Multi-LMentry: Can Multilingual LLMs Solve Elementary Tasks Across Languages?

Anonymous ACL submission

Abstract

As large language models (LLMs) continue to improve, their evaluation increasingly centers on complex, high-level tasks, often at the expense of systematically assessing fundamental capabilities. To address this gap, recent work proposed LMentry, a compact benchmark comprising tasks that are trivial for humans but remain surprisingly difficult for LLMs. However, LMentry is limited to English, leaving its insights linguistically narrow. In this paper, we present Multi-LMentry, a ground-up recreation of LMentry that enables systematic evaluation of LLMs on basic reasoning and understanding tasks across nine diverse languages. Multi-LMentry includes English and expands to Catalan, German, Spanish, Basque, Galician, Korean, Italian, and Brazilian Portuguese, emphasizing the importance of cross-lingual and low-resource settings. To validate that Multi-LMentry is still trivial for humans, we demonstrate that L2 speakers with only elementary proficiency achieve near-perfect scores in a low-resource language, namely, Basque. Through extensive experiments, we reveal that state-of-the-art open-weight multilingual LLMs still fall short of human performance on elementary tasks in many languages. Our results expose new failure modes that remain hidden in monolingual evaluation, underscoring the need for rigorous, language-diverse "unit tests" of core model abilities.

1 Introduction

LLMs have shown remarkable performance across various complex tasks, including open-domain question answering, summarization, and reasoning. However, such success often overshadows fundamental model capabilities that underlie more complex reasoning processes. LMentry (Efrat et al., 2023) was proposed as a compact benchmark to test these "basic skills," using tasks that are trivial for humans—such as selecting which word is longer or producing a short sentence containing a target word—and systematically revealing surprising failure cases in LLMs that, at that time, represented the state of the art in the field, such as GPT-3 (Brown et al., 2020) and InstructGPT (Ouyang et al., 2022). Yet, LMentry is monolingual (English-only) and evaluates model outputs primarily through a regex-based automatic scoring scheme, a methodology prone to both undercounting and overcounting errors due to its rigid nature.

To fill this gap, we introduce **Multi-LMentry**, a new resource built from scratch to cover eight new languages in addition to English.[1] Specifically, we target both high-resource (e.g., German, Spanish) and low-resource languages (e.g., Basque, Galician), allowing us to identify performance bottlenecks and language-specific weaknesses in these simple tasks. Furthermore, many of the newly included languages exhibit unique morphological, orthographic, and syntactic properties, increasing the range of challenges and exposing previously overlooked model failures.

Our experiments demonstrate that, while leading open-weight multilingual LLMs demonstrate high performance on complex multilingual benchmarks, they still struggle to achieve human-level consistency in these "elementary" tasks, showing significant performance gaps across languages. Serving as a "unit test" for LLMs, Multi-LMentry offers a clearer perspective on the reliability of models before scaling them up or deploying them in real-world multilingual and cross-lingual applications. We release Multi-LMentry to the community and encourage its use as a baseline diagnostic tool, complementing larger multilingual benchmarks by focusing on the essential building blocks of language understanding.

Contributions. The main contributions of our work are:

[1] We release code and data at `anonymous.url/for/submission` licensed under Apache 2.0.

- **Multilingual extension of LMentry:** We present new versions of the original LMentry tasks in Catalan, German, Spanish, Basque, Galician, Korean, Italian, and Portuguese, covering both high- and low-resource languages.

- **Open-source benchmark:** We release Multi-LMentry for use by the community, providing a ready-to-use, extensible evaluation suite for testing elementary yet necessary abilities that all multilingual LLMs should have.

- **Extensive LLM Evaluation:** Using our proposed benchmark, we conducted a comprehensive evaluation of a wide range of LLMs, highlighting their performance on elementary-level tasks across nine different languages, showing important limitations of state-of-the-art LLMs over such tasks.

2 Related Work

The evaluation of LLMs has traditionally centered on performance over benchmark datasets targeting complex tasks. While these benchmarks offer insights into large-scale model capabilities, they can also mask some of their underlying weaknesses. In response, recent benchmarks have shifted toward more interpretable, small-scale tasks, which often expose unexpected brittleness in models previously considered "near-perfect" on ostensibly simpler objectives.

LMentry and "Elementary" Language Tasks. LMentry (Efrat et al., 2023) specifically addresses this gap by focusing on tasks trivial for humans with elementary-level language skills. Its findings highlight that even state-of-the-art English LLMs exhibit substantial errors when faced with tasks that a typical elementary student would solve flawlessly (e.g., counting letters, identifying which word in a short list belongs to a certain category). However, LMentry is designed to be an English-only benchmark. In contrast, Multi-LMentry refines the benchmark content by extending tasks to eight new different languages whenever possible, while also improving the evaluation setup and methodology.

Multilingual Benchmarks. While there exist robust multilingual evaluations such as XTREME (Hu et al., 2020), XGLUE (Liang et al., 2020), and XNLI (Conneau et al., 2018), these datasets focus on higher-level tasks such as cross-lingual natural language inference, question answering, and classification. They do not specifically isolate or test fundamental linguistic capabilities like those in LMentry, nor do they target orthographic or morphological phenomena that can derail performance on simpler tasks. Moreover, typical multilingual benchmarks often include only a small number of low-resource languages, limiting the investigation of cross-lingual generalization in truly resource-scarce settings. Multi-LMentry aims to fill these gaps by focusing on trivial tasks that stress minimal language understanding across diverse languages, including Basque and Galician, which feature unique linguistic properties.

Regex-Based Evaluation. To evaluate generative models on open-ended benchmarks, automatic scoring with manually defined regular expressions (regex) remains a widely adopted approach. Regex-based evaluation is fast, inexpensive, interpretable, and straightforward to implement, making it a practical choice for many tasks. Benchmark creators typically define a set of regex rules that capture expected output formats, as seen in datasets like GSM8K (Cobbe et al., 2021). Recent work has highlighted both the strengths and limitations of regex-based evaluation pipelines (Molfese et al., 2025), often comparing them against LLM-based answer extraction methods (Yu et al., 2025). These studies suggest that well-crafted, regex-based scoring can perform competitively, while avoiding the computational cost and potential biases introduced by large language models.

LLM-as-a-Judge Evaluation. An increasingly popular alternative is to use LLMs-as-a-Judge (Zheng et al., 2023), where strong models are prompted with carefully designed evaluation guidelines. However, state-of-the-art closed-source judges are expensive to deploy, whereas high-quality open-source judges are predominantly developed and validated in English (Lee et al., 2024; Kim et al., 2024). This limits their applicability and reliability in multilingual settings. Although recent efforts have extended LLM-based judges to multiple languages (Pombal et al., 2025), significant gaps remain. For example, some major languages like Brazilian Portuguese are not represented at all, and the performance of open-source judges is often suboptimal in mid- and low-resource languages such as Catalan, Basque, or Korean (e.g., Barnes et al., 2025). Furthermore, best-performing multilingual judges typically require backbones with significant model size (e.g., 14B parameters), which increases deployment costs.

Given these limitations, and with the goal of offering a lightweight and easy-to-use evaluation suite, Multi-LMentry adopts a manually curated set of multilingual regex rules. We further supplement these automated approaches with manual annotation for English and Basque, allowing us to perform a more nuanced study that balances scalability, flexibility, and reliability.

Overall, by focusing on minimal tasks in multiple languages and refining the scoring methodology, Multi-LMentry helps ensure that basic linguistic aptitudes are no longer overlooked in the race toward ever more sophisticated AI benchmarks.

3 Multi-LMentry

In this section, we present the details of Multi-LMentry, our manual multilingual extension of the original LMentry framework. We describe the task design and adaptation process, the evaluation methodology, and the statistics of the benchmark.

3.1 Languages in Multi-LMentry

Multi-LMentry extends the original LMentry framework (Efrat et al., 2023) by expanding its elementary-level language tasks across eight additional languages beyond English: Basque, Catalan, Galician, German, Italian, Portuguese (Brazil), Spanish, and Korean. This multilingual expansion provides a valuable tool for evaluating the cross-linguistic capabilities of LLMs.

We categorize these languages according to the availability of publicly accessible resources, including large corpora and language-specific models (Nguyen et al., 2024). This classification yields three distinct groups: (1) low-resource languages (Catalan, Galician, Basque, and Brazilian Portuguese); (2) mid-resource languages (Italian and Korean); and (3) high-resource languages (English, Spanish, and German). The inclusion of low-resource languages is particularly significant, as it allows us to investigate the performance of LLMs in settings where training data is scarce. This is crucial for understanding how well these models can generalize across languages with limited resources and linguistic diversity.

3.2 Task Design and Adaptation

The original LMentry framework comprises 25 elementary-level tasks, ranging from simple sentence construction to contextual word selection. Table 1 presents the complete task inventory with corresponding example prompts. Each of these tasks has been systematically implemented across all nine languages in our extended framework, except for a few tasks that were not applicable to certain languages, e.g., the *Rhyming word* task; we provide a comprehensive overview of all the tasks that we implemented in each language and the corresponding number of samples in Table 54.

The original LMentry tasks were designed to operate within specific linguistic constraints, and we preserved these constraints in our multilingual extension. Each task was designed to be: (1) easily solvable by native speakers, (2) independent of domain-specific knowledge, (3) concise, and (4) suitable for straightforward automatic evaluation. To create each task of Multi-LMentry for each language, we involved native speakers of the target languages to ensure that data collection and task design were linguistically sound and culturally relevant. This process involved the following steps:

- **Data Creation:** It is important to note that, while the goal of Multi-LMentry is to create a multilingual version of LMentry, Multi-LMentry is not a direct translation of the original LMentry. Indeed, to ensure we avoid any errors, ambiguity, or potential biases, we manually recreated the data in each language.

- **Task Adaptation:** We adapted the original LMentry tasks to fit the linguistic characteristics of each target language. This involved modifying the task prompts and evaluation criteria to ensure they were appropriate for the target language, e.g., the *Ends with letter* task was adapted to account for the different alphabetic systems and orthographic rules of each language.

During the task adaptation process, annotators were encouraged to provide feedback on the task design and implementation, allowing for iterative improvements. Each task originally includes three distinct templates to ensure broader linguistic coverage. As part of our annotation process, we refined the original English templates to better align with our experimental setting. These adapted templates were then manually translated into the eight target languages. The final English templates are provided in Appendix A.3.

Task	Example
Sentence containing word	Write a sentence that contains the word "cats":
Sentence not containing word	Write a sentence that doesn't contain the word "happy":
Word containing letter	Write a word that contains the letter "s":
Word not containing letter	Write a word that doesn't contain the letter "t":
Most associated word	Of the words "skirt", "pants", "jacket", "dog", and "jeans", what is the word most commonly associated with "animals"?
Least associated word	Of the words "banana", "motorcycle", "mango", "lemon", and "strawberry", what is the word least associated with "fruit"?
Any words from category	Are any of the words "rabbit", "car", "cat", "mouse", or "bird" types of vehicles? Answer either "yes" or "no".
All words from category	Are all the words "chair", "bed", "table", "desk", and "sofa" types of furniture? Answer either "yes" or "no".
First alphabetically	In an alphabetical order, which of the words "book" and "water" comes first?
More letters	Which word has more letters, "city" or "drink"?
Less letters	Which word has fewer letters, "day" or "computer"?
Bigger number	Which number is bigger, 147 or 246?
Smaller number	Which number is smaller, 278 or 802?
Rhyming word	Which word rhymes with the word "try", "food" or "cry"?
Homophones	Of the two words "eight" and "mouth", which one sounds more like "ate"?
Word after in sentence	In the sentence "The door was pushed open", which word comes right after the word "was"?
Word before in sentence	In the sentence "You may pick any flower", which word comes right before the word "any"?
Sentence starting with word	Write a sentence that starts with the word "trains":
Sentence ending with word	Write a sentence that ends with the word "today":
Word starting with letter	Write a word that starts with the letter "e":
Word ending with letter	Write a word that ends with the letter "h":
First word of the sentence	What is the first word of the sentence "Everyone hoped that she would sing"?
Last word of the sentence	What is the last word of the sentence "There is a bench for you to sit on"?
First letter of the word	What is the first letter of the word "apples"?
Last letter of the word	What is the last letter of the word "piano"?

Table 1: LMentry consists of 25 short tasks which are trivial to humans. Each task has three different templates. Templates are phrased either as an instruction, or as a question. Templates are instantiated with *arguments* (in blue). Full details on task templates and arguments are in Appendix A.3.

3.3 Evaluation Methodology

Trivial Tasks should be Trivial for LLMs. We build Multi-LMentry for zero-shot evaluation: no in-context samples should be given to facilitate the task for the models under evaluation, and no further training or fine-tuning should be performed on the models, in line with the original LMentry framework. This approach allows us to assess the models' performance on the tasks without any prior exposure to the specific task formats or examples. By maintaining a zero-shot evaluation setup, we ensure that our benchmark remains a true test of the models' capabilities, rather than a measure of their ability to memorize or adapt to specific examples or prompts. Indeed, few-shot examples and fine-tuning are known to bias predictions (Si et al., 2023; Molfese et al., 2025).

What is the Answer? Due to the trivial nature of the task, the answers are often very simple. The fact that the answers have a very simple structure allows us to use regex-based evaluation, adopting a similar approach to the original LMentry framework as well as other generative benchmarks, such as GSM8K and MATH. In this way, we can evaluate the models' performance without imposing any constraints on the answer's structure. In Multi-LMentry, we define approximately 10 regex patterns per task, which are used to evaluate the models' outputs. The regex patterns are specifically designed for each language independently to accommodate variations in the answers while still ensuring that they meet the criteria for correctness.

Can Regex Patterns Capture All Valid Answers? It is important to note that regex-based evaluation is not without its limitations. While regex patterns can effectively capture a wide range of valid answers, they may also miss some correct outputs that do not conform to the predefined patterns. For example, in languages with rich morphology or complex syntactic structures, the same answer may be expressed in multiple ways, making it challenging to create comprehensive regex patterns that cover all possibilities.

To assess the reliability of our evaluation methodology, we conducted targeted agreement analysis with human annotations in English and Basque, the

Language	Cohen's Kappa	Accuracy
English	79.8	90.8
Basque	53.9	81.4

Table 2: Cohen's Kappa and Accuracy agreement between regex and annotated examples for English and Basque, averaged over all the tasks.

latter being a morphologically and syntactically richer language. We validated the scoring mechanism by examining $1,000$ randomly sampled predictions per task across random models, avoiding bias toward any particular model family or architecture. Results are reported in Table 2, with additional results per task in Appendix A.6. In English, we observe an accuracy agreement of 90.8% and a Cohen's Kappa of 79.8% (usually interpreted as in between *Almost Perfect Agreement* and *Substantial Agreement*), indicating reliable performance of regex patterns. For Basque, we observe an accuracy agreement of 81.4% and a Cohen's Kappa of 53.9% (*Moderate Agreement*).

Therefore, especially in lieu of the results discussed in Section 5, we can conclude that regex-based evaluation is a reliable and efficient method for evaluating the performance of LLMs on elementary-level tasks in multiple languages, especially given the current limitations of LLM-as-a-Judge approaches, as discussed in Section 2 and the cost of using closed-source LLMs as judges.[2]

Metrics. We evaluated the models' predictions using the canonical accuracy metric (*number of correctly predicted labels*) and LMentry-Score (LMS), which is computed multiplying the *accuracy* by the *robustness* score, over each task. LMentry-Score was defined in the original paper of LMentry (Efrat et al., 2023), where the authors defined the robustness as the mean over four robustness aspects: (1) argument order, (2) argument content, (3) template, and (4) adjacent tasks, over all the LMentry tasks. Robustness is computed as the difference in the accuracy scores between tasks clustered by the mentioned aspects. A detailed categorization of robustness aspects is reported in Appendix A.2.

[2]At the time of writing, the cost of using OpenAI's o1-mini to evaluate 10 models on all the languages is around $5,000. This is a significant cost for a single evaluation, especially considering that the same evaluation would need to be repeated for each new model or language.

Language	Lang. Code	Num. Samples
Basque	EU	105,279
Brazilian Portuguese	PT_BR	106,980
Catalan	CA	108,090
English	EN	110,703
Galician	GL	106,062
German	DE	110,040
Italian	IT	105,690
Korean	KO	95,799
Spanish	ES	107,601
Total		956,244

Table 3: Samples per language in Multi-LMentry.

Human Baselines. To assess the fundamental simplicity of our multilingual benchmark, we conducted manual evaluations with native speaker annotators for each language, who analyzed 20 examples per task, and confirmed that human native speakers have nearly-perfect ability to solve the benchmark. Moreover, to further assess the elementary-level accessibility of our proposed multilingual version of LMentry, we recruited three beginner-level non-native learners of Basque to complete a significative subsample of the benchmark (details available in Appendix A.4). Even then, these second language learners were able to solve the task with 96% accuracy, demonstrating that (1) our benchmark represents a unique collection of tasks that are trivial for humans yet suitable for assessing the basic linguistic capabilities of LLMs, and (2) data scarcity shall not be a discriminating factor for truly "intelligent" LLMs.

Statistics. Multi-LMentry is composed of a set of 25 tasks per language. Each task is formulated in different ways or *subtasks* to assess the robustness aspect of evaluated models (see Appendix A.2). The number of total samples per language is reported in Appendix 3. Multi-LMentry represents a rich benchmark of nearly 1M samples equally divided by 9 languages. Additional stats can be found in Appendix A.7.

