Continued Pretraining and Interpretability-Based Evaluation for Low-Resource Languages: A Galician Case Study

Pablo Rodríguez, Silvia Paniagua Suárez, Pablo Gamallo, Susana Sotelo Docio
Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS-USC)
{pablorodriguez.fernandez, silvia.paniagua.suarez,
pablo.gamallo, susana.sotelo.docio}@usc.gal

Abstract

Recent advances in Large Language Models (LLMs) have led to remarkable improvements in language understanding and text generation. However, challenges remain in enhancing their performance for underrepresented languages, ensuring continual learning without catastrophic forgetting, and developing robust evaluation methodologies. This work addresses these issues by investigating the impact of Continued Pretraining (CPT) on multilingual models and proposing a comprehensive evaluation framework for LLMs, focusing on the case of Galician language. Our first contribution explores CPT strategies for languages with limited representation in multilingual models. We analyze how CPT with Galician corpora improves text generation while assessing the trade-offs between linguistic enrichment and task-solving capabilities. Our findings show that CPT with small, high-quality corpora and diverse instructions enhances both task performance and linguistic quality. Our second contribution is a structured evaluation framework based on distinguishing task-based and language-based assessments, leveraging existing and newly developed benchmarks for Galician. Additionally, we contribute new Galician LLMs, datasets for evaluation and instructions, and an evaluation framework.

1 Introduction

As researchers continue to push the boundaries of what large language models (LLMs) can achieve, several key limitations have emerged that deserve further investigation: how to improve the ability of LLMs to handle underrepresented languages (Kiulian et al., 2024; Nguyen et al., 2024), how to ensure that they learn new knowledge and abilities without forgetting the knowledge and abilities they already possess (Alexandrov et al., 2024; Shi et al., 2024), and how to enhance evaluation so as to measure their generalization capabilities across different tasks. Our work addresses these challenges by exploring the impact of continued pretraining (CPT) and proposing a comprehensive framework for evaluating LLMs.

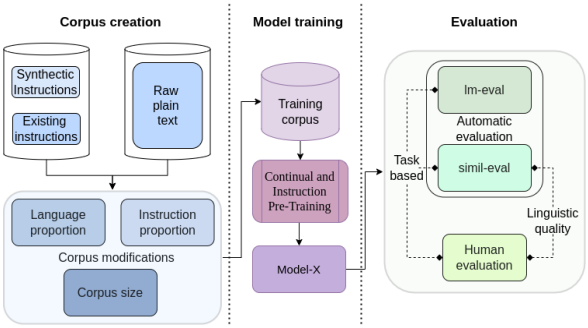


Figure 1: The proposed approach involves three stages: corpus creation using a mix of instructions and raw text, model training via continual and instruction-based pretraining, and evaluation through automatic and human assessments.

The first contribution of this work is an investigation into the impact of CPT on LLM performance, particularly for languages that are underrepresented in the base multilingual model, such as Galician. We will explore different strategies for doing CPT of Galician corpora in multilingual base models that have little or no knowledge of Galician. We hypothesize that CPT with text in such a minority language improves the ability of the base multilingual model to generate text in that language. However, if instructions are not incorporated into the training corpus during CPT, the model risks losing its ability to perform task-solving activities effectively. By comparing the effects of CPT with and without instructions, we aim to provide insights into balancing linguistic enrichment and task-solving capabilities.

The second contribution of our work is a comprehensive framework for evaluating LLMs, *simil-eval*[1], designed to assess their performance across

[1]Available in https://github.com/proxectonos/simil-eval

Findings of the Association for Computational Linguistics: ACL 2025, pages 4622–4637

the three dimensions introduced in Chang et al. (2024): *what to evaluate*, *how to evaluate*, and *where to evaluate*. For the first question, *what to evaluate*, we distinguish between two different aspects: the tasks that the model performs and the language that the model generates. Task-based evaluation measures the ability of a model to respond to questions, solve problems, and perform activities. Complementing this, language-based evaluation assesses the quality of the generated text, emphasizing coherence, fluency, and grammaticality. These two aspects are in line with the distinction made in cognitive sciences between functional linguistic competence, that is, non-language-specific skills required for real-life language use, and formal linguistic competence, that is, knowledge of linguistic rules and patterns (Mahowald et al., 2024). In the present work, we will evaluate several LLMs in Galician to examine the relationship between task performance and language quality and determine whether correlations exist between these two aspects. To tackle the second question, *how to evaluate*, we will analyze and make use of different strategies and metrics, including log-likelihood-based evaluation, which evaluates performance based on the probability assigned to correct answers, similarity-based evaluation, which compares generated outputs against expected responses using metrics such as cosine similarity or semantic overlap, and human-based evaluation. All these '*hows*' are essential to have a holistic view of the quality of the evaluated models. Finally, for the *where to evaluate*, we will explore all the benchmarks that are freely available to evaluate LLMs in Galician language. Some of them have been developed for the present study, as they were not available until now.

By addressing these two main contributions, our work offers insights into the interplay between CPT, instruction pre-training, and evaluation methodologies. More precisely, we found that CPT on small, high-quality corpora provided with diverse instructions achieves competitive performance in task-based evaluations while enhancing linguistic quality in generative tasks. Although this study focuses on Galician, the proposed methodology (Figure 1) and evaluation framework are language-agnostic and can be applied to other underrepresented languages with similar resource constraints, providing a general approach to developing more robust and adaptable LLMs with improved linguistic competence and task-solving capabilities.

Other relevant specific contributions are the following: i) the development of new generative LLMs with 8B parameters for the Galician language with CPT, ii) three new Galician datasets for LLM evaluation, iii) synthetic datasets instructions in Galician, and iv) an evaluation framework, called *simil-eval*, with a set of similarity-based metrics. All contributions are released under free and open-source licences.

This work is organized as follows. Section 2 describes, on the one hand, studies on the relationship between CPT and instructions and, on the other, the characteristics and challenges of the main evaluation frameworks. Section 3 introduces the main features of the models we have elaborated with CPT, our similarity-based evaluation framework, and the datasets we have used and elaborated for evaluation along with the human evaluation protocols. Then, in Section 4, we describe the experiments carried out, show the results and discuss the main findings. We finish in Section 5 with some concluding remarks, future work, and limitations.

2 Related work

In this section we will review the most relevant studies on the relationship between CPT and the use of instructions and will analyze different methods to evaluate LLMs.

2.1 Instructions and CPT

The enhancement of LLMs through CPT and instruction fine-tuning has been extensively explored in recent studies. One of the challenges of these studies lies in providing models with general instruction-following capabilities.

