Journal of the Brazilian Computer Society, 2025, 31:1, doi: 10.5753/jbcs.2025.5766

Enhancing Large Language Models for Underrepresented Varieties: Pretraining Strategies in the Galician-Portuguese Diasystem

Pablo Rodríguez [**University of Santiago de Compostela** | ***pablorodriguez.fernandez@usc.gal***]
Pablo Gamallo [**University of Santiago de Compostela** | ***pablo.gamallo@usc.gal***]
Daniel Santos [**University of Évora** | ***dfsantos@uevora.pt***]
Susana Sotelo [**University of Santiago de Compostela** | ***susana.sotelo.docio@usc.gal***]
Silvia Paniagua [**University of Santiago de Compostela** | ***silvia.paniagua.suarez@usc.es***]
José Ramom Pichel [**University of Santiago de Compostela** | ***jramon.pichel@usc.gal***]
Pedro Salgueiro [**University of Évora** | ***pds@uevora.pt***]
Vítor Nogueira [**University of Évora** | ***vbn@uevora.pt***]
Paulo Quaresma [**University of Évora** | ***pq@uevora.pt***]
Marcos Garcia [**University of Santiago de Compostela** | ***marcos.garcia.gonzalez@usc.gal***]
Senén Barro [**University of Santiago de Compostela** | ***senen.barro@usc.gal***]

Centro de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Santiago de Compostela, Galicia, Spain.

Received: 28 March 2025 • **Accepted:** 17 July 2025 • **Published:** 14 October 2025

Abstract This study presents a systematic exploration of strategies for pretraining generative Large Language Models (LLMs) within the Galician-Portuguese diasystem, by focusing on two underrepresented varieties of this diasystem, namely European Portuguese and Galician. We investigate the impact of combining versus separating linguistic varieties during continued pretraining, the trade-offs between large-scale noisy data and smaller high-quality corpora, and the potential gains from incorporating instruction-based data during the training phase instead of in post-training (e.g., instruction tuning). Our findings show that the inclusion of language varieties in training enhances both task-solving performance and linguistic quality in text generation, especially when leveraging curated linguistic resources. By integrating technical experimentation with sociolinguistic insight, this work underscores the importance of equitable and context-aware LLM development in multilingual and minority-language settings.

Keywords: Large Language Models, Continual Pretraining, European Portuguese, Galician

1 Introduction

Large language models (LLMs) are essential for processing text across language varieties. While it is well-established that factors like the proportion of pre-training data and model size influence performance, additional critical elements can significantly impact their effectiveness, such as linguistic characteristics of the corpora on which they are trained [Bagheri Nezhad *et al.*, 2025], including whether the LLM is able to represent in a balanced way all the varieties of each of the languages it models [Grieve *et al.*, 2025]. Languages with multiple varieties pose unique challenges for model development, as training data often disproportionately represent the most hegemonic varieties while underrepresenting others. This issue is particularly evident in the case of the Portuguese language, where Brazilian Portuguese tends to dominate large-scale datasets, often at the expense of European Portuguese, African varieties of Portuguese, and Galician.

At the crossroads between sociolinguistics and linguistic technologies, a fundamental question remains underexplored: What is actually being modeled by LLMs? Since each language is an abstract entity that manifests and materializes itself through its varieties, LLMs, and language models in general, are actually models of dynamically changing varieties rather than of static languages fixed to a single norm. From a sociolinguistic perspective, LLMs should model all varieties in a balanced way. However, two problems arise from the data itself. On the one hand, LLMs are trained on large, often poorly defined corpora (e.g., CommonCrawl), making it unclear which specific language varieties they are modeling. On the other hand, as mentioned above, LLMs inherently learn to represent the dominant varieties present in their training data, leading to an overrepresentation of the most hegemonic varieties within a diasystem. In contrast, less common varieties are either absent or underrepresented, for instance, Spanish of Costa Rica, English of India, or European Portuguese and Galician vs. Portuguese of Brazil. This sociolinguistic perspective provides crucial insights for the development, evaluation, and deployment of LLMs [Grieve *et al.*, 2025].

This article analyzes the performance of different generative models trained on two underrepresented varieties of Portuguese, namely European Portuguese and Galician. The goal is to compare their text generation and task resolution efficiency on datasets created explicitly for these linguistic varieties. Through a systematic methodology, we investigate the most effective strategies for continued pretraining of LLMs in the Galician-Portuguese diasystem, ensuring improved language generation performance and the mitigation of catastrophic forgetting in solving tasks. In sum, our study aims to identify the most effective training strategies for developing generative LLMs that can handle the linguistic nuances of the Galician-Portuguese diasystem while maintaining robust task-solving capabilities.

To address these issues, this study systematically examines the impact of different pretraining strategies on LLM performance across European Portuguese and Galician. Our approach explores four key research questions: (RQ1) Is it more effective to combine Galician and European Portuguese into a single model, or should separate models be trained for each variety? (RQ2) Does a large, noisy corpus yield better results than a smaller, high-quality corpus? (RQ3) Should instruction-based training be incorporated during the pretraining phase to enhance performance prior to instruction tuning? (RQ4) Does continued pretraining of models with underrepresented varieties improve the ability to generate text in these varieties with respect to the base multilingual LLM? By answering these questions, our aim is to contribute to the advancement of language modeling for underrepresented language varieties and to provide insights into optimal pretraining methodologies for linguistic diasystems.