4 Experimental Setup

We evaluated a diverse set of instruction-tuned LLMs, having found in preliminary tests that, unsurprisingly, base models tend to complete the text rather than answering the question. As shown in Table 4, our selection includes models ranging from 360M to 14B parameters to assess how performance on elementary tasks scales with model

Model Name	Size	Languages
Closed Multilingual Models		
meta-llama/Llama-3.2-1B	1B	Multi.
Qwen/Qwen2.5-1.5B-Instruct	1.5B	Multi.
meta-llama/Llama-3.2-3B	3B	Multi.
Qwen/Qwen2.5-3B-Instruct	3B	Multi.
meta-llama/Llama-3.1-8B-Instruct	8B	Multi.
Qwen/Qwen2.5-14B-Instruct	14B	Multi.
microsoft/phi-4	14B	Multi.
Open Multilingual Models		
utter-project/EuroLLM-1.7B-Instruct	1.7B	Multi.
BSC-LT/salamandra-2b-instruct	2B	Multi.
BSC-LT/salamandra-7b-instruct	7B	Multi.
occiglot/occiglot-7b-eu5-instruct	7B	Multi.
utter-project/EuroLLM-9B-Instruct	9B	Multi.
Language Specific		
HuggingFaceTB/SmolLM2-360M-Instruct	360M	EN
HuggingFaceTB/SmolLM2-1.7B-Instruct	1.7B	EN
occiglot/occiglot-7b-es-en-instruct	7B	ES
swap-uniba/LLaMAntino-3-ANITA-8B-Inst-DPO-ITA	8B	IT
HiTZ/Latxa-Llama-3.1-8B-Instruct	8B	EU

Table 4: Size and languages of selected models.

size. We categorize the evaluated models into three distinct categories:

Open-Weights Multilingual LLMs: The authors declared that these LLMs are trained on multilingual data but the sources and composition of the pre-training and post-training data is not known. Among these models, we select Qwen-2.5 (Team, 2024), the Llama-3 (Grattafiori et al., 2024) family, and Phi-4 (Abdin et al., 2024), since they represent the state of the art at the time of writing.

Open-Data Multilingual LLMs: The authors declared that these LLMs are trained on multilingual data and that sources and composition *are* documented. Moreover, we prioritize models with permissive licenses, including EuroLLM (Martins et al., 2024), Salamandra (Gonzalez-Agirre et al., 2025), and Occiglot-eu.

Language-Specific LLMs: The authors trained (or adapted) the models for a specific language. We select the following models from the available ones: SmolLM2 (Allal et al., 2025), a family of small models trained on English data; Occiglot-es-en, a Spanish adaption of Mistral-7B; ANITA (Polignano et al., 2024) an Italian adaptation of Llama-3.1-8B-Instruct; and Latxa (Sainz et al., 2025), a Basque adaptation of the same backbone model.

This categorization helps us analyze how multilingual and language-specific LLMs perform on Multi-LMentry across different languages. This approach can provide insights into how training data composition affects model capabilities on elementary-level tasks across diverse languages.

5 Results and Discussion

In what follows, we divide our analysis over different aspects: first, we provide an overview of the general results, showing the brittleness of current multilingual LLMs; second, we analyze the performance of various LLMs with different size; third, we discuss the gap between open-weights and open-data LLMs.

5.1 Main Results

We report LMS and accuracy for each model over the nine languages in Multi-LMentry in Table 5. We observe that, on average, current LLMs perform achieve stronger results in English. With an average LMS score of 49.5% and average accuracy of 60.9%, there is a significant gap between English and all the other languages (at least 9 points in LMS and 13 points in accuracy compared to the second language). This result reinforces the idea that there is still a strong bias toward English among LLMs, which still exists even for the simple tasks in Multi-LMentry.

What is less straightforward are the scores of the other languages. We observe that the second-best performing language on average is Korean, which may counterintuitive, especially since we do not include any Korean-specific LLM or any LLM that has been reported to have been trained on significant quantities of Korean data. We hypothesize that this result may depend on certain linguistic features – linked to elementary tasks – specific to Korean. Despite being a high-resource language, German is the most challenging one, with an average LMS of 17.5% and an average accuracy of 21.0%. All models struggle to reach satisfying results, as the best LMS is only 32.9%. We argue that the nature of the language itself impacts the scores of the models, as elementary-level syntactic and semantic linguistic realizations in German can be more challenging than in other languages due to its complex morphology, compound word structures, and flexible word order. Other languages exhibit varying performance levels. Notably, Spanish achieves an average LMS of 37.7%, which can be attributed to the availability of Spanish data on the Web. In contrast, Basque, being a low-resource language, attains a very low LMS, around 19.8%.

Interestingly, language adaptation plays a signif-

Model	EN		ES		DE		IT		KO		CA		GL		EU		PT_{BR}	
	LMS	Acc.	LMS	Acc.	LMS	Acc.	LMS	Acc.	LMS	Acc.	LMS	Acc.	LMS	Acc.	LMS	Acc.	LMS	Acc.
Closed Multilingual Models																		
Llama-3.2-1B	58.7	70.4	41.8	50.8	14.5	16.2	25.2	30.9	48.1	54.4	29.1	35.7	24.2	28.1	10.1	11.2	26.4	35.3
Qwen2.5-1.5B	42.2	57.1	33.0	42.7	18.4	21.4	25.7	32.5	52.0	57.8	33.2	40.4	25.7	30.5	10.2	11.7	25.1	30.8
Llama-3.2-3B	71.0	84.3	49.2	63.4	20.2	24.6	28.8	37.5	49.2	55.6	37.1	52.1	30.8	35.9	33.8	42.3	42.8	53.8
Qwen2.5-3B	56.6	71.1	40.5	53.6	23.2	29.9	34.2	45.2	53.0	59.1	37.7	49.0	27.3	34.6	15.0	18.3	36.6	47.1
Llama-3.1-8B	77.5	88.3	51.7	66.9	16.5	19.3	40.2	50.0	51.6	59.9	37.4	50.0	31.2	38.0	36.6	45.1	40.2	52.9
Qwen2.5-14B	77.6	88.6	53.2	66.6	27.4	33.9	37.9	48.8	57.0	60.4	44.3	57.2	33.8	41.2	29.2	37.0	43.6	55.5
phi-4	78.2	89.2	56.0	67.7	23.8	29.2	38.3	47.8	59.5	62.5	49.1	60.3	37.8	45.4	34.1	43.0	42.8	53.0
Open Multilingual Models																		
EuroLLM-1.7B	32.9	42.7	30.1	36.6	13.0	15.3	32.2	39.2	49.2	55.9	17.0	21.4	23.7	28.2	8.8	9.3	25.3	31.6
salamandra-2b	25.2	30.0	19.1	22.2	17.5	19.7	17.6	20.6	4.7	4.9	21.3	22.7	17.4	19.8	14.9	17.7	20.2	22.2
salamandra-7b	34.5	44.4	25.8	30.9	20.1	24.6	25.7	32.0	11.1	12.0	28.7	35.3	26.9	31.3	21.1	25.8	27.6	34.8
occiglot-7b-eu5	26.6	40.0	28.7	37.5	13.1	15.7	21.4	27.2	15.8	17.9	15.6	18.1	9.4	10.4	9.0	10.1	24.1	30.3
EuroLLM-9B	43.9	60.2	37.5	48.4	13.6	16.7	20.5	27.1	51.5	56.1	31.0	40.6	23.0	28.7	12.1	13.8	27.3	35.2
Language Specific Models																		
SmolLM2-360M	31.6	40.3	24.1	30.0	8.0	8.9	18.0	21.0	48.9	56.5	7.8	8.3	17.1	20.3	5.0	5.2	19.4	23.0
SmolLM2-1.7B	43.8	59.0	32.8	44.6	14.3	17.2	29.0	38.1	37.3	45.2	28.4	38.7	24.5	32.9	5.9	6.2	34.2	44.3
occiglot-7b-es-en	38.9	52.5	33.7	43.4	8.5	9.4	19.7	24.1	22.2	24.7	14.9	17.2	6.7	7.0	8.3	9.0	20.6	25.9
ANITA-8B	70.9	83.5	62.3	76.8	32.9	42.6	61.3	75.1	36.5	42.6	59.1	71.0	50.2	63.8	37.2	47.0	59.0	74.2
Latxa-8B	70.4	83.8	45.7	61.3	18.1	21.0	33.6	43.7	51.2	56.5	33.0	42.8	30.8	37.7	50.9	62.4	33.3	43.6
Average	49.5	60.9	37.7	47.7	17.5	21.0	29.3	36.7	40.6	45.5	30.1	37.8	25.5	30.8	19.8	24.0	31.5	39.8

Table 5: Results per model across language on Multi-LMentry tasks. Results are highlighted by language family: high-resource , mid-resource , and low-resource .

Model	EN	ES	DE	IT	KO	CA	GL	EU	PT_{BR}
Closed Multilingual Models									
Llama-3.2-1B	83.3	82.3	89.1	81.7	88.4	81.5	86.4	89.9	74.8
Qwen2.5-1.5B	73.8	77.1	85.9	79.1	90.0	82.2	84.2	86.8	81.7
Llama-3.2-3B	84.2	77.7	82.1	76.7	88.5	71.3	85.7	79.7	79.4
Qwen2.5-3B	79.6	75.5	77.6	75.8	89.7	77.0	78.7	82.0	77.7
Llama-3.1-8B	87.8	77.2	85.6	80.4	86.2	74.7	82.3	81.1	75.9
Qwen2.5-14B	87.6	80.0	81.0	77.6	94.5	77.4	82.0	79.0	78.6
phi-4	87.7	82.7	81.4	80.2	95.2	81.4	83.3	79.4	80.7
Open Multilingual Models									
EuroLLM-1.7B	77.0	82.1	84.8	82.3	88.0	79.7	84.1	93.7	80.0
salamandra-2b	84.1	85.9	88.5	85.7	96.7	93.7	87.8	84.3	90.7
salamandra-7b	77.8	83.7	82.0	80.3	91.9	81.2	86.2	81.8	79.4
occiglot-7b-eu5	66.4	76.6	83.6	78.9	88.3	86.6	90.4	89.0	79.5
EuroLLM-9B	72.9	77.3	81.1	75.5	91.8	76.4	80.2	87.4	77.5
Language Specific Models									
SmolLM2-360M	78.5	80.3	90.4	85.7	86.6	93.4	84.6	96.6	84.4
SmolLM2-1.7B	74.2	73.4	83.3	76.0	82.5	73.3	74.6	95.3	77.3
occiglot-7b-es-en	74.2	77.6	90.0	82.0	89.7	86.4	94.9	92.6	79.3
ANITA-8B	85.0	81.1	77.4	81.6	85.7	83.3	78.7	79.2	79.6
Latxa-8B	84.0	74.6	86.0	77.0	90.6	77.2	81.7	81.5	76.3
Average	80.4	79.7	84.5	80.1	89.6	81.3	84.2	86.0	79.8

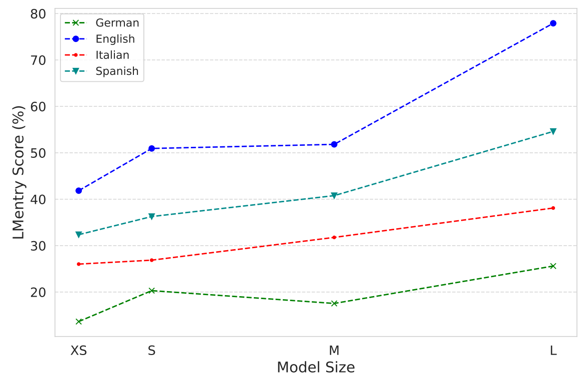
Table 6: Robustness results per model across language on Multi-LMentry tasks. Results are highlighted by language family: high-resource , mid-resource , and low-resource .

icant role: models adapted to a specific language often report improved results in the target language and also in the languages that are linguistically close to the target language. For example, ANITA achieves strong results in Italian, but also performs well in Catalan, Galician, Spanish and Portuguese, which are all Romance languages related to Italian. We can observe a similar trend for Latxa in Basque. This result highlights the importance of adaptation toward specific languages to improve elementary-level language abilities.

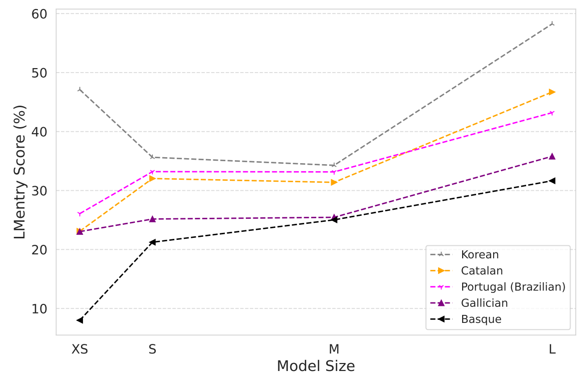
Robustness Scores. The robustness scores for each model are reported in Table 6. Overall, the models exhibit strong robustness across all languages, with average scores exceeding 80%. These findings indicate that, regardless of the language, the models remain consistently resilient across the four robustness dimensions: (1) argument order, (2) argument content, (3) template variation, and (4) adjacent tasks. Notably, manual inspection of less-than-perfect robustness scores reveals the presence of spurious correlations that models rely on to solve elementary-level tasks, e.g., word length, word frequency, most frequent senses.

5.2 Does Model Size Matter?

We distinguish between four categories of LLMs based on their parameter count: *i)* Extra Small (XS), with fewer than 2B parameters; *ii)* Small (S), with fewer than 7B parameters; *iii)* Medium (M), with fewer than 14B parameters; and *iv)* Large (L), corresponding to models with 14B parameters. Fig. 1 shows a clear overall trend: larger models generally yield better performance in English, which aligns with expectations. However, the scaling does not always hold true for other languages. For instance, in Italian, the S models only show a modest improvement over XS models, while M models do not outperform S models by a significant margin, especially for open-data models. This suggests that there is still a wide gap between linguistic capabilities and model size, and that the performance of LLMs on elementary-level tasks is not solely determined by the number of parameters.

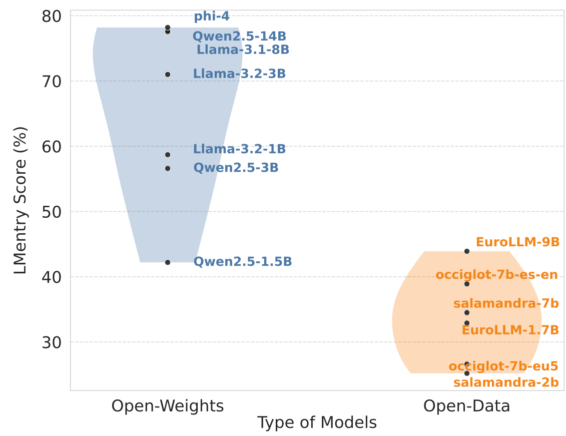


(a) Averaged LMentry Scores for English, German, Spanish, and Italian

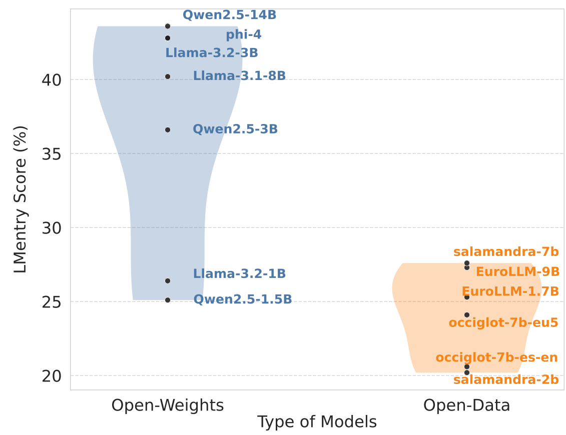


(b) Averaged LMentry Scores for Korean, Catalan, Galician, Basque, and Portuguese (Brazilian)

Figure 1: Averaged LMentry scores over different model sizes. On the x-axis, model sizes are reported in four different scales: Extra Small (XS, $<$ 2B), Small (S, $<$ 7B), Medium (M, $<$ 14) and Large (L, $\geq$ 14B)



(a) English language



(b) Portuguese (Brazilian) language

Figure 2: LMentry-Scores averaged across two languages, English (high-resource) and Brazilian Portuguese (low-resource). Left: *open-weights* models. Right: *open-data* models.

5.3 Open-Weights vs Open-Data LLMs

To assess how model openness affects multilingual performance, we compared LMS across English (high-resource) and Brazilian Portuguese (low-resource). As shown in Fig. 2, clear performance gaps exist between open-weights and open-data models. This finding highlights the value of benchmarks such as Multi-LMentry and the significant gap that open research must bridge to reach the performance level of commercial models on elementary linguistic tasks across languages.

6 Conclusion

This paper introduces Multi-LMentry, the first multilingual benchmark for elementary level tasks, covering nine diverse languages: English, Spanish, German, Italian, Catalan, Galician, Basque, Brazilian Portuguese, and Korean. Our validation confirms the benchmark's accessibility, as even elementary-level L2 Basque speakers solve these tasks with near-perfect accuracy. Despite their success on complex tasks, our evaluation of 17 state-of-the-art LLMs reveals that none achieve human-comparable performance on these elementary tasks. Results consistently show lower performance in non-English languages, with a significant gap between open-weights and open-data models. These findings challenge claims about LLM capabilities and highlight the need for continued research in multilingual contexts. Future work should expand Multi-LMentry to include additional low-resource languages, where performance gaps are most pronounced. We hope that this benchmark will serve as a valuable resource for researchers and practitioners, enabling them to better understand the limitations and capabilities of LLMs in multilingual settings.

