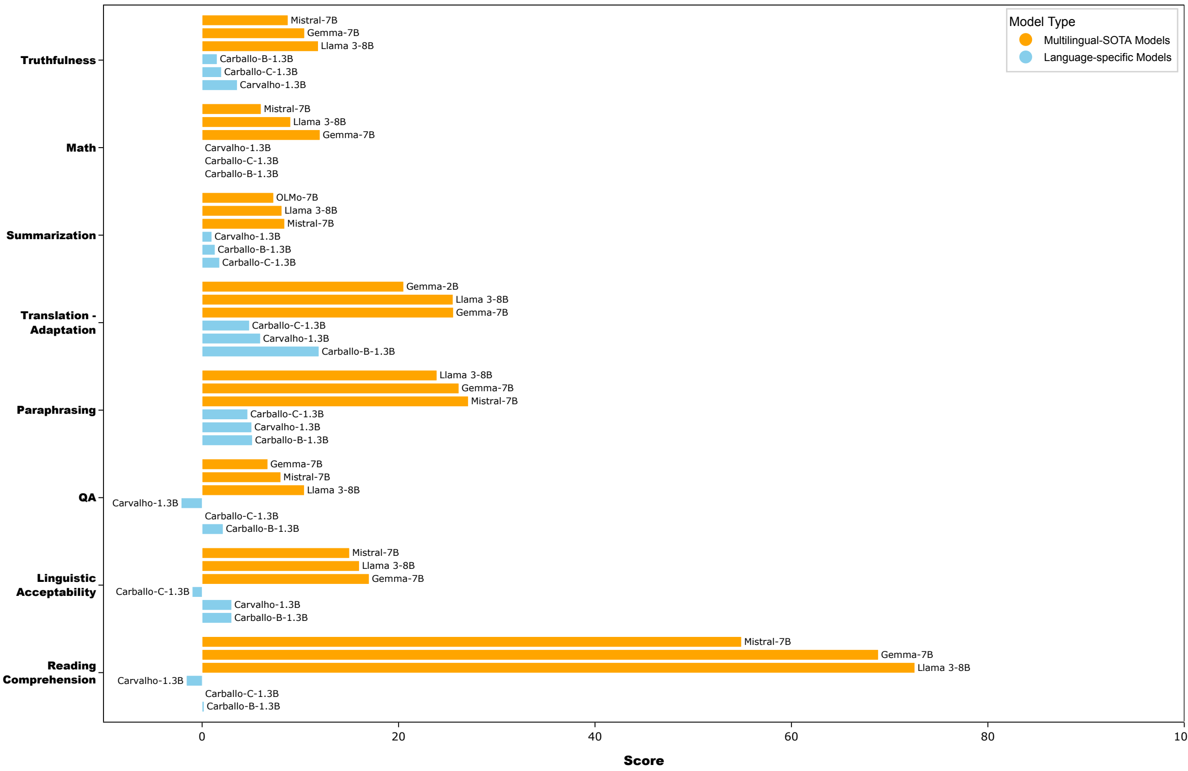

Figure 4: Five-shot performance using NPM scores on GalicianBench tasks. For each category, we report the three best performing monolingual Galician-specific models (in blue) and the three best performing multilingual-SOTA models (in orange).