Some approaches focus on how to improve the pre-training of the model to allow the introduction of new knowledge and, at the same time, learning to answer questions and multitask through instructions. Cheng et al. (2024) proposed Instruction Pre-Training, augmenting raw text with instruction-response pairs synthesized by a generative model. This enables LLMs to pre-train on simulated multitask corpora, aligning learning objectives with task-specific behaviors. Training from scratch with this method enhances base model performance and improves fine-tuning gains. Similarly, Nayak et al. (2024) developed synthetic instruction datasets for zero-shot adaptation, refining high-quality synthetic instruction generation for Instruction Pre-Training. Jindal et al. (2024) examined trade-offs

between CPT and instruction fine-tuning, identifying efficient strategies to maintain knowledge and instruction-following capabilities with minimal cost. Revisiting CPT, Shi and Lipani (2023) introduced Prompt-based Continued Pretraining, integrating task-related texts with prompt templates during unsupervised pretraining. This approach significantly enhances fine-tuning performance across tasks, outperforming state-of-the-art methods in simplicity and effectiveness.

Recent work by Hu et al. (2024) on MiniCPM proposes scalable instruction tuning strategies, including an annealing mechanism for better task alignment during training. This highlights the growing interest in optimizing instruction signal integration at various model scales, a principle also relevant to our approach.

The integration of insights from these studies informs our work, particularly regarding the hypothesis that CPT in underrepresented languages can improve linguistic capabilities while risking a decline in task-solving skills if instructions are omitted. So, our present study builds upon prior research by combining the strengths of instruction synthesis and instruction pretraining to explore the nuanced effects of CPT with structured instructions on LLM performance.

2.2 Challenges in the evaluation of LLMs

Evaluating LLMs is challenging due to their sensitivity to prompt design, formatting, and task-specific instructions. Even minor variations impact outcomes, making reproducibility and interoperability essential. Weber et al. (2023) notes that prompt-based methods yield inconsistent performance, limiting generalization. To address this, several libraries simplify evaluation. Biderman et al. (2024) describe lm-eval as a flexible benchmarking framework, supporting interoperability and custom leaderboards. This aligns with a growing recognition in recent work (Ming et al., 2024; Ji et al., 2025) of the need to expand multilingual capabilities and ensure that evaluation frameworks fairly reflect linguistic diversity, particularly in underrepresented and low-resource language settings. Built on lm-eval, IberoBench (Baucells et al., 2025) offers a multilingual benchmark for Basque, Catalan, Galician, European Spanish, and Portuguese, covering 62 tasks and 179 subtasks for standardized 0-shot and 5-shot evaluations. Zhu et al. (2024) introduce PromptBench, designed for adversarial and dynamic evaluations with modular tools for custom pipelines.

While effective for structured tasks, these tools struggle with free-text evaluation. Metrics like BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) often misalign with human judgments, failing to capture coherence, relevance, and fluency (Deutsch et al., 2022). BERTScore (Zhang et al., 2020) and MoverScore (Zhao et al., 2019) improve contextual understanding but still lack nuance in assessing originality.

When automated metrics fall short, human evaluation becomes essential. Novikova et al. (2017) argue that human judgments better capture coherence, fluency, and creativity, making them the gold standard. Chern et al. (2024) explore LLM-based evaluation (LLM-as-a-Judge), leveraging their linguistic capabilities, though further validation is needed to ensure alignment with human assessments. In the present work, as their performance on low-resource languages like Galician is often inferior to their performance on high-resource languages due to limited training data, they will not be used.

Along with the lm-eval platform, which will be of particular help in evaluating structured tasks with probabilistic metrics, we will also make use, especially in text generated by the models, of similarity-based measures. In addition, we will calculate the degree of correlation between these two types of metrics with each other, as well as with respect to human evaluation.

3 Methodology and datasets

3.1 CPT and Instruction Pre-Training

To conduct a proper study, we use models obtained from the same base model by applying different corpora during the CPT phase. Llama-3.1-8B[2] was chosen as the base model due to its strong performance in multilingual settings. Since training only in a new language can lead to catastrophic forgetting of other language capabilities (Koloski et al., 2024), we include in the training corpus not only Galician texts but also texts in Spanish and English (present in the original Llama-3.1-8B), as well as Catalan and European Portuguese, given their linguistic proximity to Galician as Ibero-Romance languages. We use different corpus combinations to assess the impact of text quality, language proportion, and instruction pretraining during CPT.

[2] https://huggingface.co/meta-llama/Llama-3.1-8B

Due to the limited availability of resources in some of the languages, the instruction datasets[3] were synthetically generated using different techniques, either by adapting existing datasets or creating new ones from scratch:

Model-large[4]: Pretrained using all available raw text corpora in Galician and the other selected languages, ensuring a balanced distribution to prevent knowledge degradation.

Model-small-1 (high-quality corpus): Trained on a high-quality Galician corpus while maintaining a balanced distribution for the other languages.

Model-small-2 (close languages): Uses a high-quality Galician corpus with a significant proportion of Portuguese data, while the rest of the languages remain balanced.

Model-small-instr-1[5] (small instruction pretraining): Enrich the corpus of Model-small-1 by incorporating examples of entity recognition and question answering exclusively in Galician.

Model-small-instr-2 (instruction pretraining for translation): Enrich the corpus of Model-small-1 by incorporating examples of translation tasks (EN-GL, ES-GL) alongside standard pretraining.

Model-small-instr-3[6] (multitasking instruction pretraining): Enrich the corpus of Model-small-instr-1 by incorporating more instruction-following tasks, such as sentiment analysis, across all languages in the corpus.

It is important to note that the corpus used to train the Model-large contains more than 20 billion words, while the rest of the models have around 350 million words of plain text and a maximum of 35 million words of instructions. More information on the proportion of languages and instructions used in each model can be found in Appendix C. In addition, the specifications used during the CPT of the models are available in Appendix D.

3.2 Evaluation datasets for Galician

Evaluating the performance of LLMs requires well-defined benchmarks that assess their ability to generate coherent, fluent text as well as their proficiency in completing structured tasks. Evaluation approaches can be broadly categorized into generative evaluation and task-based evaluation, each serving distinct objectives. Generative evaluation focuses on assessing the linguistic quality of open-ended model outputs, measuring aspects such as fluency, coherence, and grammatical accuracy rather than correctness (Sheng et al., 2024). This type of evaluation is particularly relevant for tasks such as dialogue generation, storytelling, and creative writing, where multiple valid outputs may exist. In contrast, task-based evaluation assesses a model's ability to complete structured, objective tasks, such as translation, summarization, mathematical reasoning, and factual question answering (Guo et al., 2023). Unlike generative evaluation, task-based evaluation relies on gold-standard references, allowing for more objective assessment using established metrics such as BLEU, ROUGE, and Exact Match (Rajpurkar et al., 2016).

Building on the distinction between generative and task-based evaluation, we focus on Galician to analyze LLM performance in a low-resource language setting. To this end, we select a subset of IberoBench datasets covering summarization (*summarization_gl*), mathematical reasoning (*mgsm_direct_gl*), truthfulness (*truthfulqa_gl*), translation (*flores_gl*), multiple-choice reading comprehension (*belebele_gl*), linguistic acceptability (*galcola*), paraphrasing (*parafrases_gl* and *paws_gl*) and question answering (*openbookqa_gl*). In addition, we incorporate three new datasets: *xstorycloze_gl* for commonsense reasoning,[7] *xnli_gl* for semantic inference,[8] and *calame_gl*,[9] the Galician version of Calame-pt (Lopes et al., 2024), which evaluates a model's ability to predict the final word of a sentence when provided with sufficient contextual information for both humans and models to make an accurate guess.