Limitations

While our Multi-LMentry benchmark represents a significant advancement in multilingual evaluation of elementary language capabilities, we acknowledge several limitations that present opportunities for future research.

- **Evaluation methodology limitations.** Our benchmark relies on manually curated regex patterns for evaluation. Although this approach provides transparency and efficiency, it faces challenges with morphologically rich languages like Basque, where we observed only moderate agreement (Cohen's Kappa of 53.9%) compared to human evaluation. Future work could explore hybrid evaluation approaches that combine regex patterns with lightweight, language-specific LLM-as-judge methods, particularly focusing on improving evaluation for low-resource languages.

- **Language coverage constraints.** While we include nine diverse languages spanning different resource levels and linguistic families, this represents only a fraction of the world's languages. Future extensions should prioritize languages with distinct linguistic properties (e.g., tonal languages, polysynthetic languages) and extremely low-resource languages that are currently underrepresented in LLM research.

- **Model size constraints.** Our analysis was constrained to models up to 14B parameters due to computational limitations. Expanding evaluation to larger models ($\geq$70B parameters) and closed-source commercial models would provide more comprehensive insights into how elementary capabilities scale with model size. Additionally, evaluating instruction-tuned models specifically aligned for multilingual understanding could reveal whether alignment techniques can narrow the performance gaps we observed.

- **Error analysis depth.** While we identify performance gaps across languages, our work would benefit from more fine-grained error analysis across specific tasks and languages. Future work should systematically categorize error patterns to better understand which elementary capabilities are most challenging for models in different linguistic contexts.

- **Reasoning approaches.** We did not include models that incorporate explicit reasoning processes (e.g., as in (DeepSeek-AI et al., 2025)). As these models have shown promise on complex tasks, investigating their performance on elementary tasks could reveal whether explicit reasoning helps or hinders performance on seemingly simple linguistic operations. Future research could compare chain-of-thought approaches with direct responses on Multi-LMentry tasks.

We believe that addressing these limitations deserve further investigation and will contribute to a more comprehensive understanding of LLM capabilities in multilingual contexts.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, et al. 2025. Smollm2: When smol goes big–data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*.

Jeremy Barnes, Naiara Perez, Alba Bonet-Jover, and Begoña Altuna. 2025. Summarization metrics for spanish and basque: Do automatic scores and llm-judges correlate with humans? *arXiv preprint arXiv:2503.17039*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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 *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro

Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

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.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, and Chenggang Zhao. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Avia Efrat, Or Honovich, and Omer Levy. 2023. LMentry: A language model benchmark of elementary language tasks. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10476–10501, Toronto, Canada. Association for Computational Linguistics.

Aitor Gonzalez-Agirre, Marc Pàmies, Joan Llop, Irene Baucells, Severino Da Dalt, Daniel Tamayo, José Javier Saiz, Ferran Espuña, Jaume Prats, Javier Aula-Blasco, et al. 2025. Salamandra technical report. *arXiv preprint arXiv:2502.08489*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, and Akhil Mathur. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

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. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353, Miami, Florida, USA. Association for Computational Linguistics.

Seongyun Lee, Seungone Kim, Sue Park, Geewook Kim, and Minjoon Seo. 2024. Prometheus-vision: Vision-language model as a judge for fine-grained evaluation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11286–11315, Bangkok, Thailand. Association for Computational Linguistics.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M Guerreiro, Ricardo Rei, Duarte M Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, et al. 2024. Eurollm: Multilingual language models for europe. *arXiv preprint arXiv:2409.16235*.

Francesco Maria Molfese, Luca Moroni, Luca Gioffrè, Alessandro Scirè, Simone Conia, and Roberto Navigli. 2025. Right answer, wrong score: Uncovering the inconsistencies of llm evaluation in multiple-choice question answering. *Preprint*, arXiv:2503.14996.

Mozilla Foundation. 2025. Common voice dataset. `https://commonvoice.mozilla.org/en/datasets`. Accessed: 2025-05-19.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2024. CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4226–4237, Torino, Italia. ELRA and ICCL.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Marco Polignano, Pierpaolo Basile, and Giovanni Semeraro. 2024. Advanced natural-based interaction for the italian language: Llamantino-3-anita. *arXiv preprint arXiv:2405.07101*.

José Pombal, Dongkeun Yoon, Patrick Fernandes, Ian Wu, Seungone Kim, Ricardo Rei, Graham Neubig, and André F. T. Martins. 2025. M-prometheus: A suite of open multilingual llm judges. *Preprint*, arXiv:2504.04953.

Oscar Sainz, Naiara Perez, Julen Etxaniz, Joseba Fernandez de Landa, Itziar Aldabe, Iker García, Aimar Zabala, Ekhi Azurmendi, German Rigau, Eneko Agirre, Mikel Artetxe, and Aitor Soroa. 2025. Instructing large language models for low-resource languages: A systematic study for basque. *arXiv preprint TBP*.

Chenglei Si, Dan Friedman, Nitish Joshi, Shi Feng, Danqi Chen, and He He. 2023. Measuring inductive biases of in-context learning with underspecified demonstrations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11289–11310, Toronto, Canada. Association for Computational Linguistics.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Qingchen Yu, Zifan Zheng, Shichao Song, Zhiyu Li, Feiyu Xiong, Bo Tang, and Ding Chen. 2025. xfinder: Large language models as automated evaluators for reliable evaluation. *Preprint*, arXiv:2405.11874.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

A Appendix

A.1 System Prompts

In out experiments we used only instruction tuned LLMs that use a chat template with a mandatory system prompt. The detailed list of system prompt is available at Table 7.

A.2 Detailed Information about Robustness-score

For the sake of clarity we report the details on how Robustness score is computed over all the tasks in Multi-LMentry, following the details defined in the original LMentry effort (Efrat et al., 2023)

Multi-LMentry measures four aspects of robustness (1) argument order, (2) argument content, (3) template, and (4) adjacent tasks. As original LMentry tasks are diverse in form, not every robustness aspect can be measured on every task.

Argument Order We measure argument order robustness on tasks where the correct answer is either of the two given arguments. These tasks are: *more letters*, *less letters*, *first alphabetically*, *rhyming word*, *homophones*, *bigger number*, *smaller number*.

Argument Content Argument content robustness is the accuracy gap between different argument subsets of the same task, where the difference between the subsets is naive from a human perspective. We measure argument content robustness on six tasks. For each of these tasks, a sub-task from each argument subset to increase the statistical power. The tasks from which Argument Content robustness is computed are listed and detailed in Table 8.

Template Robustness Template Robustness is measured on all LMentry tasks, for each language.

Adjacent tasks is a pair of similar tasks which differ in a specific aspect, e.g., sentence containing word and sentence not containing word, or more letters and less letters. Adjacent tasks robustness is the mean accuracy gap over all the pairs of adjacent tasks. We consider the following pairs of adjacent tasks:

- *any words from category*, *all words from category*
- *most associated word*, *least associated word*
- *more letters*, *less letters*
- *bigger number*, *smaller number*
- *word after in sentence*, *word before in sentence*
- *sentence starting with word*, *sentence ending with word*
- *word starting with letter*, *word ending with letter*
- *first word of the sentence*, *last word of the sentence*
- *first letter of the word*, *last letter of the word*
- *sentence containing word*, *sentence not containing word*
- *word containing letter*, *word not containing letter*
- *sentence containing word*, *word containing letter*
- *sentence not containing word*, *word not containing letter*
- *sentence starting with word*, *word starting with letter*
- *sentence ending with word*, *word ending with letter*
- *first word of the sentence*, *first letter of the word*

Language	System Prompt
English (EN)	You are a helpful assistant. Answer concisely and correctly.
Italian (IT)	Sei un assistente utile. Rispondi in modo conciso e corretto.
Catalan (CA)	Ets un assistent útil. Repon de forma concisa i correcta.
Spanish (ES)	Eres un asistente útil. Responde de forma concisa y correcta.
Basque (EU)	Laguntzaile erabilgarri bat zara. Erantzun laburki eta zuzen.
Galician (GL)	Vostede é un asistente útil. Responde de forma concisa e correcta.
Korean (KO)	당신은 유용한 보조 역할을 수행해야합니다. 간결하고 정확하게 답변해 주세요.
German (DE)	Sie sind ein hilfreicher Assistent. Antworten Sie präzise und richtig.
Portuguese-BR (PT_BR)	Você é um assistente prestativo. Responda de forma concisa e correta.

Table 7: System prompts used for each language in our multilingual benchmark.

Task	Argument Subset	
more letters	$\lvert \texttt{len}(w_1) - \texttt{len}(w_2) \rvert \geq 3$	(one of the words is longer than the other by at least 3 letters)
	$\lvert \texttt{len}(w_1) - \texttt{len}(w_2) \rvert = 1$	(one of the words is longer than the other by exactly one letter)
less letters	$\lvert \texttt{len}(w_1) - \texttt{len}(w_2) \rvert \geq 3$	(one of the words is shorter than the other by at least 3 letters)
	$\lvert \texttt{len}(w_1) - \texttt{len}(w_2) \rvert = 1$	(one of the words is shorter than the other by exactly one letter)
first alphabetically	$\lvert w_1[0] - w_2[0] \rvert \geq 13$	(the first letters of w_1 and w_2 are at least 13 letters apart alphabetically)
	$\lvert w_1[0] - w_2[0] \rvert > 1$	(the first letters of w_1 and w_2 are different)
	$\lvert w_1[0] - w_2[0] \rvert = 1$	(the first letters of w_1 and w_2 are different, but consecutive (e.g. c,d or p,o))
	$\lvert w_1[0] - w_2[0] \rvert = 0$	(w_1 and w_2 have the same first letter)
any words from category	None of the 5 words belong to the category	
	1 of the 5 words belongs to the category	
	2 of the 5 words belong to the category	
all words from category	All 5 words belong to the category	
	4 of the 5 words belong to the category	
	3 of the 5 words belong to the category	
rhyming word	The answer is orthographically similar to the query, and orthographically dissimilar from the distractor	
	The answer is orthographically dissimilar from the query, and orthographically similar to the distractor	

Table 8: The argument subsets of the tasks on which argument content robustness is measured.

- *last word of the sentence*, *last letter of the word*

A.3 Templates per Tasks

All Multi-LMentry tasks are accompanied by multiple templates. These templates are listed in the following tables, from Table 9 to Table 33. The table descriptions also provide information about the types of data used to construct the benchmark. Only the English templates are shown; templates for other languages were manually translated from the English versions.

A.4 Evaluation with Elementary L2 learners

To assess the elementary nature of our benchmark, we conducted a human evaluation study with non-native Basque learners. We recruited 3 annotators, all female, aged 25-30, and with native Spanish language backgrounds. None were from the Basque Country. All annotators had formal training in linguistics and possessed A2-level (CEFR) proficiency in Basque, representing beginner to elementary learners of the language. The evaluation used a test set comprising 5 randomly sampled examples from each task variant, totaling 195 examples. These examples were randomly distributed among the three annotators, and were presented in the same format as they were presented to the LLMs. The annotators correctly answered 188 out of 195 examples (96.41%). This performance by individuals with elementary Basque language skills validates our benchmark's design as representing fundamental linguistic capabilities, supporting its appropriateness for evaluating the basic linguistic capabilities of LLMs.

A.5 Results per Language

The detailed results for each language are presented in Table 34 to Table 51, showing the accuracy of each model across all tasks and their variations. The corresponding average scores were discussed in Section 5.

Sentence Containing Word
Write a sentence that contains the word “*word*”:
Write a sentence using the word “*word*”:
Write a sentence containing the word “*word*”:

Table 9: The templates of the *sentence containing word* task. $word$ is a basic word (CEFR level A1 or A2 for each language).

Sentence Not Containing Word
Write a sentence that doesn’t contain the word “*word*”:
Write a sentence without using the word “*word*”:
Write a sentence not containing the word “*word*”:

Table 10: The templates of the *sentence not containing word* task. $word$ is a basic word (CEFR level A1 or A2, for each language).

A.6 Agreement per Tasks

The agreement per task is reported in Tables 52 and 53 for English and Basque, respectively. As shown by the results, English exhibits very high agreement across most tasks. This high level of agreement is generally preserved in Basque, with the exception of a few tasks—namely, ends with word, sentence containing, and start with word—where model responses tend to be more variable. These discrepancies in agreement scores highlight the need for further research in evaluation methodologies, particularly to enable more robust evaluation frameworks for low-resource languages.

A.7 Detailed Samples per Task

Table 54 reports the number of samples for each task—and their variations—across different languages. A "—" indicates that a given task is not implemented for a particular language. These variations arise from language-specific constraints; for instance, some tasks could not be implemented due to particularities for a specific language, such as the rarity of homophones in Basque, which makes it difficult to construct an appropriate benchmark.

Word Containing Letter
Write a word that contains the letter "$letter$":
Write a word using the letter "$letter$":
Write a word containing the letter "$letter$":

Table 11: The templates of the *word containing letter* task. $letter$ is one of the possible letters for each alphabet in the target language. This number vary with the language.

Word Not Containing Letter
Write a word that doesn't contain the letter "$letter$":
Write a word without using the letter "$letter$":
Write a word not containing the letter "$letter$":

Table 12: The templates of the *word not containing letter* task. $letter$ is one of the possible letters for each alphabet in the target language. This number vary with the language.

Most Associated Word
Of the words "w_1", "w_2", "w_3", "w_4", and "w_5", what is the word most commonly associated with "$category$"?
What is the word most related to the word "$category$" from the words "w_1", "w_2", "w_3", "w_4", and "w_5"?
Of the following words, choose the word most commonly associated with the word "$category$" - "w_1", "w_2", "w_3", "w_4", "w_5":

Table 13: The templates of the *most associated word* task. $category$ and w_1 through w_5 are taken from set of categories manually curated for each language. One w_i is from $category$, and the other four words are not.

Least Associated Word
Of the words "w_1", "w_2", "w_3", "w_4", and "w_5", what is the word least associated with "$category$"?
What is the word least related to the word "$category$" from the words "w_1", "w_2", "w_3", "w_4", and "w_5"?
Of the following words, choose the word least associated with the word "$category$" - "w_1", "w_2", "w_3", "w_4", "w_5":

Table 14: The templates of the *least associated word* task. $category$ and w_1 through w_5 are taken from set of categories manually curated for each language. Four w_is are from $category$, and the remaining word is not.

Any Words From Category
Are any of the words "w_1", "w_2", "w_3", "w_4", and "w_5" types of $category$? Answer either "yes" or "no".
Do any of the words "w_1", "w_2", "w_3", "w_4", and "w_5" represent $category$? Answer either "yes" or "no".
Does the list ["w_1", "w_2", "w_3", "w_4", "w_5"] contain any $category$? Answer either "yes" or "no".

Table 15: The templates of the *any words from category* task. The number of w_is that belong to the category is either 0 or 1.

All Words From Category
Are all the words "w_1", "w_2", "w_3", "w_4", and "w_5" types of $category$? Answer either "yes" or "no".
Do the words "w_1", "w_2", "w_3", "w_4", and "w_5" all represent $category$? Answer either "yes" or "no".
Does the list ["w_1", "w_2", "w_3", "w_4", "w_5"] contain only $category$? Answer either "yes" or "no".

Table 16: The templates of the *all words from category* task. $category$ and w_1 through w_5 are taken from a set of categories manually curated for each language. The number of w_is that belong to the category is either 4 or 5. For the second and third templates, we use "items of clothing" instead of "clothes".

First Alphabetically
In an alphabetical order, which of the words "w_1" and "w_2" comes first?
In an alphabetical order, which word comes first, "w_1" or "w_2"?
Of the words "w_1" and "w_2", which word comes first alphabetically?

Table 17: The templates of the *first alphabetically* task. Both w_1 and w_2 are basic word (CEFR level A1 or A2, for each language).

More Letters
Which word has more letters, "w_1" or "w_2"?
Which word is longer, "w_1" or "w_2"?
Of the words "w_1" and "w_2" which one has more letters?

Table 18: The templates of the *more letters* task. Both w_1 and w_2 are basic word (CEFR level A1 or A2, for each language). w_1 and w_2 never have the same number of letters. In addition, to prevent any ambiguity, w_i never contains the same letter more than once. To illustrate, "horse" is a valid w_i, but "ball" or "present" are not.

Less Letters
Which word has fewer letters, "w_1" or "w_2"?
Which word is shorter, "w_1" or "w_2"?
Of the words "w_1" and "w_2" which one has fewer letters?

Table 19: The templates of the *less letters* task. Both w_1 and w_2 are basic word (CEFR level A1 or A2, for each language). w_1 and w_2 never have the same number of letters. In addition, to prevent any ambiguity, w_i never contains the same letter more than once. To illustrate, "horse" is a valid w_i, but "ball" or "present" are not.

Bigger Number
Which number is bigger, n_1 or n_2?
Of the numbers n_1 and n_2, which is bigger?
From the numbers n_1 and n_2, write the bigger number:

Table 20: The templates of the *bigger number* task. Both n_1 and n_2 are integers from the range [10, 999] (inclusive). n_1 and n_2 are never equal.