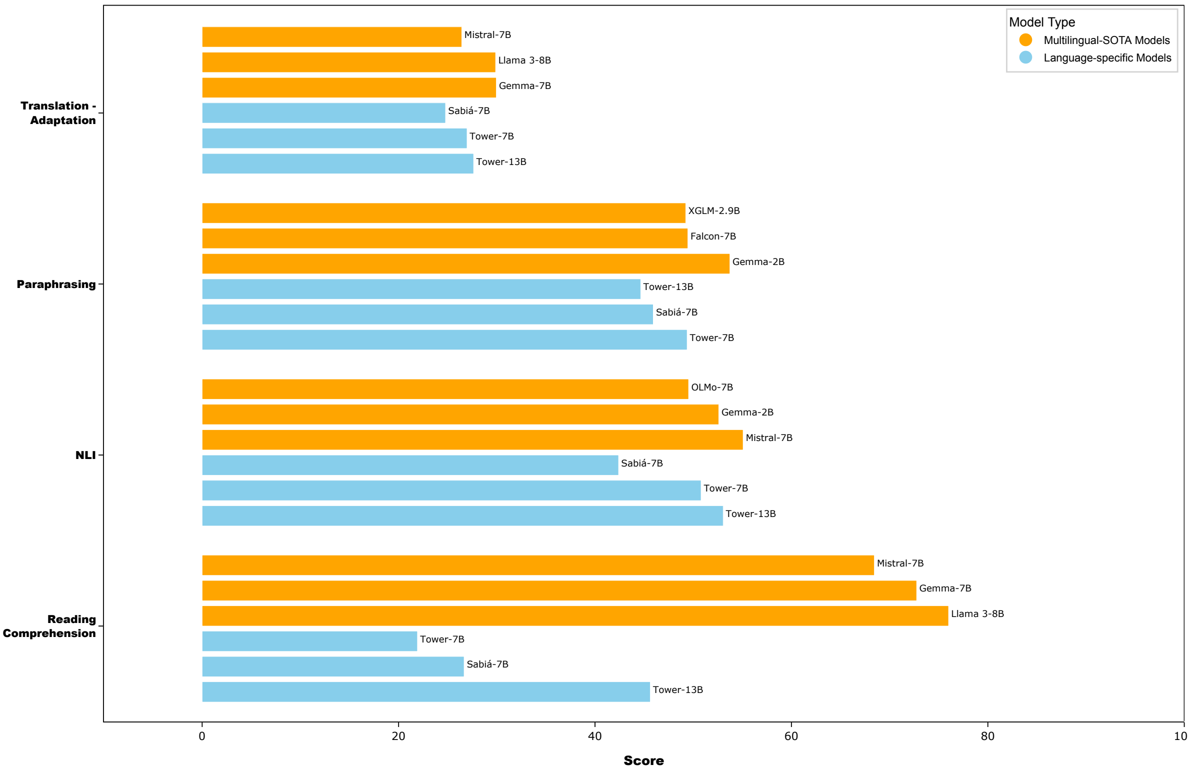

Figure 5: Five-shot performance using NPM scores on PortugueseBench tasks. For each category, we report the three best performing monolingual Portuguese-specific models (in blue) and the three best performing multilingual-SOTA models (in orange).