Among the main contributions of this work, we highlight: i) the training of a new generative model of Portuguese, specialized in European Portuguese and Galician, by continued pretraining on the multilingual LLM Llama 3.1-8B; ii) the creation of high-quality corpora in both language varieties, and iii) the construction of instruction datasets in both varieties. Models[1], corpora and datasets[2] are available under free licenses.

The remainder of this paper is organized as follows: Section 2 explores several papers on LLMs focusing on the Portuguese diasystem, but also on Iberian languages. Then, in Section 3, we will briefly describe the close relationship between Galician and the Lusophony in order to justify the inclusion of Galician in the Portuguese diasystem. Section 4 details our methodology, including training data selection, model configurations, and evaluation strategies. Section 5 presents the experiments conducted to compare different training strategies and discusses the results, highlighting the effectiveness of different approaches. Finally, Section 6 concludes the paper and outlines future research directions.

2 Related Work

This section reviews existing LLMs created to represent the Portuguese diasystem and the broader cultural context of Iberian languages. Recent studies highlight that language models inherently represent specific linguistic varieties rather than singular standardized forms. Frequently, dominant language varieties become overrepresented in standard training datasets, causing underrepresentation of minority or less dominant varieties, and thus raising important issues regarding linguistic fairness and potential biases. Understanding these sociolinguistic dynamics is essential for the effective development, assessment, and use of language models [Grieve *et al.*, 2025]. Beyond technical concerns, [Helm *et al.*, 2024] introduce the concept of language modeling bias, arguing that current language technologies are often shaped by structural design choices that systematically privilege certain languages over others. This bias can lead to epistemic injustice, particularly when marginalized language communities are excluded from meaningful self-representation in AI systems. Their analysis highlights the importance of moving beyond merely increasing language coverage and instead embracing approaches grounded in linguistic diversity and co-design with affected communities. Additionally, even in multilingual models, effective generalization across closely related language varieties is not guaranteed. [Bafna *et al.*, 2024] show that LLMs exhibit performance degradation on unseen closely-related languages and dialects relative to their high-resource language neighbors. Their findings underscore the necessity of explicitly modeling diasystems and evaluating performance at the variety level to mitigate such performance degradation.

Several recent academic works have introduced generative models adapted explicitly for Iberian languages, notably Portuguese and Galician. For example, GlorIA is an openly available generative model developed using a comprehensive dataset consisting of European Portuguese texts from various domains, including literature, news, and formal publications, aimed at enhancing linguistic accuracy and representational balance for this particular variety [Lopes *et al.*, 2024]. Likewise, the Gervasio-7B-ptpt model aims to advance text-generation performance for European Portuguese by undergoing additional pretraining on carefully selected regional corpora, resulting in improved linguistic coherence and more precise language representation compared to broader multilingual or general Portuguese-focused models [Santos *et al.*, 2024]. Additionally, the Carvalho pt-gl-1.3B model investigates the effectiveness of simultaneously training generative models on both Galician and Portuguese datasets, capitalizing on their linguistic proximity. This work, specifically aims to enhance the model's performance and linguistic coverage by creating a unified corpus of these two closely linked language varieties [Gamallo *et al.*, 2024b]. Likewise, the Carballo-Bloom model was developed solely for Galician, utilizing a carefully selected corpus designed to accurately represent Galician linguistic features and cultural elements. By concentrating exclusively on Galician, this model provides a strong baseline for assessing generative capabilities related to this distinct linguistic variety [Gamallo *et al.*, 2024a].

Beyond models specifically targeting single or closely related linguistic varieties, multilingual generative models have also emerged, aiming for broader linguistic coverage. An influential example is BLOOM, a large-scale open-access multilingual language model trained on 46 languages, including Portuguese but not Galician. Despite this exclusion, BLOOM demonstrated strong performance on translation quality of Galician, likely benefiting from linguistic transfer due to Galician's proximity to Portuguese and other Romance languages included in the training corpus [Scao *et al.*, 2022]. Similarly, the Llama family of models has recently become influential as a multilingual foundation, particularly due to

[1] `https://huggingface.co/Nos-PT`

[2] `https://github.com/proxectonos/instruction_datasets`

their extensive multilingual training data and capability for efficient adaptation to specific languages and tasks [Touvron *et al.*, 2023]. It is also important to highlight the Salamandra model family [Gonzalez-Agirre *et al.*, 2025], pretrained on 35 languages spoken across Europe, including both Galician and European Portuguese. Although these two varieties represent a relatively small portion of the overall training data, their inclusion is proportionally higher than in most existing multilingual models, making Salamandra a notable contribution to the representation of lesser-resourced Iberian languages.

Despite these important advances, critical open questions remain regarding optimal training strategies, corpus quality considerations, and the representation of specific linguistic varieties within generative language models. These challenges are especially pronounced for closely related but distinct varieties such as Galician and European Portuguese. Addressing precisely these concerns, our research investigates effective pretraining strategies and evaluates their impact on generative performance for these underrepresented language varieties.

3 Language Varieties of the Portuguese Diasystem

Galician and Portuguese are two different names for the same linguistic heritage [Carvalho, 1979; Cintra, 1971], with the former reflecting its kingdom of origin and the latter representing the kingdom that historically expanded it worldwide, similar to the distinction between Castilian and Spanish. Both varieties trace their roots to medieval Galician-Portuguese [Lopes, 2010] and are spoken in different states: Galician in the autonomous region of Galicia within the Kingdom of Spain, and Portuguese in the Republic of Portugal as well as in eight other countries, each with its own linguistic characteristics.