Smaller Number
Which number is smaller, n_1 or n_2?
Of the numbers n_1 and n_2, which is smaller?
From the numbers n_1 and n_2, write the smaller number:

Table 21: The templates of the *smaller number* task. Both n_1 and n_2 are integers from the range [10, 999] (inclusive). n_1 and n_2 are never equal.

Rhyming Word
Which word rhymes with the word "$query$", "w_1" or "w_2"?
Which is a rhyme of the word "$query$", "w_1" or "w_2"?
Of the words "w_1" and "w_2", which one rhymes with "$query$"?

Table 22: The templates of the *rhyming word* task. $query$, w_1, and w_2 are all basic words (CEFR level A1, A2, , for each language). One of w_1 and w_2 rhymes with $query$, and the other does not.

Homophones
Which word sounds like the word "$query$", "w_1" or "w_2"?
Of the two words "w_1" and "w_2", which one sounds more like "$query$"?
Which is a homophone of the word "$query$", "w_1" or "w_2"?

Table 23: The templates of the *homophones* task. The homophones were manually curated by native speaker annotators for each language. The authors further curated the data to avoid pronunciation ambiguities, the number of homophones vary a lot for each language, this tasks is not implemented for Basque. One of w_1 and w_2 is a homophone of $query$, and the other (the distractor) is not.

Word After In Sentence
In the sentence "$sentence$", which word comes right after the word "$word$"?
In the sentence "$sentence$", which word immediately succeeds the word "$word$"?
Which word comes right after "$word$" in the sentence "$sentence$"?

Table 24: The templates of the *word after in sentence* task. $sentence$ is a sentence taken from a language specific corpus in `CommonVoice Delta 15 validated` (Mozilla Foundation, 2025). $query$ is a word from $sentence$, and never its last word.

Word Before In Sentence
In the sentence "$sentence$", which word comes right before the word "$word$"?
In the sentence "$sentence$", which word immediately precedes the word "$word$"?
Which word comes right before "$word$" in the sentence "$sentence$"?

Table 25: The templates of the *word before in sentence* task. $sentence$ is a sentence taken from a language specific corpus in `CommonVoice Delta 15 validated` (Mozilla Foundation, 2025). $query$ is a word from $sentence$, and never its first word.

Sentence Starting With Word
Write a sentence that starts with the word "$word$":
Write a sentence whose first word is "$word$":
Write a sentence starting with "$word$":

Table 26: The templates of the *sentence starting with word* task. $word$ is a basic word (CEFR level A1 or A2, for each language) that the authors determined can start a sentence.

Sentence Ending With Word
Write a sentence that ends with the word "$word$":
Write a sentence whose last word is "$word$":
Write a sentence ending with "$word$":

Table 27: The templates of the *sentence ending with word* task. $word$ is a basic word (CEFR level A1 or A2, for each language) that the authors determined can end a sentence.

Word Starting With Letter
Write a word that starts with the letter "$letter$":
Write a word whose first letter is "$letter$":
Write a word starting with "$letter$":

Table 28: The templates of the *word starting with letter* task. $letter$ is one of the available letters for each language.

Word Ending With Letter
Write a word that ends with the letter “$letter$”:
Write a word whose last letter is “$letter$”:
Write a word ending with “$letter$”:

Table 29: The templates of the *word ending with letter* task. $letter$ is one of the available letters for each language.

First Word Of The Sentence
What is the first word of the sentence “$sentence$”?
In the sentence “$sentence$”, what is the first word?
Write the first word of the sentence “$sentence$”:

Table 30: The templates of the *first word of the sentence* task. $sentence$ is a sentence taken from a language specific corpus in `CommonVoice Delta 15 validated` (Mozilla Foundation, 2025), using the same filtering procedure as in the *word after in sentence* task.

Last Word Of The Sentence
What is the last word of the sentence “$sentence$”?
In the sentence “$sentence$”, what is the last word?
Write the last word of the sentence “$sentence$”:

Table 31: The templates of the *last word of the sentence* task. $sentence$ is a sentence taken from a language specific corpus in `CommonVoice Delta 15 validated` (Mozilla Foundation, 2025), using the same filtering procedure as in the *word after in sentence* task.

First Letter Of The Word
What is the first letter of the word “$word$”?
Which letter does the word “$word$” start with?
Write the first letter of the word “$word$”:

Table 32: The templates of the *first letter of the word* task. $word$ is a basic word (CEFR level A1 or A2, for each language).

Last Letter Of The Word
What is the last letter of the word “$word$”?
Which letter does the word “$word$” end with?
Write the last letter of the word “$word$”:

Table 33: The templates of the *last letter of the word* task. $word$ is a basic word (CEFR level A1 or A2, for each language).

Model	sentence_containing	sentence_not_containing	word_containing	word_not_containing	most_associated_word	least_associated_word	any_words_from_category	all_words_from_category	first_alphabetically	more_letters	less_letters	bigger_number	smaller_number	rhyming_word	homophones	word_after	word_before	starts_with_word	ends_with_word	starts_with_letter
Llama-3.2-1B	81.87	98.9	40.4	13.13	0.13	0.17	54.97	54.23	53.33	31.5	50.83	30.1	31.07	0.2	3.61	3.47	0.03	15.27	21.9	55.13
Qwen2.5-1.5B	79.33	87.73	74.75	81.82	1.13	1.23	55.6	59.3	24.43	39.37	47.33	45.43	5.43	0.2	2.22	9.9	2.83	7.07	8.47	71.79
Llama-3.2-3B	92.93	98.47	61.62	22.22	0.0	16.43	52.73	55.47	81.13	78.77	59.5	11.43	0.1	15.87	5.56	7.17	0.1	16.17	57.5	78.21
Qwen2.5-3B	84.5	91.5	80.81	94.95	0.8	4.17	49.2	63.2	56.6	64.5	72.1	56.67	53.03	6.9	3.06	17.77	3.93	15.13	27.2	80.77
Llama-3.1-8B	94.27	97.43	53.54	79.8	0.03	27.0	64.27	64.7	97.27	91.03	95.13	66.47	65.63	6.03	3.89	32.7	0.57	16.53	65.0	69.23
Qwen2.5-14B	92.13	97.13	92.93	95.96	0.13	19.77	53.23	74.67	85.5	78.1	94.37	66.67	48.93	18.93	0.0	21.3	3.97	16.6	77.3	84.62
phi-4	91.47	97.7	21.21	88.89	0.0	29.07	64.5	71.87	97.9	92.9	92.07	66.8	66.6	0.83	8.06	13.83	18.97	16.9	52.5	78.21
EuroLLM-1.7B	85.6	31.27	29.29	28.28	41.53	19.57	62.07	68.5	62.1	79.33	71.6	87.2	86.07	0.0	0.0	10.4	8.57	2.9	1.83	32.05
salamandra-2b	65.87	73.83	51.52	81.82	1.4	0.2	48.1	47.5	12.0	14.77	5.33	0.73	1.27	1.33	0.0	8.17	0.63	5.8	1.87	61.54
salamandra-7b	77.43	77.3	64.65	72.73	7.77	7.5	48.03	46.83	36.5	18.67	23.33	8.33	41.53	5.67	4.44	44.97	5.6	7.83	4.73	75.64
occiglot-7b-eu5	88.5	89.13	21.21	29.29	0.0	21.17	40.0	53.9	60.77	53.37	50.47	47.33	30.13	0.0	0.0	2.27	0.1	15.83	0.0	7.69
EuroLLM-9B	97.13	49.2	9.09	20.2	0.0	2.17	48.63	55.03	58.33	69.5	72.23	55.27	40.5	2.4	1.11	2.67	0.9	15.87	8.4	2.56
SmolLM2-360M	84.6	65.9	25.25	32.32	6.13	7.63	5.23	4.77	25.37	22.23	16.6	81.23	38.37	0.03	1.11	0.27	0.0	14.83	27.13	33.33
SmolLM2-1.7B	86.5	72.27	79.8	49.49	8.1	10.93	52.63	52.33	53.83	50.6	46.23	49.73	45.37	9.47	13.89	5.03	0.23	13.17	10.8	82.05
occiglot-7b-es-en	93.6	9.07	24.24	44.44	0.03	12.2	3.03	34.8	55.8	42.2	26.8	77.8	88.27	0.0	0.0	6.87	0.0	16.4	0.13	0.0
ANITA-8B	93.93	99.8	95.96	89.9	83.07	86.1	61.9	67.83	80.1	37.63	56.4	97.43	98.8	37.9	71.94	75.13	29.83	16.83	90.27	84.62
Latxa-Llama-3.1-8B	92.2	87.23	23.23	50.51	0.03	23.83	60.83	59.1	65.0	90.73	93.9	66.03	49.0	0.1	0.0	40.03	20.57	16.37	43.57	71.79

Table 34: Detailed accuracy results (%) for a subset of tasks in the Italian language

Model	ends_with_letter	first_word	last_word	first_letter	last_letter	first_alphabetically_far_first_letter	first_alphabetically_different_first_letter	first_alphabetically_consecutive_first_letter	first_alphabetically_same_first_letter	any_words_from_category_5_distractors	any_words_from_category_4_distractors	any_words_from_category_3_distractors	all_words_from_category_0_distractors	all_words_from_category_1_distractors	all_words_from_category_2_distractors	rhyming_word_orthographically_similar	more_letters_length_diff_3plus	more_letters_length_diff_1	less_letters_length_diff_3plus	less_letters_length_diff_1
Llama-3.2-1B	1.85	17.73	19.53	33.3	59.53	48.27	51.3	52.27	50.73	66.07	43.63	51.83	40.63	64.63	71.83	0.3	35.37	29.6	56.23	45.83
Qwen2.5-1.5B	9.26	6.37	1.27	60.87	28.6	24.37	26.33	27.5	25.47	92.33	23.63	53.3	83.43	34.63	56.13	0.07	47.6	36.23	54.47	44.2
Llama-3.2-3B	3.7	8.2	7.23	62.0	45.77	76.13	76.8	77.67	65.17	94.27	16.03	29.87	22.27	82.13	94.63	17.23	87.0	69.67	58.77	53.13
Qwen2.5-3B	24.07	30.3	27.57	87.2	33.43	56.33	58.73	54.6	52.83	99.33	2.6	11.53	38.7	86.1	97.83	6.63	82.93	57.77	83.33	61.8
Llama-3.1-8B	7.41	20.73	43.47	44.73	42.43	97.03	97.67	97.33	80.93	78.73	52.83	78.0	69.17	42.7	67.17	5.97	98.93	82.43	99.77	87.63
Qwen2.5-14B	16.67	9.0	2.8	38.37	31.83	87.0	88.57	82.8	66.97	96.87	14.33	34.77	62.33	87.43	97.97	20.23	85.9	70.2	99.2	86.7
phi-4	14.81	21.5	21.43	33.27	32.67	99.37	97.57	92.83	78.97	86.57	45.17	63.97	76.27	64.23	85.93	0.9	99.73	83.87	98.57	81.93
EuroLLM-1.7B	5.56	51.6	18.27	75.57	19.63	64.1	63.77	63.5	63.7	22.87	99.7	99.83	99.93	33.07	33.03	0.0	80.93	79.7	71.63	71.73
salamandra-2b	5.56	4.23	4.67	15.1	0.6	13.5	14.1	12.33	11.47	100.0	0.0	0.03	0.0	100.0	100.0	1.03	17.5	12.4	4.87	4.43
salamandra-7b	29.63	9.4	5.03	37.6	39.53	34.63	37.0	39.3	34.33	99.97	0.07	0.13	0.5	98.17	99.4	6.07	17.23	16.47	23.0	23.37
occiglot-7b-eu5	1.85	0.43	1.23	33.13	31.0	63.17	60.37	57.4	55.57	74.53	9.3	18.33	47.73	51.13	66.63	0.0	65.13	45.3	53.8	51.27
EuroLLM-9B	3.7	1.27	0.27	33.63	27.13	59.57	54.37	52.33	50.4	98.6	2.83	8.73	23.63	89.37	97.07	2.2	84.13	58.5	83.5	62.77
SmolLM2-360M	7.41	25.13	0.6	0.07	0.0	24.6	25.57	25.43	25.43	5.77	6.27	5.83	3.63	4.5	5.57	0.0	20.83	20.37	17.37	17.27
SmolLM2-1.7B	27.78	34.87	2.03	70.0	25.37	55.3	53.97	47.93	43.27	13.37	89.8	91.57	92.67	11.83	12.4	9.63	50.17	47.23	49.0	46.83
occiglot-7b-es-en	7.41	0.8	1.97	34.47	21.73	55.27	53.73	53.17	46.97	4.5	2.37	4.9	53.9	15.27	24.6	0.0	40.5	37.13	30.63	23.7
ANITA-8B	59.26	92.5	83.47	97.8	88.7	81.27	77.5	72.07	68.83	94.87	30.73	52.6	76.6	48.4	73.23	40.67	43.93	27.7	66.67	44.47
Latxa-Llama-3.1-8B	9.26	38.1	39.0	40.6	10.57	65.87	65.03	64.6	54.1	74.23	49.3	74.17	86.83	22.9	44.2	0.07	98.27	81.8	99.77	84.0

Table 35: Detailed accuracy results (%) for a subset of tasks in the Italian language

Model	sentence_containing	sentence_not_containing	word_containing	word_not_containing	most_associated_word	least_associated_word	any_words_from_category	all_words_from_category	first_alphabetically	more_letters	less_letters	bigger_number	smaller_number	rhyming_word	rhyming_word	word_after	word_before	starts_with_word	ends_with_word	starts_with_letter
Llama-3.2-1B	80.47	96.7	92.59	33.33	0.33	2.4	54.17	51.87	55.57	48.7	59.47	53.53	63.37	39.0	32.55	58.07	10.3	16.67	26.03	51.28
Qwen2.5-1.5B	79.43	92.7	95.06	91.36	1.4	24.53	51.4	61.47	16.97	31.17	33.53	50.0	23.7	23.9	3.39	40.47	7.77	6.4	7.4	64.1
Llama-3.2-3B	91.37	96.0	88.89	71.6	15.73	42.63	50.63	56.37	82.23	88.37	85.97	58.57	66.13	45.53	37.89	61.73	22.07	16.1	55.8	55.13
Qwen2.5-3B	87.53	84.13	100.0	90.12	1.8	15.73	49.87	69.9	58.87	76.47	77.27	86.17	33.6	17.0	11.72	49.43	13.4	15.43	30.27	71.79
Llama-3.1-8B	93.87	96.47	81.48	87.65	0.1	58.47	60.07	58.3	63.07	86.9	90.4	64.8	42.33	71.2	48.44	70.43	38.5	17.23	72.23	60.26
Qwen2.5-14B	92.13	96.2	100.0	67.9	0.17	39.63	52.5	78.53	37.4	90.5	90.1	66.63	78.63	41.43	12.76	68.4	46.03	16.97	77.8	76.92
phi-4	90.7	95.77	46.91	87.65	7.27	87.1	56.37	73.9	85.83	93.23	91.57	100.0	97.63	10.87	6.47	78.37	49.9	17.5	53.73	53.85
EuroLLM-1.7B	95.33	12.23	54.32	24.69	23.7	11.97	55.2	52.8	36.27	59.4	51.63	58.67	56.9	0.0	0.0	25.63	10.33	0.6	1.87	43.59
salamandra-2b	74.83	52.97	56.79	72.84	1.2	0.53	50.67	48.4	15.4	18.07	17.17	0.07	0.1	0.5	0.3	3.47	1.4	5.87	3.2	34.62
salamandra-7b	82.87	59.97	59.26	56.79	1.87	7.47	50.6	57.87	13.4	34.7	16.03	39.4	44.27	1.43	0.0	10.77	5.33	6.9	3.6	64.1
occiglot-7b-eu5	89.6	17.47	3.7	51.85	0.3	27.53	50.93	57.3	7.27	38.27	47.17	93.03	70.87	0.0	0.0	49.27	8.97	16.9	0.03	32.05
EuroLLM-9B	95.63	67.33	6.17	29.63	0.07	47.53	50.13	61.83	57.93	67.67	64.87	66.97	80.1	46.77	2.56	50.8	14.07	16.73	7.63	0.0
SmolLM2-360M	87.77	42.03	56.79	32.1	7.7	8.67	63.2	62.17	14.3	31.77	31.0	66.33	56.03	8.03	2.34	1.97	2.0	14.03	24.3	32.05
SmolLM2-1.7B	90.2	53.57	58.02	28.4	18.77	12.9	65.73	64.0	45.67	54.43	47.57	82.17	61.0	36.9	23.13	15.77	5.67	13.63	15.8	82.05
occiglot-7b-es-en	94.63	14.43	11.11	41.98	0.3	45.53	52.2	56.33	45.33	45.17	56.97	89.37	88.23	0.0	0.0	58.9	15.77	16.0	0.1	2.56
ANITA-8B	94.47	99.4	100.0	97.53	78.77	88.7	56.27	62.93	76.77	48.73	64.8	94.93	91.47	42.57	63.28	73.57	29.83	17.43	92.53	79.49
Latxa-Llama-3.1-8B	89.93	92.13	66.67	59.26	0.77	37.4	56.97	53.73	44.57	86.07	81.27	65.7	20.73	66.73	46.53	90.33	33.33	16.07	55.4	70.51

Table 36: Detailed accuracy results (%) for a subset of tasks in the Spanish language