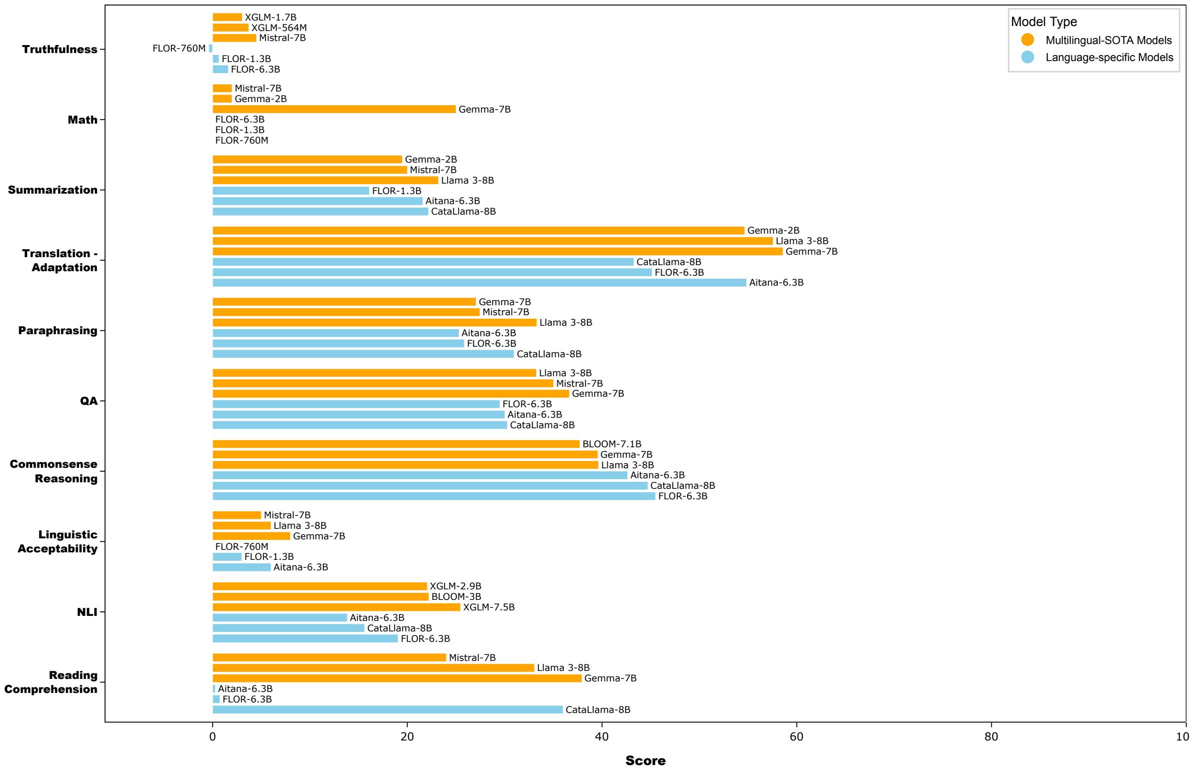

Figure 6: Zero-shot performance using NPM scores on CatalanBench tasks. For each category, we report the three best performing monolingual Catalan-specific models (in blue) and the three best performing multilingual-SOTA models (in orange).

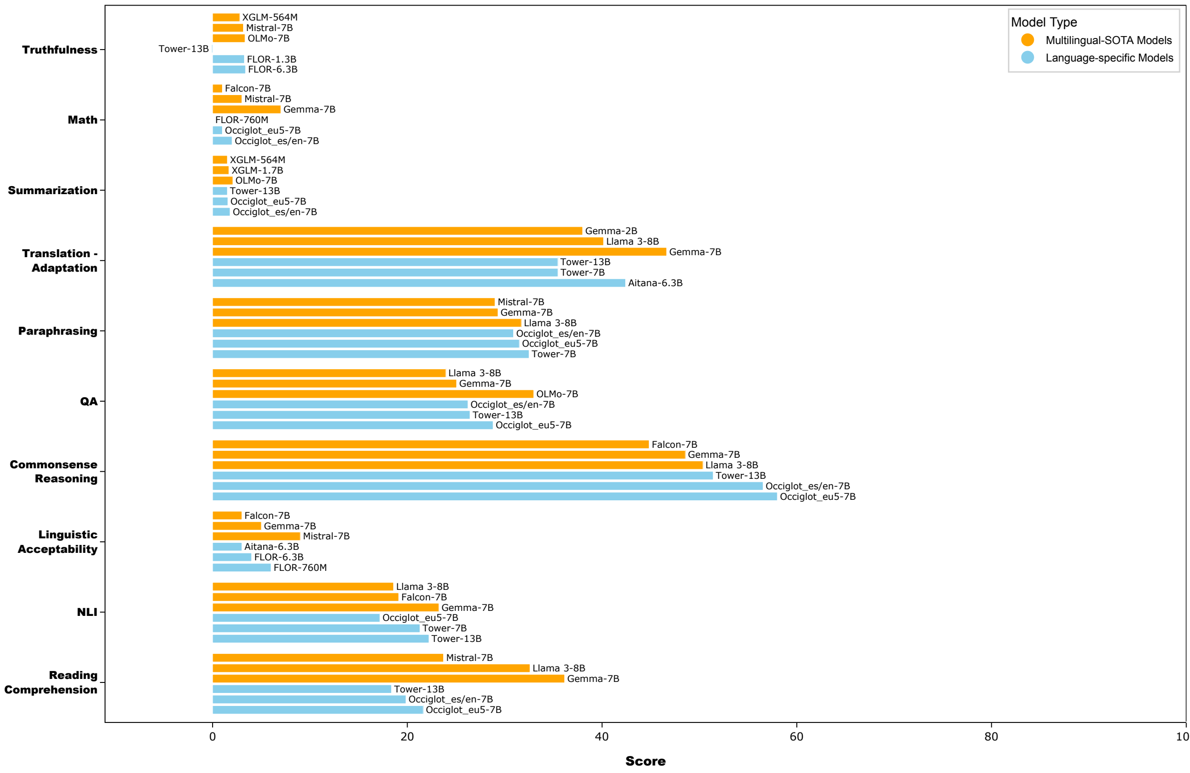

Figure 7: Zero-shot performance using NPM scores on SpanishBench tasks. For each category, we report the three best performing monolingual Spanish-specific models (in blue) and the three best performing multilingual-SOTA models (in orange).