3.3 Metrics

When evaluating the performance of autoregressive language models, we employ different automatic metrics tailored to each task type. For multiple-choice tasks, log-likelihood is the primary metric, assessing the model's ability to assign higher probabilities to correct answers. Surprisal, another probabilistic measure, is used to evaluate the predictability of generated text sequences. Finally, seman-

[3] https://github.com/proxectonos/instruction_datasets

[4] https://huggingface.co/proxectonos/Llama-3.1-Carballo

[5] https://huggingface.co/proxectonos/Llama-3.1-Carballo-Instr1

[6] https://huggingface.co/proxectonos/Llama-3.1-Carballo-Instr3

[7] https://huggingface.co/datasets/proxectonos/xstorycloze_gl

[8] https://huggingface.co/datasets/proxectonos/xnli_gl

[9] https://github.com/proxectonos/calame-gl

tic similarity measures (such as BERTScore and MoverScore) are applied to both multiple-choice and generative tasks to assess output quality beyond lexical overlap, capturing contextual and thematic coherence. Complementing these automatic metrics, we also incorporate human qualitative evaluations, which are essential for assessing fluency, coherence and overall text quality.

In the next subsections, we will describe each of these evaluation metrics in detail.

3.3.1 Probabilistic metrics

We consider two probabilistic metrics for the evaluation of the models: log-likelihood and surprisal.

Log-likelihood computes the probability assigned by the model to a sequence of tokens. Given an input consisting of tokens $x = (x_0, x_1, ..., x_{n-1})$, and a target sequence of tokens $y = (y_0, y_1, ..., y_{m-1})$, the *log-likelihood* is defined as:

$$\log P(y|x) = \sum_{i=0}^{m-1} \log P(y_i|x, y_0, \ldots, y_{i-1}) \quad (1)$$

where $P(y_i|x, y_0, \ldots, y_{i-1})$ represents the probability assigned by the language model to the i-th token in the target sequence, conditioned on the input sequence x and the preceding tokens in y.

The *surprisal* of a token y_n in a sequence $y = (y_0, y_1, \ldots, y_{n-1})$ is closely related to the log-likelihood. It can be defined as the negative log-probability of the token, i.e., $S(y_n|y) = -\log P(y_n|y)$, where the reference and target sequences are the same. This metric is particularly useful for evaluating the quality of a text, as the surprisal of a token reflects how unexpected or surprising that token is within its context in the sequence.

3.3.2 Similarity-based metrics

To assess whether the generated text adequately responds to a task, we use three similarity-based metrics that leverage embeddings to determine how closely a model's output aligns with reference texts: 1) *Cosine Similarity*: this metric measures the similarity between the last-layer embeddings of two text fragments, such as a model-generated response and its reference counterpart. Higher values indicate greater resemblance. 2) *BERTScore*: this metric uses a modified cosine measure with BERT embeddings to capture deeper semantic relationships than basic cosine. 3) *MoverScore*: by utilizing BERT-based embeddings, this metric estimates the effort required to transform one text into another. A lower transformation effort suggests a higher degree of similarity.

To evaluate these metrics, we developed a new framework, *simil-eval*, designed for simplicity and interpretability while also improving transparency in model assessment. In this framework, the model is prompted to complete a text, and its output is compared to a reference, which can be adapted for various tasks. This approach is particularly effective for multiple-choice tasks since, instead of selecting an option by letter or number, the model generates a response that is then compared to the available choices using similarity measures, and the most similar option is selected as the correct answer (examples in Appendix A). While lm-eval serves a similar purpose by relying on log-likelihood, similarity-based evaluations tend to be more intuitive for human interpretation.

Beyond similarity-based metrics, our framework also incorporates surprisal, previously introduced as a probabilistic metric. To further explore its effectiveness, we adapted the *galcola* dataset, originally designed for linguistic acceptability judgment (de Dios-Flores et al., 2023). In this adaptation, we compute the surprisal of sentence pairs—one grammatically correct and the other unacceptable, defining a new metric, *difsur*, in Equation 2, where x_a is an acceptable sentence and x_{na} is its unacceptable counterpart.

$$\textit{difsur} = \frac{S(x_{na}) - S(x_a)}{\max\{S(x_a), S(x_{na})\}} \times 100 \quad (2)$$

As difsur computes the relative difference in surprisal between a grammatical (x_a) and ungrammatical (x_{na}) version of the same sentence, normalized by the maximum surprisal value, it is able to ensure a more fair and scale-independent evaluation than using just surprisal on either grammatical or ungrammatical sentences. Indeed, difsur explicitly evaluates whether the model distinguishes correct from incorrect syntax, penalizing models that assign similar probabilities to both, showing higher values in models with more correction in writting. This makes it especially valuable for the evaluation of LLMs in applications where grammatical correctness is critical, providing a more comprehensive assessment of a model's linguistic competence beyond traditional similarity measures.

3.3.3 Human evaluation metrics

To complement automated evaluations, we conducted human assessments focusing on linguistic quality and task-specific performance.

Linguistic quality: To evaluate the perceived quality of model-generated text from a linguistic perspective, we conducted a human perception experiment based on the methodology from Gamallo et al. (2024), with some modifications adapted to our objectives. Unlike the original study, which compared authentic and synthetic continuations, our evaluation focused solely on model-generated outputs. This modification was necessary to enable a direct model-to-model comparison, ensuring that variations in linguistic quality and coherence resulted exclusively from model differences rather than disparities between human and machine-generated text. Evaluators were presented with continuations from different models for the same context, minimizing biases from authentic references. Despite this adjustment, we preserved key elements of the original methodology, such as text selection, splitting strategies, and counterbalancing, to maintain rigor.

The evaluation was conducted by four native Galician speakers, each tasked with assessing the outputs of a single model. Guided by detailed annotation instructions, they evaluated model-generated continuations according to two predefined categories: 1) *Form error:* the text is fully comprehensible but contains structural or grammatical issues easily identifiable by a non-expert reader. (2) *Content error:* The text has inconsistencies in meaning, either in relation to the context or within the text itself, which may hinder comprehension.

For form errors, evaluators reported the number of errors identified in each continuation. Content errors were assessed with a binary classification, marking any detected inconsistency as erroneous, even if some parts of the text remained topic-accurate.

Task-specific: To assess the models' performance on open-ended tasks like summarization, mathematical reasoning, and truthfulness, we conducted a separate human evaluation. This evaluation was based on model-generated results from the automated lm-eval assessment, using IberoBench's open-ended tasks for Galician. The evaluation was carried out using Label Studio (Tkachenko et al., 2020-2022), an open-source annotation platform that provided an intuitive interface for annotators and allowed easy export of results in a structured format.