Despite some divergences between European Portuguese and Galician, mainly due to Spanish influence on the latter, both Romance varieties share grammatical features differentiating them from other Ibero-Romance languages [Duarte, 2024]. They share significant phonological, lexical, and grammatical similarities, forming a closely related linguistic continuum in the western part of the Iberian Peninsula.

These unique characteristics have led to various perspectives on the linguistic relationship between Galician and European Portuguese, shaped by historical and political factors [Herrero-Valeiro, 2003; Collazo, 2014]. These influences have played a significant role in the evolving configuration of Galician, which has varied in its similarity to Portuguese or Castilian across different historical periods and standardization processes [Pichel *et al.*, 2021].

The view advocating for a linguistic reunion, known as the reintegrationist movement [Durão, 2008; Paz Felix, 2020], argues that Galician is a variety of the language known globally as Portuguese [Muhr, 2013; Dayán-Fernández and O'Rourke, 2020], emphasizing their shared medieval origin and high mutual intelligibility. Supporters of this perspective call for aligning Galician more closely with Portuguese orthographic and linguistic norms, strengthening its connection to the Lusophone world. This approach suggests that Galician and Portuguese should be treated as a single language in computational models and educational policies.

In contrast, the linguistic segregation stance, supported by institutions such as the Galician autonomous government and the Royal Galician Academy, regards Galician as a distinct language [Ramallo and Rei-Doval, 2015] that evolved separately under the influence of Spanish [Monteagudo and Santamarina, 1993]. While acknowledging the historical connection with Portuguese, this view highlights the lexical, phonetic, and grammatical differences that justify Galician's independent linguistic identity. Galician's official orthography follows Spanish-influenced norms, rather than those of Portuguese, and linguistic policies in Galicia often emphasize preserving a unique Galician identity.

As none of these academic views is completely hegemonic, a diasystemic approach [Romero, 1999; del Olmo and da Cunha, 2017] offers a middle ground for Galician-Portuguese strategies in natural language processing, recognizing Galician and European Portuguese as closely related yet distinct varieties within a linguistic continuum. This perspective acknowledges mutual intelligibility and shared features while also considering the differences that have emerged over time. In language modeling, the fact of considering these two varieties as part of an inclusive diasystem leads to conceive training strategies that favor linguistic transfer. This contrasts with the vision of separate languages which leads to strategies requiring separate training data for independent and monolingual continued pretraining.

Our research highlights the importance of completing the Portuguese diasystem by giving more presence and visibility to two of its closest varieties, Galician and European Portuguese, which will have an impact on the improvement of generative AI applications for the Portuguese language.

4 Methodology

In this section, we present a comprehensive methodology for building generative LLMs for the Galician-Portuguese diasystem. Our approach focuses on exploring the optimal strategies for continued pretraining LLMs that can effectively give rise to models with the following characteristics: on the one hand, they improve the generation capabilities in these two language varieties with regard to the base model on which the continued pretraining is performed and, on the other, they prevent catastrophic forgetting as they do not forget the strong ability to solve tasks underlying the base LLM. To this end, we investigate several key research questions: Is it more effective to combine Galician and European Portuguese varieties into a single model, or should separate models be trained for each variety? Does a large, noisy corpus yield better results than a smaller, high-quality corpus? Should instruction-based training be incorporated during the pretraining phase, prior to instruction tuning?

In the following subsections, we provide a detailed description of the methodologies employed, including the construction of training corpora, instruction datasets, and evaluation frameworks. This systematic exploration aims to identify the most effective strategies for developing generative LLMs for the Galician-Portuguese diasystem, contributing to the advancement of natural language processing for underrepresented languages varieties.

4.1 Strategies for Continued Pretraining

This subsection details our methodology for continued pretraining, which leverages diverse datasets for Galician and European Portuguese to determine the optimal training configurations. Specifically, we investigate three key factors: (i) whether training a single model on both varieties yields better results than training them separately, (ii) whether a large, noisy corpus is more effective than a smaller, high-quality one, and (iii) whether incorporating instruction-based data before instruction tuning enhances performance.

I. Variety Combination
We examine whether it is more effective to train a single model on both Galician and European Portuguese data or to train separate models for each variety. These two approaches reflect different priorities: capturing cross-variety generalizations versus preserving linguistic specificity.

- **Combined models:** Models trained on data from both varieties, aiming to learn shared linguistic patterns across the diasystem.
- **Separate models:** Models trained individually on Galician or European Portuguese, focusing on the unique characteristics of each variety.

II. Corpus Size and Quality
This dimension investigates whether model performance benefits more from larger but noisier data or from smaller, higher-quality corpora. The trade-off here lies between data quantity and linguistic precision.

- **Large, noisy corpora:** Models trained on corpora collected from diverse web sources, such as social media and online forums. While these corpora provide broad coverage and linguistic diversity, they may contain noise, inconsistencies, and lower-quality text.
- **Small, high-quality corpora:** Model trained on carefully curated datasets comprising literary works, academic publications, news articles, and official documents. Although smaller in size, these corpora are characterized by their linguistic accuracy, coherence, and cultural relevance, ensuring high-quality input for pretraining.

III. Instruction-Based Pretraining
We also assess the effect of introducing instruction-based data during pretraining, prior to any instruction tuning. This tests whether early exposure to task-oriented language improves downstream performance, bringing out the emergent ability to solve tasks in few-shot scenarios.

- **With instruction data:** Models are exposed to prompts and task structures during pretraining, enabling them to learn how to follow prompts and solve tasks more effectively. This approach aims to enhance the models' ability to generalize across diverse tasks and improve their performance in instruction-following scenarios.
- **Without instruction data:** Models follow a traditional pretraining approach, focusing solely on language modeling without explicit task-specific instructions. This baseline allows us to evaluate the impact of instruction-based pretraining on model performance.