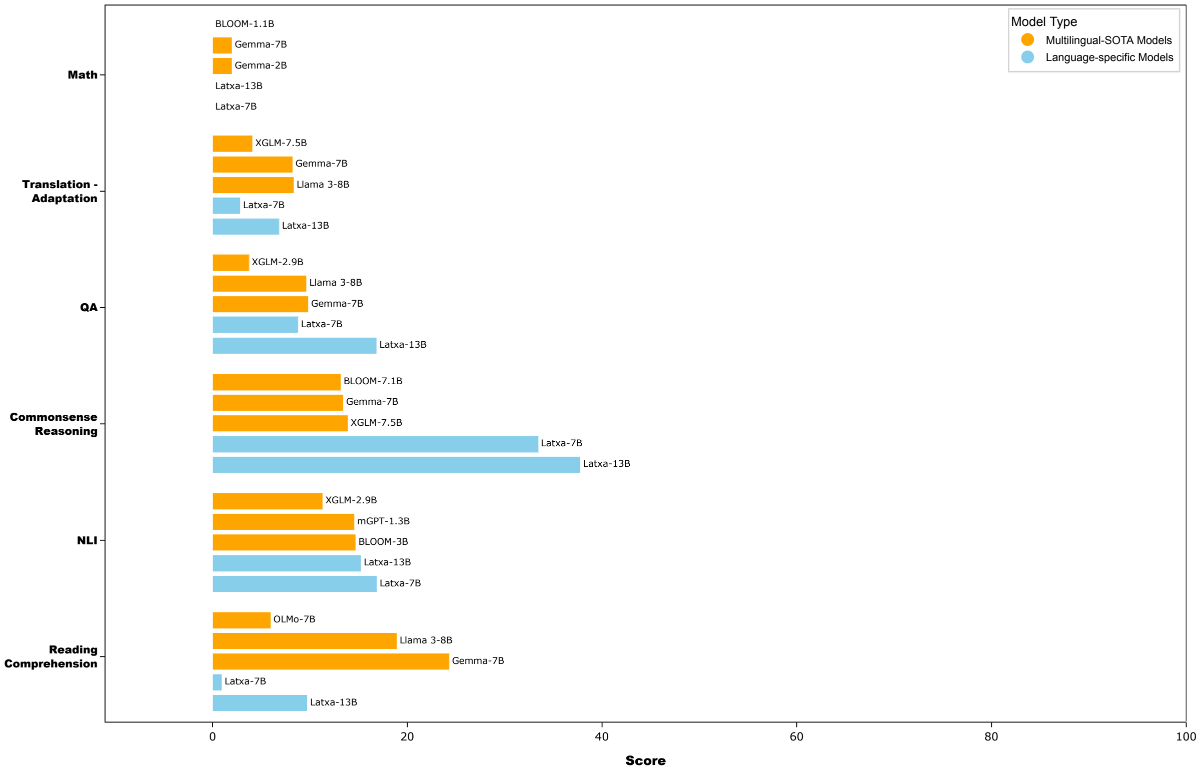

Figure 8: Zero-shot performance using NPM scores on BasqueBench tasks. For each category, we report the three best performing monolingual Basque-specific models (in blue) and the three best performing multilingual-SOTA models (in orange).