Model	ends_with_letter	first_word	last_word	first_letter	last_letter	first_alphabetically_far_first_letter	first_alphabetically_different_first_letter	first_alphabetically_consecutive_first_letter	first_alphabetically_same_first_letter	any_words_from_category_5_distractors	any_words_from_category_4_distractors	any_words_from_category_3_distractors	all_words_from_category_0_distractors	all_words_from_category_1_distractors	all_words_from_category_2_distractors	rhyming_word_orthographically_similar	more_letters_length_diff_3plus	more_letters_length_diff_1	less_letters_length_diff_3plus	less_letters_length_diff_1
EuroLLM-1.7B	9.8	57.31	36.06	97.4	39.73	34.03	37.6	36.93	37.0	60.83	50.13	57.8	83.27	21.43	24.3	0.0	62.7	56.83	48.8	53.77
EuroLLM-9B	1.96	95.84	84.25	99.47	95.3	57.13	55.2	53.73	46.93	98.13	2.8	11.0	45.9	69.67	85.4	46.8	84.63	58.57	75.07	59.83
ANITA-8B	96.08	92.71	88.48	98.9	90.27	77.5	74.6	70.4	70.57	95.37	14.93	36.23	92.07	29.9	59.9	43.3	61.43	36.73	72.2	55.73
Latxa-Llama-3.1-8B	11.76	96.47	92.11	99.47	98.13	44.93	44.4	45.5	35.1	94.77	16.2	42.87	92.7	12.4	36.7	67.53	98.83	64.47	99.0	65.33
Llama-3.1-8B	17.65	99.77	95.74	99.5	98.6	65.17	64.0	63.53	51.17	96.4	20.73	39.5	88.17	23.3	49.9	70.97	99.4	71.13	99.97	80.27
Llama-3.2-1B	1.96	86.68	62.47	99.0	92.27	49.27	55.77	59.97	48.63	84.2	22.8	29.97	47.73	54.23	60.23	38.4	54.17	41.27	64.9	53.9
Llama-3.2-3B	21.57	93.41	84.82	99.5	96.47	79.43	76.73	80.93	64.63	99.47	2.1	7.7	18.4	79.37	90.93	44.43	99.33	75.63	94.8	72.4
Qwen2.5-14B	39.22	99.73	97.77	99.67	96.97	39.5	39.57	36.1	29.4	98.87	5.23	14.9	63.23	89.67	98.77	42.17	99.2	76.8	97.67	80.6
Qwen2.5-1.5B	0.0	76.92	62.77	89.3	33.47	18.23	18.27	15.93	17.23	99.27	3.83	15.3	63.67	55.5	74.43	23.0	34.9	30.23	36.03	32.43
Qwen2.5-3B	19.61	97.74	63.6	66.1	52.73	62.13	61.93	54.93	53.9	100.0	0.3	2.17	34.97	89.5	99.07	15.77	89.23	63.87	86.43	67.87
phi-4	13.73	99.93	95.74	99.77	99.07	89.57	88.7	86.9	64.8	96.67	15.23	29.27	74.6	63.4	87.03	9.6	100.0	86.07	97.27	85.8
salamandra-2b	1.96	11.66	6.16	42.53	34.47	12.53	14.4	16.33	14.1	2.1	98.43	99.17	98.53	0.2	0.27	0.37	15.93	17.17	17.63	17.27
salamandra-7b	29.41	19.45	26.07	56.17	23.93	12.5	14.57	15.17	17.57	97.13	6.43	17.43	46.63	58.0	76.0	1.63	41.43	32.07	17.73	15.23
occiglot-7b-eu5	3.92	87.78	65.37	51.47	67.23	7.97	7.37	7.3	6.37	98.03	2.93	12.37	33.9	62.57	77.33	0.0	43.87	34.63	50.3	47.17
SmolLM2-360M	9.8	16.95	11.39	46.23	22.07	12.67	13.5	15.07	13.1	35.57	89.43	89.57	79.5	32.9	32.83	8.1	31.93	31.37	27.7	30.63
SmolLM2-1.7B	64.71	61.97	12.95	74.07	27.03	42.73	45.97	44.23	40.83	53.1	76.33	86.9	96.5	31.93	33.8	36.03	58.13	48.37	47.03	50.27
occiglot-7b-es-en	1.96	88.28	85.15	88.77	86.53	43.17	43.7	42.7	38.03	98.47	5.37	18.77	41.23	57.07	73.87	0.0	55.67	38.43	67.2	47.27

Table 37: Detailed accuracy results (%) for a subset of tasks in the Spanish language

Model	sentence_containing	sentence_not_containing	word_containing	word_not_containing	most_associated_word	least_associated_word	any_words_from_category	all_words_from_category	first_alphabetically	more_letters	less_letters	bigger_number	smaller_number	rhyming_word	rhyming_word	word_after	word_before	starts_with_word	ends_with_word	starts_with_letter
EuroLLM-1.7B	88.97	23.37	46.91	0.0	29.8	10.9	13.17	27.13	41.9	58.43	41.73	67.9	62.53	0.0	0.0	2.4	2.43	0.7	0.43	25.64
EuroLLM-9B	89.77	56.13	11.11	0.0	0.57	33.13	52.77	59.43	32.27	63.4	57.83	96.5	85.1	17.17	0.38	22.1	7.93	16.5	10.03	0.0
ANITA-8B	88.5	99.27	98.77	90.12	40.6	69.8	60.23	59.2	58.4	43.67	43.5	89.53	96.93	33.67	48.48	54.73	21.4	17.0	87.73	62.82
Latxa-Llama-3.1-8B	84.6	90.7	49.38	29.63	1.07	21.1	52.07	50.57	7.5	88.4	83.33	99.53	97.2	37.6	1.01	5.6	1.17	13.3	28.7	67.95
Llama-3.1-8B	88.97	97.9	79.01	7.41	14.17	10.0	58.07	51.73	32.67	42.07	29.9	99.8	99.47	20.67	6.44	2.63	1.0	16.0	57.03	65.38
Llama-3.2-1B	83.93	82.13	35.8	2.47	0.07	0.0	80.4	88.53	6.17	23.5	20.63	83.1	83.33	6.17	0.88	1.23	1.33	14.3	8.5	43.59
Llama-3.2-3B	85.17	97.7	77.78	71.6	3.6	3.07	57.2	54.57	44.93	6.4	1.0	99.27	99.2	7.73	0.0	1.4	0.9	14.27	32.6	64.1
Qwen2.5-14B	91.23	95.53	96.3	62.96	0.37	17.5	58.43	67.53	27.0	88.1	58.53	66.83	66.67	4.2	0.0	9.93	2.77	16.03	68.57	71.79
Qwen2.5-1.5B	71.23	94.53	98.77	82.72	0.43	0.37	56.9	56.97	25.63	10.47	20.1	64.1	48.57	0.03	0.51	0.7	0.77	7.37	7.1	38.46
Qwen2.5-3B	77.67	88.03	98.77	71.6	0.1	1.3	52.17	54.4	22.5	32.8	33.0	42.1	22.07	3.07	4.29	2.83	1.27	15.33	11.83	66.67
SmolLM2-1.7B	82.87	40.9	59.26	45.68	14.6	12.1	45.2	51.1	21.23	36.37	33.43	75.13	48.3	29.6	26.64	3.73	2.9	7.63	14.6	56.41
SmolLM2-360M	89.5	18.83	50.62	50.62	0.03	1.77	31.13	36.17	23.83	4.17	6.7	59.83	45.23	1.93	5.56	0.07	0.1	14.53	24.63	12.82
occiglot-7b-es-en	86.93	27.2	0.0	0.0	0.0	0.07	0.1	0.5	1.33	0.03	0.0	41.6	2.93	0.0	0.0	0.17	0.0	14.83	0.23	0.0
occiglot-7b-eu5	82.83	26.9	0.0	0.0	0.0	5.8	15.53	27.37	0.13	3.1	10.83	46.73	17.63	0.0	0.0	3.97	3.3	13.33	0.17	0.0
phi-4	88.87	99.27	64.2	83.95	20.1	6.87	61.3	67.77	45.47	95.23	44.83	100.0	99.5	0.1	0.0	46.9	15.0	17.03	38.9	66.67
salamandra-2b	65.1	68.3	72.84	69.14	4.5	3.1	48.6	49.0	6.63	6.7	7.1	1.03	0.0	5.17	3.79	3.9	0.77	5.57	1.93	41.03
salamandra-7b	76.1	61.4	69.14	62.96	5.7	13.03	63.9	54.7	25.63	19.67	27.43	55.5	50.93	16.33	8.71	13.27	4.53	7.33	5.87	48.72

Table 38: Detailed accuracy results (%) for a subset of tasks in the Galician language

Model	ends_with_letter	first_word	last_word	first_letter	last_letter	first_alphabetically_far_first_letter	first_alphabetically_different_first_letter	first_alphabetically_consecutive_first_letter	first_alphabetically_same_first_letter	any_words_from_category_5_distractors	any_words_from_category_4_distractors	any_words_from_category_3_distractors	all_words_from_category_0_distractors	all_words_from_category_1_distractors	all_words_from_category_2_distractors	rhyming_word_orthographically_similar	more_letters_length_diff_3plus	more_letters_length_diff_1	less_letters_length_diff_3plus	less_letters_length_diff_1
EuroLLM-1.7B	23.33	62.57	11.27	39.9	24.23	40.6	41.6	39.47	38.33	15.9	10.9	12.8	56.5	0.1	0.07	0.0	58.23	55.17	40.5	42.67
EuroLLM-9B	3.33	1.8	0.1	0.4	0.0	30.93	32.53	32.0	30.43	97.3	7.93	20.6	21.0	93.3	98.87	18.13	76.13	50.97	65.33	55.0
ANITA-8B	76.67	82.63	66.6	71.57	33.47	64.2	57.13	55.33	56.23	93.63	27.63	55.23	80.43	37.83	55.27	34.87	52.93	38.0	53.03	34.67
Latxa-Llama-3.1-8B	10.0	13.77	3.93	4.93	0.0	9.67	8.17	8.17	9.0	95.33	8.6	20.9	27.97	66.8	80.5	39.73	97.57	71.67	93.67	71.4
Llama-3.1-8B	23.33	26.17	2.4	16.3	0.23	34.43	33.37	32.4	25.63	82.03	35.63	58.43	48.73	56.0	69.13	21.0	50.9	34.47	32.63	26.3
Llama-3.2-1B	13.33	2.13	0.07	19.83	0.0	5.77	5.93	6.1	5.43	79.3	79.97	76.1	79.17	93.0	94.07	5.4	23.47	21.47	20.37	18.37
Llama-3.2-3B	33.33	13.43	8.77	17.57	2.83	44.8	42.53	46.57	39.97	93.37	21.37	13.33	12.43	89.73	95.63	7.9	6.73	3.63	1.53	0.5
Qwen2.5-14B	43.33	2.17	1.27	12.2	0.37	30.6	27.4	24.63	22.83	96.73	19.87	44.97	49.27	90.43	96.7	4.63	95.77	75.47	62.57	51.63
Qwen2.5-1.5B	23.33	13.27	4.37	32.9	3.97	24.5	24.4	27.37	25.6	94.9	17.13	37.53	79.27	43.27	60.4	0.0	11.1	8.2	18.7	19.53
Qwen2.5-3B	46.67	29.47	14.73	52.37	21.1	25.37	23.2	21.53	20.3	99.23	4.7	12.73	27.8	73.9	86.47	3.13	35.47	25.53	38.27	31.23
SmolLM2-1.7B	30.0	32.67	2.2	44.93	4.63	18.27	19.67	21.53	21.1	88.1	0.47	0.93	0.0	100.0	100.0	29.93	40.4	34.8	35.3	32.93
SmolLM2-360M	10.0	0.47	11.23	6.7	0.3	24.1	23.63	25.43	23.9	0.0	62.2	61.5	75.77	0.03	0.0	1.87	4.43	3.9	6.27	6.3
occiglot-7b-es-en	0.0	0.1	0.0	0.0	0.0	1.23	0.7	0.8	1.1	0.03	0.13	0.17	0.4	0.13	0.17	0.0	0.03	0.0	0.0	0.0
occiglot-7b-eu5	0.0	1.67	0.1	0.0	0.0	0.17	0.03	0.13	0.03	25.27	5.0	8.3	30.13	19.77	20.83	0.0	2.8	2.57	10.93	10.03
phi-4	13.33	28.07	28.97	1.77	0.0	47.07	45.83	43.5	34.47	88.63	37.3	62.97	70.07	66.07	85.4	0.03	98.83	87.43	47.27	42.27
salamandra-2b	10.0	9.73	3.4	8.03	0.0	5.67	4.63	5.9	3.43	2.07	96.9	97.73	99.9	0.2	0.1	4.67	5.93	5.97	7.1	8.07
salamandra-7b	46.67	7.13	7.97	28.3	0.6	21.37	21.9	22.47	22.33	88.47	43.9	70.27	65.8	49.67	58.5	16.0	20.8	18.73	27.9	24.03

Table 39: Detailed accuracy results (%) for a subset of tasks in the Galician language

Model	sentence_containing	sentence_not_containing	word_containing	word_not_containing	most_associated_word	least_associated_word	any_words_from_category	all_words_from_category	first_alphabetically	more_letters	less_letters	bigger_number	smaller_number	rhyming_word	rhyming_word	word_after	word_before	starts_with_word	ends_with_word	starts_with_letter
EuroLLM-1.7B	93.17	15.17	14.81	0.0	2.07	1.97	51.1	33.97	32.73	26.43	25.67	46.3	35.2	1.0	0.68	3.97	3.47	4.0	5.7	5.13
EuroLLM-9B	92.17	46.3	6.17	0.0	0.73	42.83	54.7	61.43	60.4	75.03	65.03	56.33	30.03	15.7	3.38	48.53	13.37	15.7	8.97	2.56
ANITA-8B	90.6	98.03	96.3	82.72	68.57	81.8	64.83	76.8	71.3	43.0	57.2	91.87	97.67	38.07	56.29	50.47	19.8	17.07	88.27	64.1
Latxa-Llama-3.1-8B	83.63	64.9	4.94	3.7	10.93	36.93	45.73	54.93	2.63	83.43	82.83	50.7	35.33	53.07	27.77	64.53	24.5	14.1	48.53	29.49
Llama-3.1-8B	88.97	92.77	14.81	3.7	29.0	33.07	27.67	51.73	85.03	84.73	87.3	57.0	56.67	69.9	7.15	86.33	28.87	15.63	56.83	17.95
Llama-3.2-1B	76.23	89.07	82.72	53.09	1.77	20.0	50.23	51.63	42.5	25.87	5.33	40.73	41.97	8.17	13.86	16.17	4.5	10.87	23.2	52.56
Llama-3.2-3B	83.27	96.2	67.9	53.09	25.5	24.27	50.3	50.8	75.83	43.6	62.7	88.83	72.17	29.47	22.45	50.53	15.83	11.37	34.5	51.28
Qwen2.5-14B	92.93	95.3	97.53	74.07	5.43	27.67	51.3	70.17	54.47	86.7	93.73	66.23	59.2	39.5	0.0	12.87	10.43	15.37	70.7	83.33
Qwen2.5-1.5B	66.33	95.9	96.3	60.49	0.23	2.53	60.4	53.33	13.47	37.37	36.77	64.53	80.13	18.13	2.48	6.3	1.23	7.93	9.67	60.26
Qwen2.5-3B	67.0	79.2	96.3	95.06	10.9	7.83	49.47	58.93	61.03	57.77	62.53	86.77	78.83	0.43	16.63	25.33	2.5	14.17	21.57	73.08
SmolLM2-1.7B	82.97	23.63	76.54	25.93	20.83	16.43	51.1	49.5	54.17	43.17	50.5	62.97	64.57	24.07	15.05	3.77	3.77	11.37	27.3	56.41
SmolLM2-360M	85.17	12.57	16.05	0.0	0.63	0.8	0.3	0.37	12.9	0.33	0.17	0.17	2.7	0.13	0.47	0.0	0.03	13.53	50.77	7.69
occiglot-7b-es-en	83.67	49.87	0.0	0.0	7.8	31.3	22.8	24.23	33.33	51.3	45.63	11.3	8.67	0.0	0.0	25.4	7.47	14.73	0.13	0.0
occiglot-7b-eu5	82.03	49.97	0.0	0.0	2.7	37.37	57.83	53.57	2.9	53.53	32.33	0.07	0.07	0.0	0.0	37.03	8.8	14.8	0.2	3.85
phi-4	90.67	98.8	61.73	85.19	10.77	29.33	59.37	67.63	97.03	60.67	56.63	99.67	97.03	8.13	0.54	84.57	38.77	17.0	46.8	60.26
salamandra-2b	69.57	65.27	60.49	77.78	2.7	0.57	57.57	53.03	16.7	13.37	23.1	2.53	0.57	2.93	1.9	7.93	2.67	5.1	4.23	42.31
salamandra-7b	79.17	58.93	70.37	66.67	10.3	12.2	52.43	57.97	16.3	46.6	43.37	28.5	38.63	10.37	0.18	41.53	7.37	7.93	5.27	61.54

Table 40: Detailed accuracy results (%) for a subset of tasks in the Catalan language