This systematic approach enables us to identify the most effective strategies for training generative LLMs in the Galician-Portuguese diasystem.

4.2 Evaluation Strategies

The performance of these models is evaluated using two complementary methodologies: one focuses on assessing the linguistic quality of the generated text, while the other evaluates the ability of the models to solve diverse tasks.

4.2.1 Linguistic Quality of Text Generation

To assess the linguistic quality of the text generated by the models, we employ two distinct strategies:

I. Automatic Evaluation
This approach relies on specific datasets and metrics designed to evaluate the quality of the generated language. The model is asked to predict a word based on the preceding context and we employ the *surprisal* metric to measure the degree of semantic coherence of the new generated token. Surprisal captures how unexpected or surprising a token is within its context in the sequence. This makes it particularly useful for assessing the sensitivity of the model to linguistic plausibility.

II: Qualitative Evaluation
This evaluation is conducted by human annotators to assess the correctness and coherence of the text generated by the LLMs. The models are prompted to generate a specific number of tokens following a text, as in an autocomplete task. The generated text is evaluated across two dimensions:

- **Formal Correctness**: The text is considered formally correct if it is grammatically accurate and free of spelling errors.
- **Content Coherence**: The text is deemed semantically coherent if it is contextually relevant to the preceding text and maintains a consistent language register.

Both dimensions are assessed using a binary evaluation scheme (e.g., correct/incorrect or coherent/incoherent). Annotators are provided with clear guidelines to ensure consistency and reliability in their assessments.

4.2.2 Task-Solving Performance

This evaluation measures the model's ability to solve tasks: automatic summaries, translation, or question answering. The evaluation is performed in two platforms: LM Evaluation Harness (lm-eval) [Gao *et al.*, 2024] and Simil-Eval [Rodríguez *et al.*, 2025], which provide standardized tasks for assessing model capabilities in diverse contexts. Both platforms rely on the datasets from the IberoBench benchmark [Baucells *et al.*, 2025], originally designed for lm-eval and adapted in some tasks for Simil-Eval. For multi-choice question-answering tasks, lm-eval relies on probabilistic log-likelihood metrics, while Simil-Eval requires models to generate answers to compute the semantic similarity (using contextualized embeddings) between the generated answer and the correct answer, measured with cosine similarity metrics. The strategies used by these two evaluation platforms can be seen as complementary, which helps to have a more solid view of the behavior of the models in this type of tasks.

4.3 Datasets and Corpora

In this subsection, we describe the datasets and corpora used for training, instruction pretraining, and evaluation. These resources fall into three main categories: training corpora, which support both unsupervised and instruction-based pretraining; instruction datasets, which provide task-oriented examples across multiple languages; and evaluation datasets, which are used to assess model performance on both generative and task-specific benchmarks.

4.3.1 Training Corpora

We divide the training corpora into two main categories, each consisting of multiple corpora (see Table 1 for their composition).

Large, noisy corpora. This category includes corpora composed of both crawled and curated web data covering a wide range of domains and languages. It encompasses the *Multilingual Noisy* corpus and its subset, the *PT-GL Noisy* corpus, both designed to provide high-volume, diverse training material despite potential variation in text quality.

The **Multilingual Noisy** corpus is a large-scale, heterogeneous collection of texts in Galician, Portuguese, Spanish, Catalan, and English, compiled from a wide range of sources and genres. It integrates both professionally edited content and automatically collected material, such as web-scraped texts and informal writing. This composition supports a broad linguistic and stylistic range, including formal and informal registers, factual and narrative modes, and domain coverage across politics, science, culture, and everyday discourse. The corpus was constructed to maintain a balanced token distribution across the five languages, while also accounting for variations in data availability and source characteristics. For Galician, the core of the corpus is derived from CorpusNÓS [de Dios-Flores *et al.*, 2024], a large-scale compilation of texts spanning journalistic, literary, and institutional domains. For Portuguese, the main sources include CETEMPúblico [Santos and Rocha, 2000] and Arquivo.pt [Gomes *et al.*, 2009], the Portuguese national web archive. While both CorpusNÓS and CETEMPúblico contain primarily structured and high-quality texts, Arquivo.pt, despite undergoing cleaning and filtering, retains typical challenges of web-crawled corpora, such as residual noise, domain imbalance, and duplicated material. The Spanish and English components include a wide range of structured resources, such as literary corpora, encyclopedic entries, academic publications, and parliamentary records, supplemented by selected user-generated content like reviews. The Catalan portion draws from a variety of domains, such as news outlets, online forums, digital libraries, and institutional websites. However, only a filtered, high-quality subset of this content was included to ensure consistency.