To conduct this evaluation, four native Galician speakers were each assigned to a specific model and tasked with assessing all three tasks associated with that model. Each evaluator reviewed a set of 50 items, which included the input prompt, the gold-standard reference, and the model-generated output. They were then instructed to classify the output as valid or invalid following these guidelines: For *summarization*, a summary was considered valid if it conveyed at least one key point of the text. Minor grammatical errors were disregarded unless they hindered readability. A summary was deemed invalid if it simply replicated the first paragraph without summarizing the text. For *truthfulness*, a response was valid if it aligned in meaning with any of the correct answers. For *mathematical reasoning*, an exact match was required for validity, meaning the response had to include the correct answer, but not necessarily a strict character-by-character match with the gold standard.

For all tasks, evaluators used the gold-standard reference as a guide rather than requiring exact matching, assessing outputs based on task-specific validity criteria. Any additional generated text was disregarded as long as the output met the specified criteria. Moreover, evaluators could also provide comments to contextualize their assessments.

4 Experiments

The objective is to evaluate language models with a significant proportion of Galician in their training data, focusing exclusively on pretrained models and excluding those with instruction tuning. The selected models vary in size and linguistic scope, from monolingual Galician models to multilingual models covering several European languages: *CPT models*: this group includes our models described in section 3.1; *Llama-3.1-8B*: the base model for CPT training, which is a multilingual model covering eight languages, including Spanish and Brazilian Portuguese; *Salamandra-2B*[10] and *Salamandra-7B*[11]: multilingual models focused on European languages including Galician, making them particularly relevant for assessing Galician language capabilities; *Carballo-Bloom*[12]: a monolingual 1.7B Galician model, offering insights into the perfor-

[10] https://huggingface.co/BSC-LT/salamandra-2b

[11] https://huggingface.co/BSC-LT/salamandra-7b

[12] https://huggingface.co/proxectonos/Carballo-bloom-1.3B

mance of a Galician-specific model compared to larger multilingual ones.

4.1 Results

Figure 2 shows the hit rates obtained in various multiple-choice tasks by the analyzed models, using MoverScore within simil-eval and log-likelihood in lm-eval. It is important to note that for *belebele_gl* and *openbookqa_gl*, a random selection would yield a hit rate of 0.25, while for *xstorycloze_gl*, it would be 0.5. In the case of *truthfulqa_gl*, establishing a reference is more challenging, as the number of possible options varies between 4 and 10 depending on the question, and it results in an average random hit rate of approximately 0.196. Similar results are observed when using alternative similarity measures, such as cosine, as shown in Figure 7 in Appendix B. This appendix also includes the results for BertScore for average similarity (Figure 8), which also follow the trend of the other similarity metrics, but are more difficult to interpret.

On the other hand, Figure 3 illustrates the discrepancies between automated evaluation results (lm-eval) and human qualitative assessment across three generative tasks. Metrics are accuracy and BLEU while the latter was divided by 100 to adjust the scale. To reduce the human evaluation burden, we chose the two small models with the best performance in quantitative evaluation. The Y-axis represents the difference between the two evaluation methods, where positive values indicate that lm-eval overestimates the performance of the model compared to human evaluation, while negative values denote the opposite. These results show that lm-eval consistently assigns lower scores than human evaluation in the summarization task across all models, except for Model-large, where both evaluations yield similar scores. This could suggest that metrics like BLEU might not be entirely suitable for assessing this type of task. As for the other tasks, differences are less pronounced, but lm-eval seems to slightly underestimate the performance of Model-small-instr-1 and Model-small-instr-3, while overestimating that of Llama-3.1-8B. These variations suggest that the reliability of lm-eval for evaluating generative tasks may depend on the specific model and the particular challenges of each task.

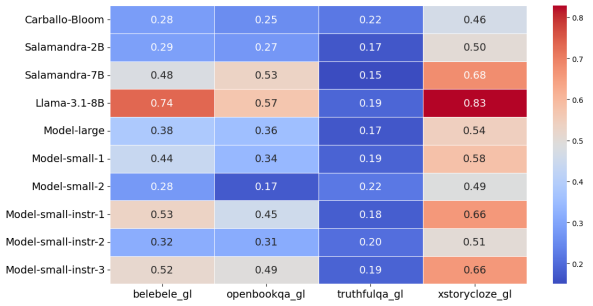
Finally, Figure 4 presents several comparisons between human error evaluation and surprisal-based metrics. In Figure 4a, the *difsur* values are shown for each automatically evaluated model. Additionally, for the models assessed by humans, the ratio of texts without form errors and without content errors is also included, scaled to 5 for readability. The higher the bars associated with a model, the better its performance, as this indicates a higher *difsur* value and a greater proportion of error-free texts. In contrast, Figure 4b displays the surprisal of the last word in *calame_gl* alongside the ratio of texts with form or content errors. In this case, lower bars are preferable, as they indicate that the model is less surprised by the expected word (*calame_gl*) and that the generated texts contain fewer errors according to human evaluation.

4.2 Discussion

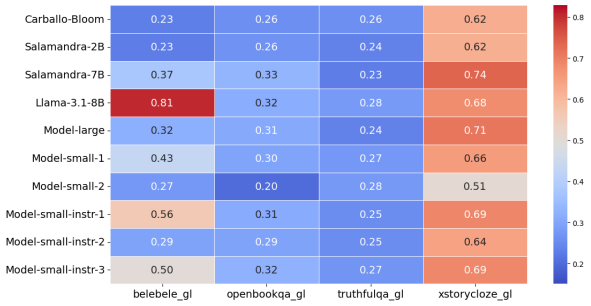
The results shown in the previous subsection allow us to infer very relevant information about, on the one hand, the use of instructions in CPT and, on the other hand, the effectiveness of the metrics tested from their comparison with human evaluations.

Concerning the task-based evaluation shown in the two plots of Figure 2, we observe that the larger base multilingual models without CPT (Salamandra-7B and Llama-3.1-8B) outperform, as expected, smaller models: multilingual base model (Salamandra-2B) and Galician Carballo-Bloom built with CPT. They also clearly outperform in almost all tasks Model-large and the two small models without instructions (Model-small-1 and 2). However, the small models with non-translation instructions, namely Model-small-instr-1 and 3, have acceptable performance close to both Llama-3.1-8B and Salamandra-7B. This demonstrates that the use of a variety of instructions in the training corpus during CPT is useful and allows the models not to lose the ability to solve tasks of different types. It is worth noting, however, that the instructions for translations used in Model-small-instr-2 only benefit the translation task, but do not help the model perform other types of tasks (see the performance on the *flores_gl* dataset in Table 2 of Appendix E).