Model	ends_with_letter	first_word	last_word	first_letter	last_letter	first_alphabetically_far_first_letter	first_alphabetically_different_first_letter	first_alphabetically_consecutive_first_letter	first_alphabetically_same_first_letter	any_words_from_category_5_distractors	any_words_from_category_4_distractors	any_words_from_category_3_distractors	all_words_from_category_0_distractors	all_words_from_category_1_distractors	all_words_from_category_2_distractors	rhyming_word_orthographically_similar	more_letters_length_diff_3plus	more_letters_length_diff_1	less_letters_length_diff_3plus	less_letters_length_diff_1
EuroLLM-1.7B	1.67	32.03	14.75	68.13	14.9	33.43	34.3	34.3	30.4	18.03	83.77	87.03	65.73	2.17	3.9	1.03	26.37	22.67	19.67	24.77
EuroLLM-9B	6.67	85.85	55.91	99.2	68.9	60.5	59.97	57.6	51.2	93.17	17.93	42.6	38.8	79.5	92.0	15.9	84.6	57.77	76.47	61.27
ANITA-8B	91.67	87.35	71.56	96.7	71.77	74.23	70.13	65.13	61.57	83.43	45.03	74.03	85.97	59.57	75.77	40.17	53.73	34.53	67.13	42.3
Latxa-Llama-3.1-8B	3.33	95.04	16.58	98.93	32.47	4.3	4.27	3.77	3.63	78.6	12.9	28.63	60.9	43.8	57.8	55.37	92.3	64.2	92.97	69.2
Llama-3.1-8B	3.33	95.74	21.71	94.87	40.43	87.6	85.13	84.1	67.17	41.63	12.57	26.03	66.83	32.93	44.13	70.7	97.93	59.03	98.5	73.57
Llama-3.2-1B	1.67	82.08	1.2	96.57	1.33	42.8	42.13	43.83	40.73	73.8	24.73	29.27	32.77	82.77	85.53	7.73	28.83	24.43	14.53	13.53
Llama-3.2-3B	8.33	84.08	27.77	99.33	72.43	79.37	72.53	73.1	62.33	99.53	2.7	10.23	1.17	99.77	100.0	30.8	51.0	35.17	69.6	53.6
Qwen2.5-14B	55.0	91.77	39.03	99.07	38.13	53.8	53.33	51.13	41.5	98.77	5.43	20.43	64.87	80.23	96.37	44.67	90.7	77.5	99.63	86.57
Qwen2.5-1.5B	13.33	62.74	26.11	91.7	43.13	13.0	11.13	10.4	11.7	83.37	33.3	50.43	61.27	54.33	64.33	18.53	34.8	32.9	39.2	36.67
Qwen2.5-3B	43.33	38.39	31.8	90.0	55.5	68.43	60.73	53.63	51.3	99.9	0.8	5.67	18.83	92.57	98.43	0.27	67.87	51.37	74.13	57.93
SmolLM2-1.7B	35.0	47.75	14.39	75.6	31.1	52.97	52.27	53.1	50.4	0.0	100.0	100.0	100.0	0.0	0.0	25.67	47.03	44.17	47.23	50.4
SmolLM2-360M	0.0	0.0	1.5	1.33	0.4	15.07	15.0	15.07	14.3	0.1	0.5	0.4	0.2	0.3	0.37	0.1	0.23	0.37	0.37	0.17
occiglot-7b-es-en	0.0	0.5	0.0	11.47	0.87	30.8	32.1	34.17	22.2	20.7	26.43	27.8	40.7	15.9	23.57	0.0	53.43	40.63	49.03	40.9
occiglot-7b-eu5	8.33	0.07	0.37	2.87	2.87	4.07	3.5	3.67	2.47	84.07	28.43	43.07	58.47	50.63	60.37	0.0	56.4	43.3	43.1	34.1
phi-4	5.0	91.77	48.32	99.67	92.83	99.2	97.77	95.7	69.7	84.17	35.03	64.5	76.5	54.0	79.63	8.43	66.13	50.4	63.0	48.67
salamandra-2b	1.67	9.86	7.56	32.37	6.7	16.63	15.8	15.97	16.5	59.13	54.37	63.9	55.7	47.77	53.2	3.3	14.6	13.37	21.23	24.13
salamandra-7b	16.67	63.7	12.45	60.3	13.4	16.0	15.9	14.8	15.8	98.6	7.53	22.9	29.8	79.1	89.97	12.0	51.3	39.3	45.23	39.27

Table 41: Detailed accuracy results (%) for a subset of tasks in the Catalan language

Model	sentence_containing	sentence_not_containing	word_containing	word_not_containing	most_associated_word	least_associated_word	any_words_from_category	all_words_from_category	first_alphabetically	more_letters	less_letters	bigger_number	smaller_number	rhyming_word	word_after	word_before	starts_with_word	ends_with_word	starts_with_letter
EuroLLM-1.7B	98.9	1.77	0.0	7.41	0.03	0.03	0.0	0.0	0.03	0.1	0.23	0.0	0.0	0.0	0.0	0.0	14.67	93.53	2.56
EuroLLM-9B	96.7	1.43	1.23	4.94	0.67	0.57	7.73	8.63	3.07	10.43	16.6	3.97	0.0	0.37	0.43	0.83	15.17	79.9	3.85
ANITA-8B	65.77	87.0	85.19	80.25	26.73	34.63	36.53	52.27	42.07	13.37	5.93	75.33	73.4	14.07	13.33	12.5	13.47	68.47	70.51
Latxa-Llama-3.1-8B	56.57	95.43	74.07	30.86	68.43	69.33	63.17	57.03	68.33	88.2	86.27	97.8	91.07	46.17	35.6	24.73	14.17	44.07	73.08
Llama-3.1-8B	60.63	86.73	33.33	4.94	19.37	7.33	46.27	54.8	46.2	85.4	79.03	87.6	82.83	41.73	12.17	17.3	15.3	61.7	21.79
Llama-3.2-1B	90.67	5.73	8.64	2.47	0.7	1.07	16.23	15.07	2.43	2.1	2.9	1.13	4.63	0.0	0.03	0.13	14.67	55.3	7.69
Llama-3.2-3B	70.93	67.13	80.25	74.07	15.47	12.6	49.3	51.7	33.93	47.77	53.27	62.8	64.6	1.23	9.93	1.13	14.9	42.37	69.23
Qwen2.5-14B	80.63	8.0	43.21	3.7	19.17	6.1	34.3	42.03	74.03	68.97	58.17	37.37	18.7	0.0	4.17	3.0	15.67	83.13	35.9
Qwen2.5-1.5B	72.73	17.83	9.88	0.0	0.2	0.23	42.73	43.2	0.23	11.93	13.37	6.93	0.23	1.77	0.0	0.0	11.43	34.73	7.69
Qwen2.5-3B	95.27	2.97	17.28	3.7	1.63	1.37	44.27	41.3	10.37	16.5	10.63	23.7	22.2	0.0	0.8	1.07	15.37	89.37	25.64
SmolLM2-1.7B	71.9	27.6	0.0	0.0	0.03	0.03	0.17	0.07	0.0	0.03	0.03	0.0	0.0	0.1	0.0	0.0	12.5	29.87	3.85
SmolLM2-360M	35.07	67.27	1.23	0.0	0.03	0.07	0.0	0.0	0.07	0.57	0.23	0.0	0.0	0.1	0.0	0.13	5.07	12.13	1.28
occiglot-7b-es-en	95.6	2.8	1.23	0.0	0.1	0.13	0.0	3.2	3.0	8.27	13.3	0.0	0.0	0.0	0.0	0.03	15.4	55.33	1.28
occiglot-7b-eu5	94.03	14.1	1.23	0.0	0.33	0.37	0.03	2.2	20.2	15.97	15.27	0.2	0.3	0.0	0.0	0.0	15.07	56.07	0.0
phi-4	70.3	73.73	8.64	12.35	39.13	38.7	55.6	57.67	89.93	56.37	48.93	64.17	59.43	0.0	3.1	3.23	15.17	46.1	29.49
salamandra-2b	25.07	78.5	81.48	2.47	3.37	2.63	46.57	48.3	3.4	3.1	14.2	12.03	1.77	5.8	0.97	0.4	4.2	7.2	56.41
salamandra-7b	20.6	93.1	61.73	54.32	12.2	11.83	55.77	50.7	5.93	10.6	14.3	35.47	27.8	3.97	2.77	1.97	4.97	11.77	55.13

Table 42: Detailed accuracy results (%) for a subset of tasks in the Basque language

Model	ends_with_letter	first_word	last_word	first_letter	last_letter	first_alphabetically_far_first_letter	first_alphabetically_different_first_letter	first_alphabetically_consecutive_first_letter	first_alphabetically_same_first_letter	any_words_from_category_5_distractors	any_words_from_category_4_distractors	any_words_from_category_3_distractors	all_words_from_category_0_distractors	all_words_from_category_1_distractors	all_words_from_category_2_distractors	rhyming_word_orthographically_similar	more_letters_length_diff_3plus	more_letters_length_diff_1	less_letters_length_diff_3plus	less_letters_length_diff_1
EuroLLM-1.7B	2.56	2.2	0.17	0.0	0.07	0.17	0.17	0.1	0.07	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.07	0.07	0.2	0.03
EuroLLM-9B	0.0	21.1	18.8	12.5	22.87	2.57	1.3	1.3	1.43	0.47	14.2	16.9	13.77	3.63	3.4	0.13	12.77	9.77	14.0	13.9
ANITA-8B	51.28	65.67	29.63	69.8	41.27	48.57	43.57	40.57	35.97	27.33	42.1	55.6	89.93	10.17	15.6	16.2	17.6	12.97	6.63	5.97
Latxa-Llama-3.1-8B	10.26	95.63	32.9	97.37	77.93	76.4	73.93	75.4	67.27	77.1	52.23	76.83	79.73	24.3	46.07	47.23	95.67	87.07	92.53	82.63
Llama-3.1-8B	2.56	77.33	30.3	65.27	42.27	51.4	48.2	46.5	42.83	59.97	32.7	46.87	72.8	27.0	43.07	40.67	90.67	84.03	81.73	77.63
Llama-3.2-1B	2.56	15.87	11.2	1.6	6.2	2.07	1.8	1.53	2.4	30.13	0.3	0.23	0.03	29.57	29.7	0.0	2.37	2.6	3.47	2.73
Llama-3.2-3B	30.77	70.93	17.57	56.83	17.67	35.63	35.7	39.63	35.27	98.57	1.43	1.53	19.53	86.03	88.3	0.87	54.2	45.9	54.67	52.9
Qwen2.5-14B	7.69	84.5	35.9	67.13	56.63	73.2	72.47	72.2	60.8	66.7	2.63	7.47	34.0	49.73	58.27	0.03	69.33	69.23	59.23	58.07
Qwen2.5-1.5B	2.56	2.17	1.33	0.1	0.17	0.07	0.13	0.33	0.47	83.4	2.7	2.33	2.17	81.93	84.43	1.73	10.0	9.8	10.8	12.87
Qwen2.5-3B	5.13	6.93	2.7	0.63	0.47	13.2	12.4	9.47	8.67	89.6	0.2	0.33	1.63	81.47	83.23	0.0	16.2	15.0	9.2	9.73
SmolLM2-1.7B	0.0	2.27	0.23	0.03	0.0	0.1	0.03	0.0	0.07	0.67	0.03	0.0	0.0	0.37	0.0	0.03	0.0	0.0	0.0	0.0
SmolLM2-360M	0.0	1.47	0.17	0.0	0.0	0.2	0.13	0.0	0.13	0.0	0.0	0.0	0.0	0.0	0.0	0.17	0.0	0.13	0.4	0.37
occiglot-7b-es-en	0.0	11.2	1.17	0.5	2.37	2.1	2.73	3.13	1.87	0.1	0.03	0.07	0.03	6.53	6.1	0.0	8.87	8.87	13.77	14.37
occiglot-7b-eu5	0.0	6.37	0.47	0.03	0.03	18.7	21.37	21.07	19.17	0.17	0.0	0.0	0.6	2.93	2.33	0.0	15.83	14.83	15.8	15.2
phi-4	5.13	67.53	38.5	62.13	86.8	93.63	92.0	86.97	68.6	91.83	21.17	32.87	44.33	58.63	70.47	0.0	60.03	59.33	45.0	47.27
salamandra-2b	7.69	1.83	0.67	15.4	1.37	3.73	2.83	4.07	3.27	17.33	78.57	77.2	95.43	3.97	5.67	6.57	4.5	4.13	11.9	13.5
salamandra-7b	7.69	5.57	2.0	54.1	14.27	7.5	7.57	6.9	5.5	91.33	24.23	40.43	78.03	26.97	30.07	4.47	12.2	10.3	12.2	12.1

Table 43: Detailed accuracy results (%) for a subset of tasks in the Basque language

Model	sentence_containing	sentence_not_containing	word_containing	word_not_containing	most_associated_word	least_associated_word	any_words_from_category	all_words_from_category	first_alphabetically	more_letters	less_letters	bigger_number	smaller_number	rhyming_word	homophones	word_after	word_before	starts_with_word	ends_with_word	starts_with_letter
EuroLLM-1.7B	85.7	31.33	3.33	2.22	0.0	0.0	52.77	49.1	0.3	43.07	43.8	48.53	0.0	0.13	0.0	0.0	0.0	1.27	2.47	10.26
EuroLLM-9B	82.9	63.8	0.0	0.0	0.0	0.0	49.53	61.53	0.1	56.73	49.0	7.0	14.37	0.0	0.0	18.57	0.17	8.33	5.07	0.0
ANITA-8B	86.83	99.73	34.44	96.67	9.07	4.67	52.13	62.93	62.6	36.6	60.4	22.07	11.97	0.07	1.51	59.53	13.73	16.23	59.1	75.64
Latxa-Llama-3.1-8B	79.0	84.97	0.0	0.0	0.0	0.0	53.6	61.2	10.03	39.83	43.1	95.93	0.07	0.0	1.45	22.7	7.27	6.0	18.57	0.0
Llama-3.1-8B	84.73	92.63	0.0	1.11	0.0	0.0	53.6	58.13	11.53	36.87	43.77	58.5	1.4	0.0	0.0	0.0	0.73	6.17	32.63	0.0
Llama-3.2-1B	71.3	89.13	0.0	3.33	0.0	0.0	52.37	51.73	0.33	17.47	24.3	31.27	35.17	0.0	0.0	15.63	0.33	2.57	5.03	5.13
Llama-3.2-3B	83.33	94.73	3.33	10.0	0.0	0.0	53.1	51.0	20.67	48.6	43.63	39.13	0.1	0.0	0.0	14.43	0.2	3.83	17.83	53.85
Qwen2.5-14B	84.17	94.17	27.78	85.56	0.0	0.0	55.5	67.07	2.07	76.47	84.33	66.67	0.0	9.1	0.0	3.53	0.37	15.97	50.63	83.33
Qwen2.5-1.5B	67.37	80.1	23.33	76.67	0.13	0.07	51.13	60.13	2.83	33.33	22.53	0.03	0.0	0.33	0.0	0.1	0.0	3.23	4.93	79.49
Qwen2.5-3B	76.63	87.6	46.67	66.67	0.0	0.0	48.93	63.23	12.4	28.77	67.9	48.37	1.4	0.0	0.02	0.63	0.0	12.67	9.07	79.49
SmolLM2-1.7B	84.67	48.53	0.0	1.11	2.1	1.5	47.03	50.57	0.2	33.8	40.03	27.53	25.3	0.0	0.23	0.37	0.03	9.0	4.33	28.21
SmolLM2-360M	94.37	12.67	4.44	0.0	0.8	1.2	15.03	10.4	0.0	11.87	11.37	11.1	11.9	0.03	0.04	0.1	0.0	11.97	4.73	17.95
occiglot-7b-es-en	83.53	17.43	0.0	0.0	0.0	0.0	4.27	38.87	0.93	13.6	16.77	42.73	0.57	0.0	0.0	0.0	0.0	12.97	0.3	1.28
occiglot-7b-eu5	79.3	61.0	0.0	1.11	0.0	0.0	48.2	59.97	0.0	54.3	51.93	24.93	0.0	0.0	0.0	0.43	0.17	6.93	0.1	1.28
phi-4	80.1	90.93	4.44	51.11	0.0	0.0	69.0	68.03	51.43	62.4	84.1	59.77	0.07	0.0	0.0	0.0	0.0	16.33	25.0	56.41
salamandra-2b	62.43	80.43	24.44	52.22	1.8	1.07	50.83	51.43	1.87	16.17	19.4	18.0	33.17	0.03	0.0	0.5	0.23	1.33	1.1	60.26
salamandra-7b	76.4	78.37	13.33	73.33	2.37	0.8	49.6	53.03	3.93	17.67	33.9	27.1	58.1	0.23	0.11	6.27	0.27	7.63	4.07	65.38

Table 44: Detailed accuracy results (%) for a subset of tasks in the German language