The **PT-GL Noisy** corpus is a targeted subset of the multilingual corpus, designed to prioritize Galician and Portuguese while preserving limited exposure to Spanish and English. Catalan was excluded entirely, as it was not a target language for the subsequent models and was not present in the original training distribution. The decision to retain Spanish and English, although in drastically reduced proportions, was intended to maintain compatibility with the initial multilingual setup while preventing these high-resource languages from dominating the training dynamics. The Galician and Portuguese portions were downsized relative to the multilingual corpus, but their proportional balance was preserved. For Portuguese, the same sources were reused, with Arquivo.pt undergoing a second filtering stage to extract smaller, higher-quality subsets. This process involved perplexity-based selection, using thresholding to discard low-quality content. Very low perplexity scores did not consistently indicate high-quality texts, while high scores often flagged noisy or incoherent material; therefore, texts with intermediate scores were selected as a more reliable proxy for quality. This filtering task was performed with the pyplexity library [Fernández-Pichel *et al.*, 2024]. Additional cleaning was performed using regular expressions to remove recurring undesired patterns identified in the corpus. The Galician portion likewise consists of a carefully selected subset of previously used corpora, with strong representation in journalistic and institutional domains. For Spanish and English, only general-domain, high-quality corpora were retained. This corpus also includes instruction datasets in Portuguese, Galician, Spanish, and English. It remains classified as noisy due to two aspects: first, the application of a data augmentation strategy, in which content was replicated multiple times to match target token counts and enhance exposure to the core languages. Second, it contains a significant proportion of text from web scraping, which, although filtered, maintains structural problems that are very difficult to detect.

Small, high-quality corpora. Unlike the large, noisy sets, these corpora are composed of carefully curated texts selected for their linguistic quality and domain consistency. Drawn from sources such as literature, academia, and official publications, this category includes the *PT-GL High-Quality*, *GL High-Quality*, and *PT High-Quality* corpora, each designed to support focused training on high-quality language data.

The **PT-GL High-Quality** corpus builds on the same sources as the noisy version but removes data augmentation and applies stricter filtering with pyplexity (a lower threshold). It retains the same language composition, with Galician and Portuguese as the main focus and reduced contributions from Spanish and English. In addition, it introduces a significantly larger volume of instruction data in Portuguese while preserving smaller amounts in the other languages.

The **GL High-Quality** corpus adapts the PT-GL configuration to center on Galician. The Galician portion is retained in full, while Portuguese, Spanish, and English are downsized to equivalent volumes. Its instruction component mirrors that of the PT-GL Noisy corpus, with Galician having the

	Multilingual Noisy	PT-GL Noisy	PT-GL High Quality	GL High Quality	PT High Quality
Plain text					
Galician	2570	1160	232	232	29
Portuguese	3000	1500	250	29	250
English	3500	145	29	29	29
Spanish	3390	145	29	29	29
Catalan	3390	-	-	-	-
Instructions					
	-	28.5	72.2	28.5	72.2

Table 1. Corpus used to train the models, detailing its composition by language and number of instructions (in millions of tokens).

largest share, and Portuguese, Spanish, and English present in smaller amounts.

The **PT High-Quality** corpus, in turn, focuses on Portuguese. It maintains the full Portuguese dataset and uses the same instruction data as the PT-GL High-Quality corpus, where Portuguese constitutes the largest portion, followed by Galician, with minimal Spanish and English. This configuration supports Portuguese-centered modeling while preserving limited multilingual robustness.

4.3.2 Instruction Datasets

The instruction datasets used in this work were created through a combination of synthetic generation, adaptation of existing resources, and repurposing of general-purpose corpora. These approaches resulted in three main types of datasets, which differ in their source, structure, and method of construction, and are described in detail below.

Model-generated datasets: A subset of the instruction corpus was synthetically created using Salamandra-7B-Instruct[3], including datasets for multiple-choice question answering in Portuguese and Galician, as well as a Galician summarization dataset. While useful for aligning the models with instructional prompts, this method may introduce limitations or biases related to the generative behavior of the model used for dataset creation.

Repurposed datasets: These are resources originally not intended for instruction-based training, but adapted into an instruction–response format. They include corpora for morphological analysis, named entity recognition, sentiment analysis, linguistic simplification, definition generation, and simple question answering. Although designed for traditional supervised learning, these datasets were reformulated by pairing inputs with task-specific prompts to simulate instructional settings.

Instruction-aligned datasets: Datasets originally developed for instruction-based tasks were adapted to match the instruction–response format used during training. These datasets contributed examples for chat-style interactions, multi-turn reasoning, reading comprehension, natural language inference (NLI), textual similarity, and general QA supervision.

Instruction data was predominantly composed of European Portuguese (approximately 60%) and Galician (around 36%), reflecting the central focus of this work and the relative availability of resources in these language varieties. Spanish and English accounted for less than 4% combined, providing minimal but intentional coverage to support some degree of multilingual generalization. A detailed overview of the instruction datasets, including their language, task type, number of entries, and creation method, is presented in Table 2.

4.3.3 Evaluation Datasets

To evaluate our language models, we will make use of two types of datasets: those used as standardized benchmarks from IberoBench to assess performance on tasks such as text classification, question answering, and summarization, among others; and CALAME, a dataset used to measure the generative ability of models, involving surprisal and human judgment of the quality and coherence of open-ended text generation [Lopes *et al.*, 2024].

Task-based datasets: Table 3 provides an overview of the benchmarks drawn from IberoBench (GalicianBench and PortugueseBench) used to evaluate structured tasks using lm-eval and Simil-Eval. This table provides details of the dataset names by language, associated task types, and the evaluation frameworks utilized. The datasets are categorized into task types such as multiple-choice, text generation, and exact match, with corresponding evaluation frameworks.