On the other hand, if we compare the results of the similarity metric in Figure 2a and the probability metric in 2b, we find that there is a strong correlation between the two. Specifically, the Pearson correlation is 0.87. From this, we deduce that it is possible to use similarity-based techniques, which are more interpretable and transparent than those based on probability, on generated responses without losing performance efficiency.



(a) MoverScore hit rate.



(b) Log-likelihood hit rate.

Figure 2: Hit rates obtained using simil-eval (Figure 2a) and lm-eval (Figure 2b).

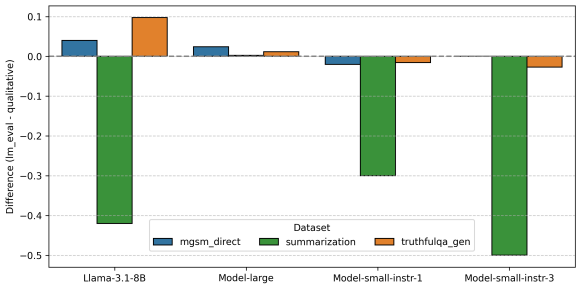


Figure 3: Difference between lm-eval and qualitative evaluation across three generation tasks

Figure 3, devoted to showing the relationship between probability (which correlates with similarity) and human evaluation, shows that there is a low correlation in the tasks that were evaluated manually. The overall Pearson correlation is 0.17. This indicates that it is not yet possible to rely on automatic metrics to evaluate the performance of the models in the evaluated tasks, and therefore, automatic methods must be rethought to improve the metrics. The fact that similarity-based metrics are more transparent and interpretable can help meet this objective.

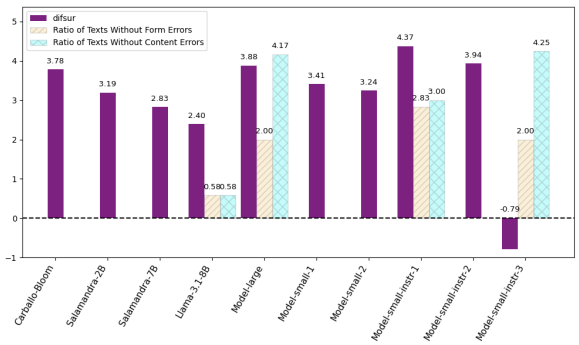
Regarding the evaluation of the linguistic quality of the generated text, it is possible to infer from Figure 4 that all the models trained with CPT in Galician, both large and small, substantially improve the base model Llama-3.1-8B. From this, we conclude that CPT is an appropriate method for adapting a model to a new language or variety. Moreover, the two plots in Figure 4 allow us to observe that none of the metrics used to evaluate linguistic quality, namely *difsur* in Figure 4a and last word surprisal in Figure 4b, correlates properly with the human evaluation, with a Pearson in both cases < 0.25. It is important to note that in the case of *difsur*, the correlation with human assessment is generally good except in one critical case, Model-small-instr-3, where the negative value, not expected in this case, causes the overall correlation to be very low. This suggests the necessity of further investigation into the automatic metric to enhance its performance or identify a more robust and stable alternative. In the case of the metric applied to *calame_gl*, which computes surprisal on the last word (Figure 4b), the main error lies in the attribution of a low score (and therefore, in this case, high quality) to the base model Llama-3.1-8B, which is not to be expected since this model has quality problems in the generation of texts in Galician. As in the previous metric, it will be necessary to investigate whether the metric can be improved or should be discarded as an automatic method to assess the linguistic quality of generated texts.

5 Conclusions

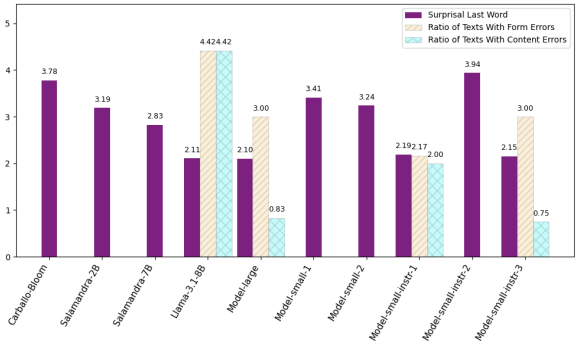
In this work we evaluated the impact of CPT on multilingual language models, comparing their performance using both automatic and human evaluation metrics. Larger base multilingual models outperform models that were adapted to a specific language with CPT in task-based evaluation, though CPT models trained with small and high-quality corpora, as well as with diverse instructions, show competitive results in task-based evaluation while they improve the linguistic quality in generative tasks. Some discrepancies between automatic and human evaluations indicate the need to refine the current metrics, both those that measure task performance and those that assess linguistic quality. In the current state of evaluation metrics, we must conclude that humans are still necessary to assess the results of LLMs more fairly.

5.1 Future work

In future work, we aim to enhance our models adapted via CPT through other tuning techniques,



(a) Comparison of difsur scores with the proportion of texts with content or form errors over 5. The higher the values, the better the quality.



(b) Comparison of surprisal of the last word in *calame_gl* with the proportion of texts without content or form errors over 5. The lower the values, the better the quality.

Figure 4: Comparison between two metrics and human assessment to compute the linguistic quality of texts.

such as instruction tuning, reasoning with reinforcement learning, or annealing, ensuring improved task performance without compromising linguistic quality. Additionally, we will refine our automatic metrics for evaluating linguistic quality, addressing potential discrepancies with human assessments. To further validate these metrics, we will leverage LLM-as-a-Judge systems, comparing their evaluations against both traditional automatic metrics and human judgments. This multi-perspective evaluation will help us develop more reliable and interpretable assessment methodologies for LLM-generated text.

Limitations

While this work makes significant contributions to understanding the impact of CPT on LLMs and proposes a comprehensive evaluation framework, several limitations arise. The study focuses primarily on Galician, a low-resource language. While the findings provide valuable insights, they may not fully generalize to other underrepresented or typologically distinct languages. Further research is needed to validate the proposed methods across a broader range of languages. Moreover, due to the scarcity of high-quality instruction datasets in Galician, synthetic data generation techniques were employed. While these methods enabled the inclusion of diverse tasks, the quality and authenticity of synthetic data may not fully replicate human-created datasets, potentially affecting the robustness of the models. Finally, while human evaluation adds valuable insights beyond what automated metrics can capture, it also has some important limitations in this study. The evaluation was carried out by four native Galician speakers, each assigned to assess outputs from a single model across 50 examples per task. While this setup ensured consistency and linguistic expertise, it also limited the diversity of perspectives and the overall scale of the evaluation. In addition, the evaluation covered only a small number of tasks and examples, which may not fully capture the range of possible model behaviors. These aspects should be taken into account when interpreting the results, and future work could benefit from involving more annotators, expanding task coverage, and increasing the number of evaluated items.