Model	ends_with_letter	first_word	last_word	first_letter	last_letter	first_alphabetically_far_first_letter	first_alphabetically_different_first_letter	first_alphabetically_consecutive_first_letter	first_alphabetically_same_first_letter	any_words_from_category_5_distractors	any_words_from_category_4_distractors	any_words_from_category_3_distractors	all_words_from_category_0_distractors	all_words_from_category_1_distractors	all_words_from_category_2_distractors	rhyming_word_orthographically_similar	more_letters_length_diff_3plus	more_letters_length_diff_1	less_letters_length_diff_3plus	less_letters_length_diff_1
EuroLLM-1.7B	0.0	3.13	0.33	0.47	3.8	0.13	0.27	0.17	0.1	24.9	77.4	80.13	99.63	0.8	1.4	0.0	45.07	40.27	39.7	44.87
EuroLLM-9B	0.0	0.8	0.0	0.0	0.0	0.0	0.03	0.0	0.03	96.73	4.6	9.63	25.47	90.57	97.53	0.0	66.5	46.53	55.33	49.23
ANITA-8B	54.17	47.32	41.66	12.2	42.93	62.77	59.47	56.23	43.43	93.87	11.0	30.73	65.07	51.4	76.2	0.07	37.67	31.1	67.03	47.37
Latxa-Llama-3.1-8B	0.0	0.03	0.13	0.17	1.97	10.3	8.57	6.97	11.77	66.4	39.1	54.97	68.13	45.27	63.53	0.0	48.43	30.4	49.3	35.9
Llama-3.1-8B	0.0	0.87	0.23	0.0	0.0	12.3	9.63	9.8	11.5	79.67	26.03	54.03	57.2	50.5	64.67	0.0	44.17	29.0	52.7	36.1
Llama-3.2-1B	0.0	0.57	0.0	0.0	0.0	0.57	0.4	0.3	0.3	50.9	54.97	58.63	28.93	72.63	73.7	0.0	18.53	17.77	22.97	22.33
Llama-3.2-3B	20.83	11.02	12.32	11.43	21.87	20.53	21.4	22.97	18.0	82.9	23.03	38.5	33.83	67.93	74.73	0.0	52.13	43.93	51.97	38.73
Qwen2.5-14B	31.94	1.63	0.37	5.93	0.0	2.07	1.6	1.73	1.7	97.2	15.17	42.6	35.63	95.13	99.6	0.0	90.97	60.63	91.23	64.0
Qwen2.5-1.5B	5.56	7.66	0.27	11.43	3.57	3.4	3.1	2.53	1.93	94.7	8.8	18.87	38.87	78.43	88.43	0.47	37.77	30.93	25.67	22.5
Qwen2.5-3B	16.67	31.54	15.32	14.5	19.13	12.03	10.47	10.2	7.63	99.87	1.37	4.8	27.5	92.57	98.33	0.0	34.9	18.27	73.63	59.63
SmolLM2-1.7B	9.72	3.13	1.53	9.63	0.2	0.23	0.13	0.17	0.23	97.0	0.13	0.07	11.27	94.27	94.33	0.13	31.07	30.13	40.53	41.8
SmolLM2-360M	1.39	0.07	1.07	0.0	0.0	0.0	0.0	0.0	0.07	0.3	29.13	29.13	13.1	0.07	0.13	0.0	11.23	11.57	10.13	10.33
occiglot-7b-es-en	0.0	0.23	1.1	0.0	0.63	0.93	0.77	0.73	0.57	8.2	0.2	0.33	13.43	61.67	62.77	0.0	12.07	8.8	17.53	13.1
occiglot-7b-eu5	1.39	0.27	0.03	0.0	0.0	0.0	0.0	0.0	0.0	91.7	5.47	14.9	36.93	71.6	85.8	0.0	63.77	53.2	50.57	49.2
phi-4	0.0	0.77	10.86	0.0	0.0	52.23	53.37	52.83	25.47	85.4	52.17	76.07	56.13	76.47	90.03	0.0	75.7	45.2	88.33	67.9
salamandra-2b	2.78	7.99	2.66	3.4	0.2	1.43	1.7	2.23	1.2	62.2	39.77	40.53	39.97	60.63	61.73	0.0	12.07	15.07	15.97	18.47
salamandra-7b	12.5	5.19	5.43	12.67	6.3	4.2	3.97	3.1	1.67	97.47	3.33	8.83	6.6	95.17	97.3	0.23	19.87	13.87	33.43	32.07

Table 45: Detailed accuracy results (%) for a subset of tasks in the German language

task	sentence_containing	sentence_not_containing	word_containing	word_not_containing	most_associated_word	least_associated_word	any_words_from_category	all_words_from_category	first_alphabetically	more_letters	less_letters	bigger_number	smaller_number	rhyming_word	homophones	word_after	word_before	starts_with_word	ends_with_word	starts_with_letter
EuroLLM-1.7B	74.04	0.2	15.38	0.0	54.93	57.8	96.5	96.17	91.47	89.97	91.73	37.73	41.0	98.63	86.36	58.87	47.53	76.5	73.43	0.0
EuroLLM-9B	62.7	0.03	5.13	2.56	72.77	70.0	99.97	99.13	99.63	99.63	98.83	99.83	99.83	82.93	75.98	54.3	32.57	31.13	65.07	0.0
ANITA-8B	23.79	85.4	52.56	78.21	24.5	19.23	94.37	63.97	50.97	43.03	58.93	79.73	85.97	48.13	26.31	22.03	17.93	31.03	25.33	0.0
Latxa-Llama-3.1-8B	34.97	1.13	51.28	3.85	68.37	67.27	96.77	78.27	99.07	97.47	95.6	97.7	87.7	98.3	73.65	50.4	38.3	23.17	24.87	0.0
Llama-3.1-8B	35.3	0.13	60.26	14.1	77.0	78.0	99.23	84.17	99.33	99.27	93.17	80.8	56.2	99.9	67.93	61.7	65.5	26.9	27.33	0.0
Llama-3.2-1B	82.35	0.23	21.79	0.0	44.77	29.17	95.13	96.43	96.63	92.57	86.2	69.87	78.07	96.27	76.55	32.5	30.6	48.47	70.4	0.0
Llama-3.2-3B	23.16	0.67	87.18	3.85	70.87	69.3	99.93	95.5	94.2	82.1	81.2	87.63	80.17	98.4	79.08	44.47	42.4	18.27	20.5	0.0
Qwen2.5-14B	23.49	0.9	15.38	0.0	86.97	92.77	99.77	96.03	97.57	94.37	97.17	99.73	99.63	83.13	88.81	91.93	55.73	27.77	33.23	0.0
Qwen2.5-1.5B	17.62	0.6	88.46	50.0	65.23	57.47	100.0	94.4	99.17	94.2	80.7	98.3	94.6	92.9	62.38	44.97	33.1	19.07	20.63	0.0
Qwen2.5-3B	18.59	1.33	94.87	57.69	83.2	70.17	100.0	98.43	81.47	94.5	85.63	96.6	83.07	97.23	83.62	50.6	29.4	19.83	20.93	0.0
SmolLM2-1.7B	53.49	0.4	32.05	0.0	22.8	14.17	57.13	62.3	99.47	92.77	86.27	63.5	64.03	99.83	80.02	23.27	39.43	40.2	31.67	0.0
SmolLM2-360M	96.8	3.1	7.69	0.0	28.8	31.83	99.93	99.6	98.73	99.63	99.67	68.63	40.47	99.27	84.31	52.8	57.23	61.77	85.63	0.0
occiglot-7b-es-en	67.8	1.47	1.28	0.0	33.2	49.4	17.87	15.83	36.33	62.5	42.77	7.37	0.53	56.97	57.48	21.27	25.07	27.93	16.6	0.0
occiglot-7b-eu5	54.79	2.47	2.56	0.0	30.83	21.77	19.33	27.23	14.3	23.8	18.8	0.07	0.0	70.87	57.64	25.73	3.7	25.07	7.37	0.0
phi-4	16.85	0.0	5.13	20.51	84.9	93.2	99.97	96.07	96.3	97.5	99.73	99.47	99.33	94.77	95.67	78.9	60.13	16.37	25.17	0.0
salamandra-2b	11.54	53.57	6.41	0.0	0.07	0.07	0.0	0.0	0.77	9.57	7.87	0.1	2.63	12.37	0.29	0.0	0.07	6.87	8.2	0.0
salamandra-7b	4.54	97.07	12.82	0.0	11.9	15.47	0.0	0.0	34.03	31.1	30.13	1.53	16.73	12.7	24.71	1.07	0.27	2.5	2.5	0.0

Table 46: Detailed accuracy results (%) for a subset of tasks in the Korean language

task	ends_with_letter	first_word	last_word	first_letter	last_letter	any_words_from_category_5_distractors	any_words_from_category_4_distractors	any_words_from_category_3_distractors	all_words_from_category_0_distractors	all_words_from_category_1_distractors	all_words_from_category_2_distractors	rhyming_word_orthographically_similar
EuroLLM-1.7B	0.0	61.7	70.0	36.87	41.3	95.77	94.97	96.47	95.97	95.97	96.13	98.17
EuroLLM-9B	0.0	53.3	43.3	17.07	35.6	99.8	100.0	100.0	99.2	98.17	99.17	87.6
ANITA-8B	0.0	36.27	22.83	46.2	27.47	91.23	95.77	97.07	100.0	29.1	46.43	54.47
Latxa-Llama-3.1-8B	0.0	65.5	45.73	65.07	47.2	93.47	100.0	100.0	99.23	58.0	77.97	98.13
Llama-3.1-8B	0.0	67.73	71.97	77.37	53.53	98.5	100.0	100.0	100.0	69.27	84.57	99.63
Llama-3.2-1B	0.0	76.07	71.67	27.47	36.4	95.53	96.17	96.47	98.47	95.93	95.27	94.2
Llama-3.2-3B	0.0	83.23	35.77	46.7	44.83	100.0	100.0	99.93	100.0	91.33	95.27	98.13
Qwen2.5-14B	0.0	58.93	52.83	55.37	57.33	99.97	99.63	99.3	99.87	91.63	93.43	88.83
Qwen2.5-1.5B	0.0	58.47	38.57	65.73	67.93	100.0	100.0	100.0	100.0	88.13	94.3	90.97
Qwen2.5-3B	0.0	56.4	41.17	64.8	47.67	100.0	100.0	100.0	100.0	97.07	98.53	97.47
SmolLM2-1.7B	0.0	65.5	55.13	27.77	18.83	8.87	99.97	99.93	99.73	25.93	25.13	99.77
SmolLM2-360M	0.0	86.53	84.5	16.93	7.47	99.97	99.97	99.97	99.7	99.43	99.8	99.2
occiglot-7b-es-en	0.0	4.0	32.2	4.1	36.1	18.17	18.0	15.27	16.7	15.07	12.17	52.6
occiglot-7b-eu5	0.0	0.0	8.67	0.37	31.7	1.77	37.47	44.53	51.87	6.17	4.37	69.2
phi-4	0.0	76.07	77.1	68.3	61.93	99.97	99.93	100.0	100.0	91.03	95.73	98.63
salamandra-2b	0.0	0.37	0.3	0.43	0.2	0.0	0.57	0.0	0.0	0.03	0.0	11.37
salamandra-7b	0.0	0.33	1.53	0.03	0.03	0.0	0.0	0.0	0.0	0.0	0.0	12.47

Table 47: Detailed accuracy results (%) for a subset of tasks in the Korean language

Model	sentence_containing	sentence_not_containing	word_containing	word_not_containing	most_associated_word	least_associated_word	any_words_from_category	all_words_from_category	first_alphabetically	more_letters	less_letters	bigger_number	smaller_number	rhyming_word	homophones	word_after	word_before	starts_with_word	ends_with_word	starts_with_letter
EuroLLM-1.7B	90.67	32.43	8.97	0.0	29.33	12.53	44.9	44.6	52.03	63.0	58.27	65.33	68.0	0.0	0.0	4.9	4.1	6.63	2.1	19.23
EuroLLM-9B	92.5	41.93	0.0	0.0	0.03	58.23	48.4	58.63	47.37	70.57	68.03	81.2	79.87	37.63	0.0	0.0	0.0	16.6	6.6	0.0
ANITA-8B	94.1	99.47	103.85	93.59	81.43	85.23	53.5	58.7	70.33	45.1	59.2	94.53	91.67	42.2	70.66	71.13	36.1	18.37	92.97	69.23
Latxa-Llama-3.1-8B	90.17	94.07	69.23	12.82	0.03	79.57	60.7	54.67	25.47	90.4	88.03	78.73	72.87	11.9	30.05	0.73	0.17	17.17	54.87	53.85
Llama-3.1-8B	93.63	96.7	79.49	5.13	0.0	88.53	58.57	59.63	76.43	90.33	90.07	98.93	98.87	36.57	22.42	0.13	0.03	17.63	71.53	69.23
Llama-3.2-1B	77.83	95.23	58.97	2.56	0.0	29.97	47.5	50.47	35.2	49.53	58.27	47.6	54.53	12.73	24.47	0.0	0.0	12.03	25.87	61.54
Llama-3.2-3B	91.8	96.73	88.46	70.51	0.0	62.03	47.6	53.67	81.83	88.47	86.2	98.8	98.7	34.87	19.19	0.03	0.0	16.77	57.67	61.54
Qwen2_5-14B	93.8	96.87	101.28	92.31	0.1	27.83	48.9	71.47	42.87	66.2	93.73	66.67	66.63	30.37	5.16	0.2	0.3	17.97	81.23	80.77
Qwen2.5-1.5B	79.67	77.37	97.44	75.64	0.87	0.7	39.07	60.3	23.13	21.33	35.17	35.27	23.77	0.93	0.06	1.8	0.7	5.87	4.03	53.85
Qwen2.5-3B	85.23	82.93	102.56	85.9	2.93	13.9	47.5	66.8	56.8	67.0	71.57	62.47	16.43	0.33	5.4	13.5	2.43	16.17	24.23	73.08
SmolLM2-1_7B	86.6	88.07	69.23	42.31	34.53	13.2	50.23	54.6	62.0	53.3	50.13	74.7	55.97	29.23	33.45	4.93	5.1	13.83	19.33	70.51
SmolLM2-360M	90.2	59.57	78.21	1.28	5.33	7.43	35.93	18.33	7.77	36.57	32.83	53.77	0.83	1.13	3.81	0.43	0.2	17.17	28.5	28.21
occiglot-7b-es-en	95.97	3.13	1.28	0.0	0.0	63.93	22.7	32.27	34.23	41.1	23.43	53.07	58.13	0.0	0.0	0.0	0.03	16.97	0.27	0.0
occiglot-7b-eu5	85.4	60.0	0.0	0.0	0.0	62.97	42.0	57.17	64.17	40.4	52.8	75.37	73.9	0.0	0.0	0.03	0.0	16.9	0.0	0.0
phi-4	92.23	93.37	34.62	78.21	0.0	93.63	56.3	68.7	96.6	92.47	89.97	100.0	97.73	7.8	3.46	0.0	0.0	18.57	53.9	71.79
salamandra-2b	74.9	44.83	82.05	46.15	5.3	1.97	54.13	52.63	15.03	17.83	23.6	2.7	0.07	4.2	1.0	2.1	0.8	7.83	4.1	56.41
salamandra-7b	82.0	48.93	66.67	70.51	13.8	14.6	47.4	54.77	46.37	11.83	26.7	44.0	54.33	21.07	8.27	16.47	2.3	10.07	4.3	64.1

Table 48: Detailed accuracy results (%) for a subset of tasks in the Portuguese (Brazilian) language

Model	ends_with_letter	first_word	last_word	first_letter	last_letter	first_alphabetically_far_first_letter	first_alphabetically_different_first_letter	first_alphabetically_consecutive_first_letter	first_alphabetically_same_first_letter	any_words_from_category_5_distractors	any_words_from_category_4_distractors	any_words_from_category_3_distractors	all_words_from_category_0_distractors	all_words_from_category_1_distractors	all_words_from_category_2_distractors	rhyming_word_orthographically_similar	more_letters_length_diff_3plus	more_letters_length_diff_1	less_letters_length_diff_3plus	less_letters_length_diff_1
EuroLLM-1.7B	5.56	56.17	20.37	78.67	23.13	52.07	52.3	52.03	46.57	20.33	66.93	67.0	72.33	15.83	15.47	0.0	69.23	60.9	51.27	54.6
EuroLLM-9B	5.56	53.53	52.33	33.1	29.0	44.4	46.2	47.2	43.03	98.77	2.33	9.3	55.17	62.13	77.83	41.03	85.7	59.07	79.43	61.47
ANITA-8B	94.44	94.7	87.17	91.9	54.37	74.5	69.67	67.3	67.17	96.67	13.83	36.63	91.53	21.27	44.2	43.03	53.47	36.53	67.13	47.93
Latxa-Llama-3.1-8B	5.56	15.23	11.57	41.03	31.2	26.83	24.53	25.13	20.73	84.8	33.7	59.77	94.33	12.4	25.17	13.0	99.23	75.8	93.6	76.87
Llama-3.1-8B	27.78	39.77	20.57	49.73	31.87	80.57	77.7	74.17	59.2	95.07	25.4	45.77	90.13	25.4	44.97	36.9	99.73	75.63	99.67	78.33
Llama-3.2-1B	13.89	37.83	18.8	38.57	29.87	35.37	34.83	35.97	32.5	100.0	0.0	0.83	3.07	97.1	98.07	13.3	57.5	42.37	62.0	54.2
Llama-3.2-3B	36.11	29.6	37.93	51.03	36.5	79.67	79.13	79.3	65.1	92.03	2.1	7.1	20.8	73.57	85.1	35.2	98.43	72.87	94.9	69.9
Qwen2_5-14B	52.78	89.7	87.67	40.83	31.17	48.4	45.17	41.97	33.6	99.77	2.23	10.43	59.73	88.87	98.37	33.27	70.9	57.0	99.73	84.73
Qwen2.5-1.5B	16.67	28.8	16.63	63.47	6.47	24.6	23.2	23.63	22.17	80.2	5.23	17.33	74.5	43.37	65.17	1.13	23.2	21.13	37.2	33.4
Qwen2.5-3B	19.44	87.6	42.63	87.4	43.67	64.67	59.37	56.23	58.07	99.97	0.13	3.6	47.67	82.8	97.0	0.3	80.3	53.87	81.67	67.67
SmolLM2-1_7B	63.89	31.7	5.27	87.17	7.87	61.37	59.47	61.53	54.33	96.03	8.07	15.97	42.0	60.7	68.97	31.07	60.73	53.2	50.8	48.33
SmolLM2-360M	30.56	5.9	7.93	20.1	2.23	8.6	8.4	8.27	6.97	6.13	69.2	61.8	31.43	0.73	0.87	1.73	40.63	37.53	33.1	33.3
occiglot-7b-es-en	5.56	71.5	76.4	33.33	14.93	30.07	30.17	30.0	29.47	45.77	0.33	1.8	19.03	40.57	51.37	0.0	49.57	33.37	24.03	23.73
occiglot-7b-eu5	11.11	30.53	21.43	32.97	29.7	64.03	64.8	63.77	52.3	83.17	7.6	25.07	52.4	50.4	59.3	0.0	46.8	35.3	56.47	49.03
phi-4	19.44	42.27	50.77	33.2	31.07	99.1	98.37	94.97	74.17	96.83	17.03	30.77	73.23	60.23	83.2	7.97	99.57	81.53	97.43	81.17
salamandra-2b	8.33	18.47	5.63	20.87	4.93	14.27	14.33	13.83	13.73	72.53	35.13	44.1	65.23	31.17	34.53	3.87	18.2	17.67	23.53	24.2
salamandra-7b	16.67	47.67	21.13	54.17	20.73	43.53	48.7	47.83	45.5	100.0	0.13	1.37	16.33	91.67	97.5	21.17	10.83	10.8	27.67	25.8