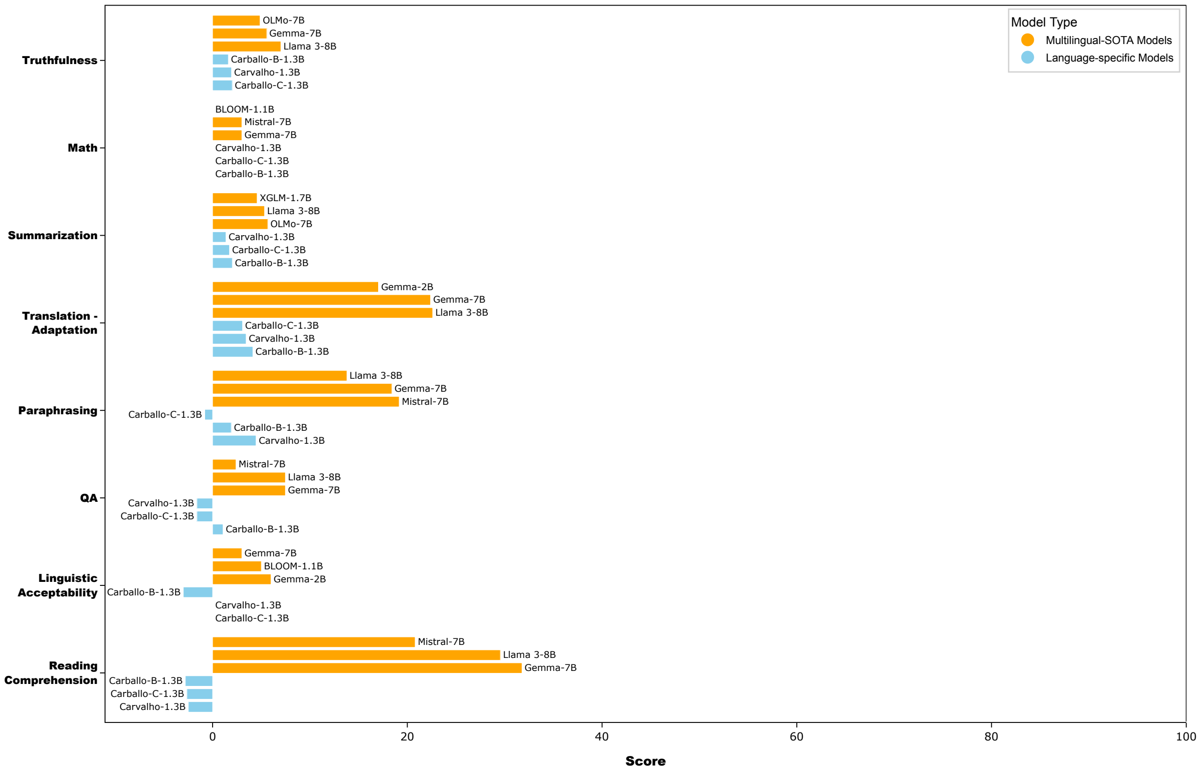

Figure 9: Zero-shot performance using NPM scores on GalicianBench tasks. For each category, we report the three best performing monolingual Galician-specific models (in blue) and the three best performing multilingual-SOTA models (in orange).