Acknowledgments

This research has received financial support from the Agencia Estatal de Investigación (LingUMT, grant PID2021-128811OA-I00), the Xunta de Galicia - Consellería de Cultura, Educación, Formación Profesional e Universidades (Centro de investigación de Galicia accreditation 2024-2027 ED431G-2023/04). Also, the ILENIA-Nós Project, which is funded by the Spanish Ministry of Economic Affairs and Digital Transformation and by the Recovery, Transformation, and Resilience Plan - Funded by the European Union - NextGenerationEU, with reference 2022/TL22/00215336.

We are grateful to CESGA (Centro de Supercomputación de Galicia) for allowing us access to their infrastructure to carry out the experiments.

References

Anton Alexandrov, Veselin Raychev, Mark Niklas Müller, Ce Zhang, Martin T. Vechev, and Kristina Toutanova. 2024. Mitigating catastrophic forgetting in language transfer via model merging. In *Con-*

ference on Empirical Methods in Natural Language Processing.

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

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

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

Daixuan Cheng, Yuxian Gu, Shaohan Huang, Junyu Bi, Minlie Huang, and Furu Wei. 2024. Instruction pre-training: Language models are supervised multitask learners. *CoRR*, abs/2406.14491.

Steffi Chern, Ethan Chern, Graham Neubig, and Pengfei Liu. 2024. Can large language models be trusted for evaluation? scalable meta-evaluation of llms as evaluators via agent debate. *ArXiv*, abs/2401.16788.

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

Daniel Deutsch, Rotem Dror, and Dan Roth. 2022. On the limitations of reference-free evaluations of generated text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10960–10977, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Pablo Gamallo, Pablo Rodríguez, Iria de Dios-Flores, Susana Sotelo, Silvia Paniagua, Daniel Bardanca, José Ramón Pichel, and Marcos Garcia. 2024. Open generative large language models for galician. *Procesamiento del Lenguaje Natural*, 73(0):259–270.

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

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies. *Preprint*, arXiv:2404.06395.

Shaoxiong Ji, Zihao Li, Indraneil Paul, Jaakko Paavola, Peiqin Lin, Pinzhen Chen, Dayyán O'Brien, Hengyu Luo, Hinrich Schütze, Jörg Tiedemann, and Barry Haddow. 2025. Emma-500: Enhancing massively multilingual adaptation of large language models. *Preprint*, arXiv:2409.17892.

Ishan Jindal, Chandana Badrinath, Pranjal Bharti, Lakkidi Vinay, and Sachin Dev Sharma. 2024. Balancing continuous pre-training and instruction fine-tuning: Optimizing instruction-following in llms. *Preprint*, arXiv:2410.10739.

Artur Kiulian, Anton Polishko, Mykola Khandoga, Yevhen Kostiuk, Guillermo Gabrielli, Łukasz Gagała, Fadi Zaraket, Qusai Abu Obaida, Hrishikesh Garud, Wendy Wing Yee Mak, Dmytro Chaplynskyi, Selma Belhadj Amor, and Grigol Peradze. 2024. From English-Centric to Effective Bilingual: LLMs with Custom Tokenizers for Underrepresented Languages. *Preprint*, arXiv:2410.18836.

Boshko Koloski, Blaž Škrlj, Marko Robnik-Šikonja, and Senja Pollak. 2024. Measuring catastrophic forgetting in cross-lingual transfer paradigms: Exploring tuning strategies. *Preprint*, arXiv:2309.06089.

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

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

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6):517–540.

Lingfeng Ming, Bo Zeng, Chenyang Lyu, Tianqi Shi, Yu Zhao, Xue Yang, Yefeng Liu, Yiyu Wang, Linlong

Xu, Yangyang Liu, Xiaohu Zhao, Hao Wang, Heng Liu, Hao Zhou, Huifeng Yin, Zifu Shang, Haijun Li, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2024. Marco-llm: Bridging languages via massive multilingual training for cross-lingual enhancement. *Preprint*, arXiv:2412.04003.

Nihal Nayak, Yiyang Nan, Avi Trost, and Stephen Bach. 2024. Learning to generate instruction tuning datasets for zero-shot task adaptation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12585–12611, Bangkok, Thailand. Association for Computational Linguistics.

Xuan-Phi Nguyen, Sharifah Mahani Aljunied, Shafiq Joty, and Lidong Bing. 2024. Democratizing llms for low-resource languages by leveraging their english dominant abilities with linguistically-diverse prompts. *Preprint*, arXiv:2306.11372.

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

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

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

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Shuqian Sheng, Yi Xu, Luoyi Fu, Jiaxin Ding, Lei Zhou, Xinbing Wang, and Chenghu Zhou. 2024. Is reference necessary in the evaluation of NLG systems? when and where? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8580–8596, Mexico City, Mexico. Association for Computational Linguistics.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, and Hao Wang. 2024. Continual learning of large language models: A comprehensive survey. *ArXiv*, abs/2404.16789.

Zhengxiang Shi and Aldo Lipani. 2023. Don't stop pretraining? make prompt-based fine-tuning powerful learner. *Preprint*, arXiv:2305.01711.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Lucas Weber, Elia Bruni, and Dieuwke Hupkes. 2023. The icl consistency test. *Preprint*, arXiv:2312.04945.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

Kaijie Zhu, Qinlin Zhao, Hao Chen, Jindong Wang, and Xing Xie. 2024. Promptbench: A unified library for evaluation of large language models. *Journal of Machine Learning Research*, (25):1–22.

A Examples of simil-eval outputs

```
ID 479 - Generated answer: placed in good soil
Score with option 1: planted in zinc pills: 0.5524
Score with option 2: plated in the sea: 0.5374
Score with option 3: placed in good soil: 0.9996
Score with option 4: made out of soil: 0.5544
Mean score with question 479: 0.66093
Score with correct option 3: 0.9996
```

Figure 5: Example output of simil-eval for MoverScore scores in a QA task.

```
--FINAL COSINE RESULTS--
Mean similarity score: 0.62045
Mean similarity score with correct options: 0.6217
Percentage of correct answers (over 1): 0.2626
```

Figure 6: Example output of simil-eval for final results of cosine score.

B Additional evaluation results

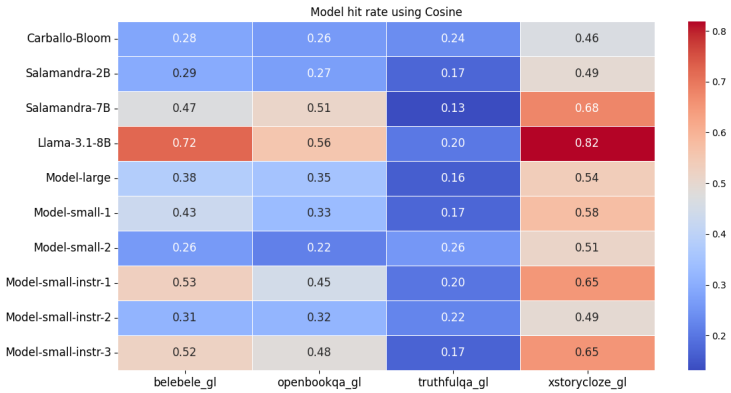


Figure 7: Cosine hit rates using simil-eval.