Table 49: Detailed accuracy results (%) for a subset of tasks in the Portuguese (Brazilian) language

Model	*sentence_containing*	*sentence_not_containing*	*word_containing*	*word_not_containing*	*most_associated_word*	*least_associated_word*	*any_words_from_category*	*all_words_from_category*	*first_alphabetically*	*more_letters*	*less_letters*	*bigger_number*	*smaller_number*	*rhyming_word*	*homophones*	*word_after*	*word_before*	*starts_with_word*	*ends_with_word*	*starts_with_letter*
EuroLLM-1.7B	99.9	0.6	12.82	64.1	77.37	21.97	82.23	61.67	52.73	62.83	53.73	67.77	69.0	0.0	0.0	22.17	8.27	0.0	2.4	20.51
EuroLLM-9B	95.4	56.0	7.69	58.97	89.83	66.77	58.27	74.27	35.63	68.83	67.03	95.83	64.4	37.83	24.71	52.77	16.37	94.77	23.7	17.95
ANITA-8B	97.27	99.9	100.0	96.15	95.1	95.9	68.0	74.47	66.77	81.1	91.37	75.27	94.77	24.67	27.71	77.83	38.07	99.8	92.93	100.0
Latxa-Llama-3.1-8B	93.23	99.8	80.77	97.44	96.17	100.0	75.2	58.73	96.7	90.67	90.77	93.2	93.3	29.27	37.42	86.7	47.27	98.87	87.07	100.0
Llama-3.1-8B	95.57	99.43	97.44	97.44	91.77	84.47	87.0	71.23	99.5	95.4	96.2	86.83	93.77	33.3	42.25	95.23	54.3	96.83	91.93	100.0
Llama-3.2-1B	81.8	99.73	98.72	98.72	78.4	76.77	64.8	52.83	86.07	59.53	55.07	81.73	52.43	37.73	44.33	70.67	13.23	60.93	31.13	100.0
Llama-3.2-3B	94.93	99.53	92.31	97.44	94.87	80.1	60.7	65.83	94.0	95.17	96.03	92.03	70.8	25.63	43.42	92.33	40.67	97.17	85.43	100.0
Qwen2.5-14B	96.87	98.03	98.72	93.59	69.17	94.43	71.03	97.2	84.1	97.03	95.0	97.53	90.5	89.07	55.62	70.27	58.77	99.47	85.43	100.0
Qwen2.5-1.5B	87.63	88.57	100.0	83.33	33.53	18.83	58.97	64.5	32.77	54.97	55.63	87.73	48.2	19.13	33.42	54.53	17.5	34.03	5.5	100.0
Qwen2.5-3B	89.17	89.9	100.0	93.59	53.43	51.47	58.03	82.23	68.17	69.03	78.9	97.07	95.57	32.77	25.67	57.33	20.23	93.23	30.47	100.0
SmolLM2-1.7B	87.67	86.93	84.62	53.85	83.8	55.8	55.33	57.63	60.4	60.1	46.27	86.87	62.13	46.4	44.0	41.4	5.0	27.0	7.1	87.18
SmolLM2-360M	89.6	52.23	65.38	28.21	29.13	14.17	49.27	52.83	47.33	44.6	34.97	53.13	55.13	23.77	38.42	7.2	3.43	62.0	5.97	85.9
occiglot-7b-es-en	87.67	86.3	25.64	25.64	62.27	90.8	43.27	62.23	56.0	82.87	69.2	45.33	72.43	0.0	0.0	75.93	19.0	94.2	0.0	2.56
occiglot-7b-eu5	86.03	96.73	20.51	37.18	38.1	43.1	50.13	68.4	54.53	59.13	50.23	49.3	40.83	0.0	0.0	0.47	1.57	97.6	0.0	6.41
phi-4	94.3	99.3	93.59	84.62	80.6	99.7	76.1	91.0	98.0	80.23	86.07	100.0	100.0	78.27	69.25	87.23	71.37	99.83	73.53	100.0
salamandra-2b	83.77	44.4	64.1	76.92	1.3	0.07	52.4	54.2	22.83	29.53	21.5	0.8	4.17	9.2	0.83	18.2	3.5	35.13	1.5	97.44
salamandra-7b	90.27	89.1	46.15	67.95	21.87	3.27	66.73	55.6	25.37	28.83	42.23	2.5	59.97	11.67	0.5	37.47	11.33	76.67	8.07	87.18

Table 50: Detailed accuracy results (%) for a subset of tasks in the English language

Model	*ends_with_letter*	*first_word*	*last_word*	*first_letter*	*last_letter*	*first_alphabetically_far_first_letter*	*first_alphabetically_different_first_letter*	*first_alphabetically_consecutive_first_letter*	*first_alphabetically_same_first_letter*	*any_words_from_category_5_distractors*	*any_words_from_category_4_distractors*	*any_words_from_category_3_distractors*	*all_words_from_category_0_distractors*	*all_words_from_category_1_distractors*	*all_words_from_category_2_distractors*	*rhyming_word_orthographically_similar*	*rhyming_word_orthographically_different*	*more_letters_length_diff_3plus*	*more_letters_length_diff_1*	*less_letters_length_diff_3plus*	*less_letters_length_diff_1*
EuroLLM-1.7B	2.9	95.87	42.4	97.73	48.43	53.53	54.9	53.87	53.9	92.4	70.9	92.27	100.0	20.37	21.57	0.0	0.0	67.0	59.67	51.03	50.2
EuroLLM-9B	18.84	98.23	85.33	99.97	95.93	36.23	35.17	35.9	35.07	100.0	16.1	53.17	99.93	44.83	89.93	38.83	29.83	83.9	55.2	76.63	64.07
ANITA-8B	98.55	95.67	97.73	98.97	98.33	67.87	64.5	63.63	62.1	100.0	35.97	67.0	99.9	45.73	93.43	27.4	15.73	90.87	80.43	97.5	85.83
Latxa-Llama-3.1-8B	49.28	99.5	98.2	97.4	98.03	97.77	97.13	96.97	79.97	100.0	51.93	86.4	100.0	14.17	54.37	28.87	23.13	97.7	83.83	97.43	84.43
Llama-3.1-8B	98.55	99.97	99.43	99.97	99.27	100.0	99.43	98.33	82.17	99.93	74.63	96.5	99.6	41.2	81.43	35.83	24.03	99.9	91.87	99.7	93.9
Llama-3.2-1B	43.48	88.23	90.87	99.93	93.63	90.6	88.07	77.7	63.2	63.03	68.7	85.0	66.2	37.83	48.3	40.17	30.53	58.87	46.5	59.33	52.8
Llama-3.2-3B	92.75	99.93	96.93	100.0	98.67	96.17	92.63	89.93	73.17	100.0	21.0	60.63	66.57	64.77	95.73	26.03	21.47	99.37	92.43	99.3	93.4
Qwen2.5-14B	78.26	99.97	97.83	100.0	97.57	85.73	84.33	79.03	69.23	100.0	42.63	79.83	100.0	94.07	100.0	91.6	74.83	99.97	93.97	99.67	93.17
Qwen2.5-1.5B	17.39	94.9	82.8	99.57	55.27	39.03	34.73	32.37	30.6	100.0	19.73	56.63	99.03	24.3	65.47	19.53	15.47	59.8	46.8	54.47	57.43
Qwen2.5-3B	49.28	84.27	77.5	99.87	80.13	74.4	67.73	57.6	53.73	99.97	16.9	42.6	97.73	64.97	97.97	35.3	26.47	80.33	66.77	88.23	75.13
SmolLM2-1.7B	21.74	91.8	77.0	84.63	60.07	59.87	57.0	55.73	52.83	99.97	11.1	50.13	99.63	11.43	34.17	45.83	43.6	69.0	52.43	45.9	48.7
SmolLM2-360M	0.0	35.73	33.8	78.57	16.37	44.73	46.8	52.17	49.8	85.5	15.03	18.83	32.8	74.0	80.67	24.3	21.53	45.13	40.83	32.97	37.6
occiglot-7b-es-en	20.29	77.47	87.33	36.57	89.03	61.8	58.03	58.87	48.03	86.87	0.07	2.27	70.37	51.1	68.33	0.0	0.0	88.67	77.03	79.0	60.8
occiglot-7b-eu5	31.88	49.7	4.9	49.17	65.03	57.57	56.67	59.1	50.47	100.0	0.83	16.37	99.77	34.37	71.1	0.0	0.0	66.33	54.13	56.9	50.27
phi-4	68.12	100.0	99.47	100.0	99.57	99.43	98.07	95.87	78.37	100.0	53.07	93.8	100.0	79.43	99.47	80.1	67.93	99.47	81.23	98.9	83.4
salamandra-2b	4.35	28.83	13.17	55.23	27.1	22.63	23.07	23.37	21.73	98.17	8.6	17.67	53.77	54.9	68.0	9.7	8.07	31.97	26.8	19.93	20.53
salamandra-7b	0.0	81.03	45.23	98.37	52.6	22.57	22.73	25.57	24.5	99.93	33.57	86.67	85.47	24.0	36.27	10.77	7.93	32.23	22.87	46.53	42.07

Table 51: Detailed accuracy results (%) for a subset of tasks in the English language

Task Name	Agreement (Accuracy %)
all_words_from_category	97.80
all_words_from_category_0_distractors	99.10
all_words_from_category_1_distractors	97.30
all_words_from_category_2_distractors	97.30
any_words_from_category	97.40
any_words_from_category_3_distractors	94.40
any_words_from_category_4_distractors	96.70
any_words_from_category_5_distractors	98.60
bigger_number	86.30
ends_with_letter	83.70
ends_with_word	96.00
first_alphabetically	91.90
first_alphabetically_consecutive_first_letter	92.50
first_alphabetically_different_first_letter	95.00
first_alphabetically_far_first_letter	93.00
first_alphabetically_same_first_letter	95.10
first_letter	89.50
first_word	94.50
global_accuracy	90.78
homophones	79.60
last_letter	95.00
last_word	93.70
least_associated_word	80.10
less_letters	93.60
less_letters_length_diff_1	94.10
less_letters_length_diff_3plus	91.80
more_letters	90.10
more_letters_length_diff_1	88.00
more_letters_length_diff_3plus	87.70
most_associated_word	79.20
rhyming_word	88.40
rhyming_word_orthographically_different	89.70
rhyming_word_orthographically_similar	86.70
sentence_containing	96.50
sentence_not_containing	92.40
smaller_number	89.30
starts_with_letter	77.50
starts_with_word	95.20
word_after	84.90
word_before	92.00
word_containing	73.90
word_not_containing	86.30

Table 52: Agreement values for regex pattern of English tasks (sorted alphabetically and shown as percentages).

Task Name	Agreement (Accuracy %)
all_words_from_category	86.60
all_words_from_category_0_distractors	76.50
all_words_from_category_1_distractors	87.10
all_words_from_category_2_distractors	88.40
any_words_from_category	84.50
any_words_from_category_3_distractors	89.40
any_words_from_category_4_distractors	86.80
any_words_from_category_5_distractors	80.30
bigger_number	85.00
ends_with_letter	90.50
ends_with_word	50.10
first_alphabetically	83.20
first_alphabetically_consecutive_first_letter	82.50
first_alphabetically_different_first_letter	81.90
first_alphabetically_far_first_letter	82.40
first_alphabetically_same_first_letter	84.10
first_letter	79.80
first_word	92.30
global_accuracy	81.43
last_letter	80.80
last_word	92.50
least_associated_word	92.40
less_letters	79.60
less_letters_length_diff_1	81.70
less_letters_length_diff_3plus	80.80
more_letters	81.40
more_letters_length_diff_1	80.80
more_letters_length_diff_3plus	80.70
most_associated_word	92.20
rhyming_word	84.10
rhyming_word_orthographically_similar	85.50
sentence_containing	39.10
sentence_not_containing	78.00
smaller_number	84.50
starts_with_letter	80.70
starts_with_word	49.60
word_after	93.10
word_before	93.00
word_containing	78.20
word_not_containing	78.70

Table 53: Agreement values for regex pattern of Basque tasks (sorted alphabetically and shown as percentages).

Task	en	it	es	de	ca	gl	eu	ko	pt_br
all_words_from_category	3000	3000	3000	3000	3000	3000	3000	3000	3000
all_words_from_category_0_distractors	3000	3000	3000	3000	3000	3000	3000	3000	3000
all_words_from_category_1_distractors	3000	3000	3000	3000	3000	3000	3000	3000	3000
all_words_from_category_2_distractors	3000	3000	3000	3000	3000	3000	3000	3000	3000
any_words_from_category	3000	3000	3000	3000	3000	3000	3000	3000	3000
any_words_from_category_3_distractors	3000	3000	3000	3000	3000	3000	3000	3000	3000
any_words_from_category_4_distractors	3000	3000	3000	3000	3000	3000	3000	3000	3000
any_words_from_category_5_distractors	3000	3000	3000	3000	3000	3000	3000	3000	3000
bigger_number	3000	3000	3000	3000	3000	3000	3000	3000	3000
ends_with_letter	69	54	51	72	60	30	39	120	36
ends_with_word	3000	3000	3000	3000	3000	3000	3000	3000	3000
first_alphabetically	3000	3000	3000	3000	3000	3000	3000	3000	3000
first_alphabetically_consecutive_first_letter	3000	3000	3000	3000	3000	3000	3000	—	3000
first_alphabetically_different_first_letter	3000	3000	3000	3000	3000	3000	3000	—	3000
first_alphabetically_far_first_letter	3000	3000	3000	3000	3000	3000	3000	—	3000
first_alphabetically_same_first_letter	3000	3000	3000	3000	3000	3000	3000	—	3000
first_letter	3000	3000	3000	3000	3000	3000	3000	3000	3000
first_word	3000	3000	3003	3003	3003	3000	3000	3000	3000
homophones	2400	360	2304	4704	2784	792	—	2448	1704
last_letter	3000	3000	3000	3000	3000	3000	3000	3000	3000
last_word	3000	3000	3003	3003	3003	3000	3000	3000	3000
least_associated_word	3000	3000	3000	3000	3000	3000	3000	3000	3000
less_letters	3000	3000	3000	3000	3000	3000	3000	3000	3000
less_letters_length_diff_1	3000	3000	3000	3000	3000	3000	3000	3000	3000
less_letters_length_diff_3plus	3000	3000	3000	3000	3000	3000	3000	3000	3000
more_letters	3000	3000	3000	3000	3000	3000	3000	3000	3000
more_letters_length_diff_1	3000	3000	3000	3000	3000	3000	3000	3000	3000
more_letters_length_diff_3plus	3000	3000	3000	3000	3000	3000	3000	3000	3000
most_associated_word	3000	3000	3000	3000	3000	3000	3000	3000	3000
rhyming_word	3000	3000	3000	3000	3000	3000	3000	3000	3000
rhyming_word_orthographically_different	3000	—	—	—	—	—	—	—	—
rhyming_word_orthographically_similar	3000	3000	3000	3000	3000	3000	3000	3000	3000
sentence_containing	3000	3000	3000	3000	3000	3000	3000	2997	3000
sentence_not_containing	3000	3000	3000	3000	3000	3000	3000	3000	3000
smaller_number	3000	3000	3000	3000	3000	3000	3000	3000	3000
starts_with_letter	78	78	78	78	78	78	78	78	78
starts_with_word	3000	3000	3000	3000	3000	3000	3000	3000	3000
word_after	3000	3000	3000	3000	3000	3000	3000	3000	3000
word_before	3000	3000	3000	3000	3000	3000	3000	3000	3000
word_containing	78	99	81	90	81	81	81	78	81
word_not_containing	78	99	81	90	81	81	81	78	81

Table 54: Number of samples present in the Multilingual LMentry for each tasks across all the languages.