CALAME: For the evaluation of the linguistic quality of generated texts, we used the CALAME dataset [Lopes *et al.*, 2024], originally developed as a benchmark for European Portuguese and adapted into Galician through translation for this work. This dataset is used in two different types of model evaluation. First, the model is asked to predict a word based on the preceding context in CALAME. However, instead of using an exact match to verify whether the predicted word coincides with a reference, as proposed by the authors of the dataset, we employ the *surprisal* metric for a quantitative evaluation of the coherence of the generated word. Second, instead of predicting the final word from a given context, the texts of CALAME were used as prompts to elicit open-ended text generation. More precisely, the models are prompted to generate 50 tokens following each example from CALAME

[3] `https://huggingface.co/BSC-LT/salamandra-7b-instruct`

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

Table 2. Overview of the instruction datasets used in our experiments, including language, dataset type, number of entries, and creation method. Datasets were constructed through three main strategies: (1) generating synthetic examples using Salamandra-7B-Instruct, (2) repurposing existing supervised resources, and (3) aligning instruction-based datasets to our training format. Due to license restrictions, Parafraseja, CASSA, and RAG cannot be publicly released.

(100 examples per model and language variety). This adaptation enabled a qualitative evaluation of the generated outputs through human annotation, focusing on both formal correctness and content coherence.

By carefully curating and defining these datasets, we ensure a robust and reproducible framework for training and evaluating generative LLMs for the Galician-Portuguese diasystem.

5 Experiments

5.1 The models

During the experimentation, we will compare the models trained following the methodology presented with others, including Galician or Portuguese, during their training phase.

In relation to the models trained by us, all of them take Llama-3.1-8B[4] as the base model. Moreover, although the corpus used varies among them, almost all of them include instructions in Galician and Portuguese, and to a lesser extent, in English and Spanish:

- **Carballo-Llama**: Multilingual model trained on multilingual noisy corpus (see Multilingual Noisy column in Table 1), with all available corpus for Galician and Portuguese including noisy and high-quality sources, and a proportional amount of English and Spanish. This multilingual corpus does not contain any instructions.
- **Carvalho All**: Model trained on a corpus of Galician and Portuguese, including noisy and high quality sources, and a small amount of English and Spanish (see PT-GL Noisy column in Table 1). This model was provided with 28.5M tokens in instructions during continued pretraining.
- **Carvalho PT-GL**: Model trained on a high-quality corpus of Galician and Portuguese, with a small amount of English and Spanish (see PT-GL High Quality column in Table 1). This model was provided with 72.2M tokens in instructions during continued pretraining.
- **Carvalho GL**: Model trained on a high-quality corpus, with an emphasis on Galician over Portuguese, and a residual amount of English and Spanish (see GL High Quality column in Table 1). This model was provided with 28.5M tokens in instructions during continued pretraining.
- **Carvalho PT**: Model trained on a high-quality corpus, with an emphasis on Portuguese over Galician, and a residual proportion on English and Spanish (see PT High Quality column in Table 1). This model was provided with 72.2M tokens in instructions during continued pretraining.

The other models evaluated are the following

- **Carballo-Bloom**: 1.3B paramater model, based on

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

Language	Dataset	Task Type	Evaluation Framework
Galician	*Multiple-choice*		
	belebele_gl	Reading Comprehension	Simil-Eval, lm-eval
	galcola	Linguistic Acceptability	lm-eval
	openbookqa_gl	Question Answering	Simil-Eval, lm-eval
	parafrases_gl	Paraphrase Recognition	lm-eval
	paws_gl	Paraphrase Recognition	lm-eval
	truthfulqa_gl (mc1)	Factual Consistency	Simil-Eval, lm-eval
	xnli_gl	Natural Language Inference	lm-eval
	xstorycloze_gl	Commonsense Narrative Understanding	Simil-Eval, lm-eval
	Text Generation		
	summarization_gl	Summarization	lm-eval
	truthfulqa_gl (gen)	Factual Generation	lm-eval
	flores_gl	Translation	lm-eval
	Exact Match		
	mgsm_direct_gl	Mathematical Reasoning	lm-eval
Portuguese	*Multiple-choice*		
	belebele_pt	Reading Comprehension	Simil-Eval, lm-eval
	xstorycloze_pt	Commonsense Narrative Understanding	Simil-Eval
	assin_entailment	Natural Language Inference	lm-eval
	paws_pt	Paraphrase Recognition	lm-eval
	assin_paraphrase	Paraphrase Recognition	lm-eval
	Text Generation		
	flores_pt	Translation	lm-eval

Table 3. Overview of evaluation datasets from IberoBench (GalicianBench and PortugueseBench), organized by language, task type, and evaluation framework. The Portuguese datasets of PortugueseBench consists of datasets for European Portuguese.

FLOR-1.3B model[5], trained on Galician data.

- **Carvalho_pt-gl-1.3B**: 1.3B parameter model, based on Cerebras-1.3B[6], trained on a combination of Galician and Portuguese data.
- **GlorIA**: 1.3B parameter model, based on GPT-Neo[7], trained on 35B Portuguese tokens from different sources.
- **Gervasio-7B-ptpt**: Model based on LLaMA-2-7B enriched with Portuguese data. It has Brazilian and European Portuguese varieties, but we only analyze the European model.
- **Salamandra-7B**: 7.7B parameter multilingual model pretrained in 35 European languages, including Galician and Portuguese.
- **Llama-3.1-8B**: Base model for the training of our models, used as a baseline to compare the effectiveness of the methodology presented. Galician is not included in the pretraining, and Portuguese is in its Brazilian variety, not in the European one.

5.1.1 Continued Pretraining Configurations

To pretrain the different models, we maintained consistent hyperparameter configurations across all experiments, making minor adjustments to account for differences in corpus size and HPC cluster specifications. The models Carballo-Llama and Carvalho GL were trained at the *Galician Supercomputing Center (CESGA)*, using 10 and 8 NVIDIA A100 GPUs, respectively. Carvalho All and Carvalho PT-GL were trained on MareNostrum V at the *Barcelona Supercomputing Center (BSC)*, utilizing 20 NVIDIA H100 GPUs. Finally, Carvalho PT was trained on the VISION cluster at the *University of Evora (UEvora)*. The variation in training locations was due to computational and scheduling constraints at each center. However, this diversity in infrastructure contributed to the robustness of our methodology, demonstrating that the proposed approach is replicable across different hardware configurations.