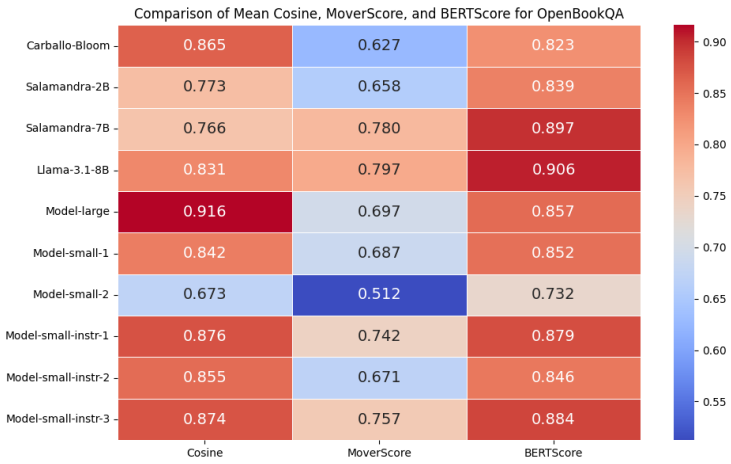


Figure 8: Comparison of the mean similarity of embeddings obtained with simil-eval using BertScore, Cosine and MoverScore in *openbookqa_gl*.

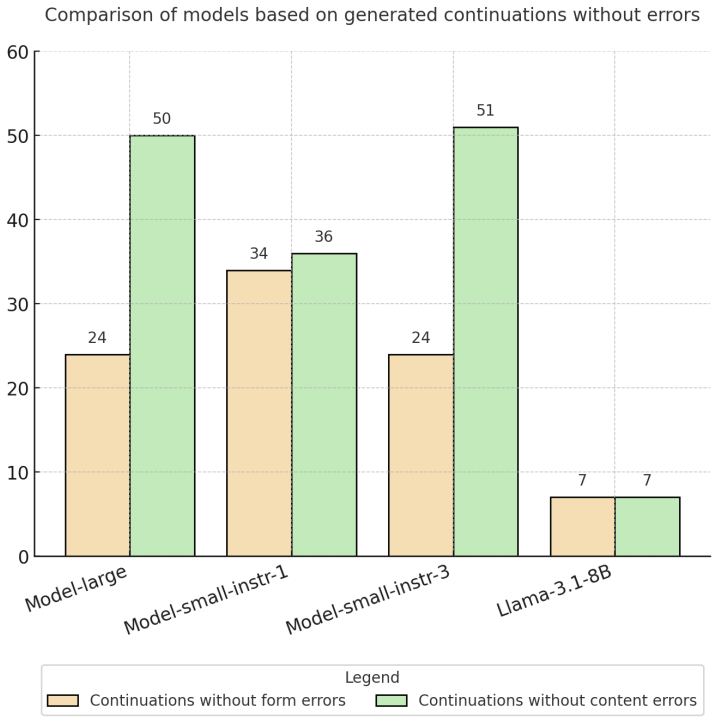


Figure 9: Number of error-free model-generated continuations in form and content out of 60 evaluated texts for each model.

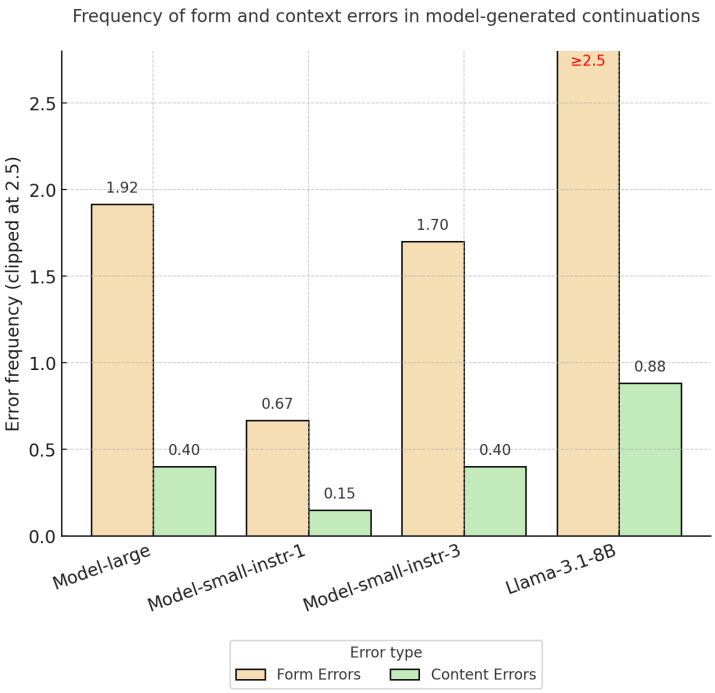


Figure 10: Normalized frequency of form and content errors in model-generated continuations. Form errors represent individual incorrect words, while content errors classify entire continuations as erroneous based on meaning. Frequencies are normalized by the total number of evaluated continuations (60) per model. Form error frequencies indicate the average number of incorrect words per generated continuation, which can exceed 1.0. In contrast, content error frequencies represent the proportion of continuations containing at least one content error, making them inherently bounded between 0 and 1.0. The y-axis is clipped at 2.5 for readability, with higher values marked accordingly.

C Corpora composition

	Model-large	Model-small-1	Model-small-2	Model-small-instr-1	Model-small-instr-2	Model-small-instr-3
Plain Text						
Galician	2570	232	232	232	232	232
Portuguese	3000	29	47	29	29	29
English	3500	29	35	29	29	29
Spanish	3390	29	16	29	29	29
Catalan	3390	29	16	29	29	29
Instructions	-	-	-	5.5	34.5	30

Table 1: Corpus distribution by language and type of text used to train the models presented in 3.1 (in million tokens).

D CPT configurations

To pretrain the models on Galician, we maintained consistent hyperparameter configurations across all experiments, except for model-large, where some adjustments were made due to its larger corpus size. Models model-small-1,2 and model-small-instr-1,2,3 were trained using two nodes, each with two NVIDIA A100 GPUs, while model-large was trained on five such nodes. To distribute the training load efficiently, we utilized DeepSpeed (Rajbhandari et al., 2020) with ZeRO stage 2. The effective batch size was 320 for model-large and 128 for the small models. The learning rate started at 10^{-4} and decayed linearly. All models were trained for a single epoch with a fixed sequence length of 2048 tokens. Finally, training was conducted using BF16 mixed precision.

E Evaluation results using lm-eval

The following table below (Table 2) presents the evaluation results obtained with lm-eval across multiple datasets. Model names were shortened due to space constraints, with these equivalences: Salam = Salamandra, Model Lg = Model-large, Model Sm = Model-small and Model SmI = Model-small-instr. New datasets introduced in the present work are highlighted in blue.