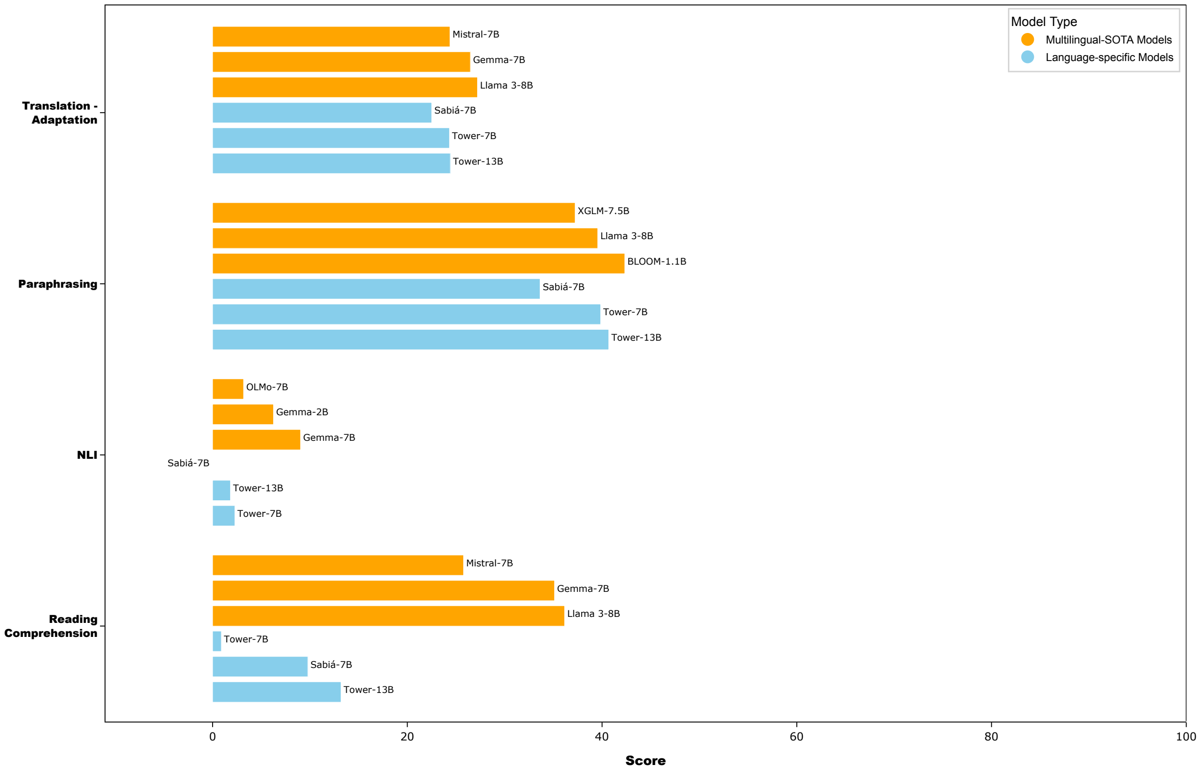

Figure 10: Zero-shot performance using NPM scores on PortugueseBench tasks. For each category, we report the three best performing monolingual Portuguese-specific models (in blue) and the three best performing multilingual-SOTA models (in orange).

E Resources and reproducibility

For running the evaluations, we used a single compute node with 4 64GB NVIDIA Hopper H100 GPUs. In all cases, we run the evaluation using data parallelism to reduce computation time, as the models fit on a single GPU. The environment for running the EleutherAI framework utilized Torch version 2.1.0a0+32f93b1 and Transformers version 4.40.2. To ensure reproducibility of results, we consistently set the random seed to 1234 across all evaluations. For the Occiglot_en/en-7B and Occiglot_eu5-7B models, we loaded them in `float16` format to reduce memory usage, as they were initially in `float32` format, which required twice the storage space. Additionally, for these models and Mistral-7B, we encountered memory issues in the execution of certain generative tasks that involve longer prompts or require the generation of relatively long texts. Specifically, for the caBREU, xlsum_es and eus_reading summary tasks, the maximum length of the model was limited to 5000 tokens to avoid these problems.

To compute the average of results per language, we took into account the normalized results of the tasks listed in table 7: 27 tasks in Catalan, 17 in Spanish, 14 in Basque, 14 in Galician, and 4 in Portuguese.

Catalan	arc_ca_challenge, arc_ca_easy, belebele_cat_Latn, cabreu_abstractive, cabreu_extractive, cabreu_extreme, catalanqa, catcola, copa_ca, coqcat, flores_ca, mgsm_direct_ca, openbookqa_ca, parafraseja, paws_ca, phrases_ca-va, phrases_va-ca, piqa_ca, siqa_ca, teca, veritasqa_gen_ca, veritasqa_mc1_ca, veritasqa_mc2_ca, wnli_ca, xnli_ca, xquad_ca, xstorycloze_ca
Spanish	belebele_spa_Latn, copa_es, escola, flores_es, mgsm_direct_es, openbookqa_es, paws_es, phrases_es-va, phrases_va-es, veritasqa_gen_es, veritasqa_mc1_es, veritasqa_mc2_es, wnli_es, xlsum_es, xnli_es, xquad_es, xstorycloze_es
Basque	belebele_eus_Latn, eus_exams_eu, eus_proficiency, eus_reading, eus_trivia, flores_eu, mgsm_direct_eu, piqa_eu, qnlieu, wnli_eu, xcopa_eu, xnli_eu, xnli_eu_native, xstorycloze_eu
Galician	belebele_glg_Latn, flores_gl, galcola, mgsm_direct_gl, openbookqa_gl, parafrases_gl, paws_gl, summarization_gl, truthfulqa_gl_gen, truthfulqa_gl_mc1, truthfulqa_gl_mc2, veritasqa_gen_gl, veritasqa_mc1_gl, veritasqa_mc2_gl
Portuguese	assin_entailment, assin_paraphrase, belebele_por_Latn, flores_pt

Table 7: Languages included in IberoBench and their corresponding tasks used to compute the average scores for each language, as shown in Tables 2 and 5.