To distribute the training load across computational nodes efficiently, we utilized DeepSpeed with ZeRO 2 stage [Rajbhandari *et al.*, 2020]. All models were trained for a single epoch with a fixed sequence length of 2048 tokens. Finally, training was conducted using BF16 mixed precision.

5.2 Results

We distinguish between purely automatic or quantitative evaluation and manual or qualitative evaluation, separating both types into two subsections.

5.2.1 Quantitative Evaluation

The automatic/quantitative assessment is also divided into two subtypes: task-based and surprisal-based evaluation. The second one directly assesses the quality of the generated text by making use of surprisal.

[5] https://huggingface.co/projecte-aina/FLOR-1.3B
[6] https://huggingface.co/cerebras/Cerebras-GPT-1.3B
[7] https://huggingface.co/EleutherAI/gpt-neo-1.3B

Task-Based Evaluation: This evaluation compares multiple LLMs across multiple-choice, text-generation (BLEU), and exact-match tasks for both European Portuguese and Galician. As introduced in Subsection 4.2.2, two evaluation platforms are used: lm-eval using IberoBench (GalicianBench and PortugueseBench) and Simil-Eval. The results provided by the former are shown in Tables 4 and 5. The results in Table 4 for Portuguese show that the two best models are clearly LLama-3.1-8.B, which dominates in belebele_pt (0.83 accuracy) and flores_pt (29.94 BLEU), and Carvalho PT-GL, the best model in three datasets: paws_pt (0.72), xstorycloze_pt (0.66), and assin_paraphrase (0.70), suggesting balanced Galician-Portuguese training benefits task-specific performance. It is also noticeable that Carvalho PT excels in assin_entailment (0.67), indicating Portuguese-focused training aids in semantic tasks. The results in Table 5 for Galician also show that the most competitive models are still Carvalho PT-GL, achieving the highest scores in galcola (0.60), parafrases_gl (0.64), paws_gl (0.72), and xnli_gl (0.54), being the second best model in other three tasks (underlined scores), and Llama-3.1-8B, which leads in belebele_gl (0.81), summarization_gl (7.99 BLEU), and truthfulqa_gl (mc2,gen). Salamandra-7B shows robustness in xstorycloze_gl (0.74) and openbookqa_gl (0.51), and Carvalho GL performs well (the second-best model) in three datasets, suggesting that a monolingual model with a single variety may still be competitive even if it does not perform at the same level as the model enriched with more varieties.

Figures 1 and 2 show the results obtained in the Simil-Eval platform. For both European Portuguese and Galician, the best model is Llama 3.1-8B, followed not far behind by Carvalho PT-GL and Salamandra 7B. Although the correlation is not exact with the IberoBench results, this evaluation complements and ratifies the previous one, confirming that Llama 3.1-8B and Carvalho PT-GL tend to have a more regular and stable behavior in the evaluated tasks.

Surprisal-Based Evaluation: To automatically measure linguistic quality in textual generation, we use the Calame dataset, initially conceived for European Portuguese [Lopes *et al.*, 2024] and partially translated into Galician. In this test, we measure the validity of the word proposed by the model, given the previous context. The lower the surprisal, the better the choice of that word. In this evaluation, the models with the best values were again Carvalho PT-GL, Llama 3.1-8B, and Salamandra 7B. Specifically, the best model in Galician is Carvalho PT-GL (surprisal 2.08), followed by Llama 3.1-8B (2.12) and Salamandra 7B (2.14). In Portuguese, the ranking is reversed: first, Salamandra 7B, followed by Llama 3.1-8B and Carvalho PT-GL. Notice that those models trained exclusively with the European variety, such as GlorIA and Gervasio 7B-ptpt, have a very high surprisal (5.14 and 4.79, respectively) in the Galician variant. This great disparity of surprisal does not occur with other monolingual models since their base is multilingual, for instance, Carballo Bloom, Carvalho GL, or Carvalho PT.