	Carballo Bloom	Salam 2B	Salam 7B	Llama 3.1-8B	Model Lg	Model Sm1	Model Sm2	Model SmI1	Model SmI2	Model SmI3
Multiple-choice tasks (accuracy)										
belebele_gl	0.231	0.229	0.374	0.807	0.320	0.431	0.272	0.563	0.293	0.500
galcola	0.498	0.497	0.533	0.588	0.524	0.489	0.488	0.553	0.508	0.576
openbookqa_gl	0.258	0.264	0.332	0.316	0.308	0.298	0.198	0.310	0.294	0.324
parafrases_gl	0.571	0.561	0.558	0.626	0.565	0.588	0.558	0.571	0.578	0.561
paws_gl	0.533	0.514	0.603	0.667	0.609	0.629	0.488	0.655	0.605	0.628
truthfulqa_gl (mc1)	0.257	0.235	0.228	0.278	0.235	0.274	0.277	0.255	0.255	0.268
truthfulqa_gl (mc2)	0.358	0.339	0.328	0.383	0.332	0.371	0.381	0.351	0.346	0.361
xnli_gl	0.480	0.478	0.505	0.501	0.500	0.496	0.398	0.520	0.477	0.509
xstorycloze_gl	0.624	0.619	0.736	0.680	0.713	0.656	0.509	0.690	0.643	0.686
Text-generation tasks (BLEU)										
summarization_gl	1.308	2.031	2.308	7.992	0.281	3.256	0.486	3.978	2.634	4.0198
truthfulqa_gl (gen)	0.858	0.453	9.219	13.734	1.182	0.411	0.662	0.411	0.705	7.326
flores_gl	11.763	9.086	12.826	2.579	20.771	14.664	0.711	22.669	16.317	3.110
Exact match tasks (accuracy)										
mgsm_direct_gl	0.000	0.024	0.028	0.060	0.044	0.024	0.016	0.020	0.036	0.040

Table 2: Evaluation results using lm-eval.

F Additional information about datasets

In addition to existing benchmark datasets, this work introduces three new Galician datasets as part of our contributions: *xnli_gl*, *xstorycloze_gl*, and *calame_gl*.

- **xnli_gl**: The Galician extension of the XNLI dataset, designed for evaluating cross-lingual sentence classification and transfer learning. The task involves classifying sentence pairs (a premise and a hypothesis) into one of three semantic categories: entailment, contradiction, or neutral. The dataset was automatically translated and subsequently revised by professional translators. It contains 5,009 test entries.
- **xstorycloze_gl**: A Galician adaptation of XStoryCloze, a commonsense reasoning dataset used for evaluating story comprehension, story generation, and script learning. Initially translated automatically from English, it was later reviewed and corrected by professional translators. It includes 1,841 entries, with 360 designated for training and 1,511 for testing.
- **calame_gl**: The Galician subset of the Calame-pt dataset, composed of short context passages with their respective final words. The task is designed to assess whether a human or a model can accurately predict the missing word while ensuring the context remains neither too specific nor overly ambiguous. This subset contains 931 entries, forming a portion of the original dataset. The dataset was initially translated from Portuguese to Galician and subsequently revised by professional translators.

Dataset Name	Type	Total Entries	Train	Test	Dev
belebele_gl	Machine Reading Comprehension (Multiple-Choice)	900	-	-	-
galcola	Linguistic Acceptability	17,088	11,957	1,710	3,418
openbookqa_gl	Question Answering (Open-Book)	1,000	-	500	500
parafrases_gl	Paraphrase Detection	2,935	2,053	295	587
paws_gl	Paraphrase Adversarial Testing	2,000	-	2,000	-
truthfulqa_gl	Truthfulness Evaluation (QA)	1,634	-	817	817
mgsm_direct_gl	Multi-Step Mathematical Reasoning	258	8	250	-
summarization_gl	Summarization (News Articles)	80,000	-	-	-
flores_gl	Machine Translation Evaluation	2,009	-	997	1,012
xnli_gl	Cross-Lingual Sentence Classification	5,009	-	5,009	-
xstorycloze_gl	Commonsense Reasoning (Story Comprehension)	1,841	360	1,511	-
calame_gl	Cloze Test (Context Completion)	931	-	-	-

Table 3: Overview of the evaluation datasets used in our experiments, including their dataset type, number of entries, and partitions.

G Instruction datasets

Language	Dataset Name	Type	Entries	Creation Method
GL	EGU (Enciclopedia Galega Universal)	Encyclopedic Knowledge	47,396	Manually Adapted
GL	RAG (Real Academia Galega)	Definitions	47,845	Manually Adapted
GL	MT (GL - ES)	Translations	275,292	Manually Adapted
GL	MT (GL - EN)	Translations	421,974	Manually Adapted
GL	SLI NER	Named Entity Recognition	8,138	Manually Adapted
GL	GalCoLA	Orthographic Correction	8,160	Manually Adapted
GL	SLI PoS TAGGING	Morphological Analysis	46,864	Manually Adapted
GL	Wikipedia Multiple-Choice QA	QA Multiple-choice	1,486	LLM-Generated
GL	CódigoCero Summarization	Summarization	342	LLM-Generated
CA	Parafraseja	Paraphrase Detection	21,984	Manually Adapted
CA	CASSA	Sentiment Analysis	6,400	Manually Adapted
PT	Wikipedia Multiple-Choice QA	QA Multiple-choice	547	LLM-Generated
PT	Extraglue-Instruct (Boolean Questions)	QA Simple	28,281	Manually Adapted
PT	Extraglue-Instruct (CB)	Concept Bottleneck	1,500	Manually Adapted
PT	Extraglue-Instruct (MultiRC)	Reading Comprehension	108,972	Manually Adapted
PT	Extraglue-Instruct (STSB)	Text Similarity	22,996	Manually Adapted
PT	Extraglue-Instruct (WNLI)	NLI (Inference)	3,810	Manually Adapted
PT	Aya (Train)	QA Simple	8,997	Manually Adapted
PT	OpenAssistant	Chat / Assistant	287	Manually Adapted
EN	Natural Instructions - NER	Named Entity Recognition	1,574	Manually Adapted
EN	QASC	QA Multiple-choice	9,980	Manually Adapted
EN	OpenAssistant	Chat / Assistant	154	Manually Adapted
ES	ALEXSIS	Linguistic Simplification	3,918	Manually Adapted
ES	COAH	Sentiment Analysis	1,816	Manually Adapted
ES	COAR	Sentiment Analysis	2,202	Manually Adapted

Table 4: Overview of instruction datasets used in our experiments, including their language, dataset type, number of entries, file size, and creation method. The datasets were created using two approaches: (1) adapting existing datasets or corpora by modifying their format to make them suitable for instruction pretraining while preserving their original intent, and (2) generating new datasets from scratch using *Salamandra-7B-Instruct*, leveraging its ability to synthesize diverse instruction-following examples. Due to license restrictions, *Parafraseja*, *CASSA*, and *RAG (Real Academia Galega)* cannot be publicly released.