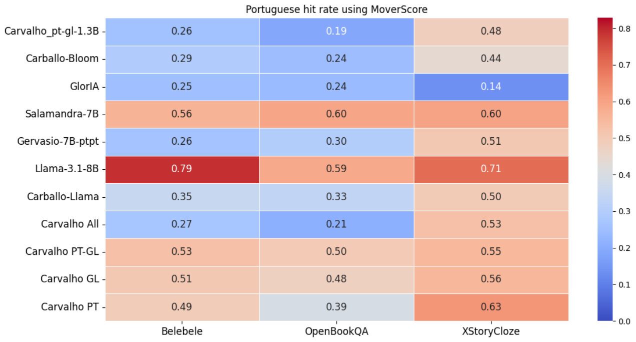

Figure 1. Portuguese hit rate using Simil-Eval

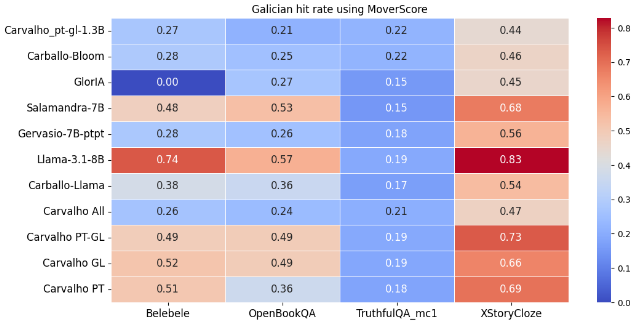

Figure 2. Galician hit rate using Simil-Eval

5.2.2 Qualitative Evaluation

As it was stated previously, the qualitative evaluation focuses on the quality of the text generated by the LLM, specifically on formal and content errors. A total of 400 texts were evaluated. 200 in European Portuguese and 200 in Galician. In each language, 100 were generated by Llama 3.1-8B and 100 by Carvalho PT-GL. The evaluators, three in total, 1 for the texts in European Portuguese and 2 for Galician, divided the texts without duplicate annotations, so there is no margin to evaluate the inter-rater reliability. Figure 4 shows the results of this evaluation. The results show that the two models generate better quality text in European Portuguese than in Galician. This is due to the fact that, on the one hand, this variety is more present in the base model and, on the other hand, it shares many normative and standardization aspects, especially in orthography, with the most represented variety, Brazilian.

If we focus on the form errors, we find that in both Galician and European Portuguese, but especially in Galician, the base model, Llama 3.1-8B, makes many more errors than Carvalho PT-GL. The main problem of the base model is that it mixes languages such as Spanish and English in many of its generations, especially when it tries to use the Galician variety. As for content errors (e.g., coherence and linguistic register), the differences between the two models are much smaller. This is because the base model already has the necessary knowledge to complete the text coherently, but lacks the ability to write correctly in the variety requested in the prompt.

It is very important to emphasize here that there is no correlation between the formal errors of this qualitative evaluation and the automatic evaluation based on surprisal. In fact, we consider that the surprisal-based evaluation measures, not the formal quality of the generated text, but rather the coherence and content of what is generated.

	Carv 1.3B	Carb Bloom	GlorIA	Salam 7B	Gerv 7B	Llama 3.1-8B	Carb Llama	Carv All	Carv PT-GL	Carv GL	Carv PT
Multiple-choice tasks (accuracy)											
belebele_pt	0.273	0.228	0.230	0.502	0.270	**0.832**	0.231	0.573	0.619	0.519	0.269
paws_pt	0.521	0.542	0.529	0.631	0.695	0.703	0.635	0.510	**0.724**	0.684	0.626
xstorycloze_pt	0.560	0.548	0.556	**0.675**	0.643	0.649	0.645	0.514	0.661	0.616	0.571
assin_entailment	0.603	0.592	0.621	0.646	0.540	0.572	0.549	0.413	0.570	0.603	**0.668**
assin_paraphrase	0.618	0.694	0.548	0.648	0.665	0.641	0.588	0.479	**0.702**	0.586	0.672
Text-generation tasks (BLEU)											
flores_pt	6.407	9.538	3.891	13.855	8.880	**29.945**	19.682	0.962	18.316	21.937	21.149

Table 4. Evaluation results for PortugueseBench using lm-eval. Model names were shortened due to space constraints, where Salam = Salamandra, Carb = Carballo, Carv = Carvalho, Gerv = Gervasio, Carv 1.3B = Carvalho_pt-gl-1.3B.

	Carv 1.3B	Carb Bloom	GlorIA	Salam 7B	Gerv 7B	Llama 3.1-8B	Carb Llama	Carv All	Carv PT-GL	Carv GL	Carv PT
Multiple-choice tasks (accuracy)											
belebele_gl	0.276	0.231	0.229	0.374	0.231	**0.807**	0.32	0.226	0.601	0.5	0.327
galcola	0.494	0.498	0.496	0.533	0.504	0.588	0.524	0.493	**0.599**	0.576	0.540
openbookqa_gl	0.234	0.258	0.222	**0.332**	0.25	0.316	0.308	0.206	0.324	0.324	0.282
parafrases_gl	0.541	0.571	0.537	0.558	0.561	0.626	0.565	0.575	**0.639**	0.561	0.602
paws_gl	0.514	0.533	0.470	0.603	0.641	0.667	0.610	0.503	**0.72**	0.628	0.602
truthfulqa_gl (mc1)	0.237	0.257	0.191	0.228	0.176	**0.278**	0.235	0.269	0.220	0.268	0.225
xnli_gl	0.449	0.480	0.349	0.505	0.425	0.501	0.500	0.397	**0.536**	0.509	0.516
xstorycloze_gl	0.598	0.624	0.453	**0.736**	0.540	0.680	0.713	0.541	0.712	0.686	0.662
Text-generation tasks (BLEU)											
summarization_gl	1.017	1.308	0.014	2.308	2.265	**7.992**	0.281	0.540	3.700	4.020	3.223
truthfulqa_gl (gen)	7.310	0.858	11.950	9.219	2.493	13.734	1.182	0.383	2.461	7.326	0.337
flores_gl	5.893	11.763	1.143	12.823	6.991	2.579	**20.772**	0.943	15.918	3.110	20.37
Exact match tasks (accuracy)											
mgsm_direct_gl	0	0	0	0.028	0	**0.06**	0.044	0.004	0.052	0.04	0

Table 5. Evaluation results for GalicianBench using lm-eval. Model names were shortened due to space constraints, where Salam = Salamandra, Carb = Carballo, Carv = Carvalho, Gerv = Gervasio, Carv 1.3B = Carvalho_pt-gl-1.3B.

5.3 Discussion

The discussion of the results will focus exclusively on answering the four questions raised in the introduction, which will be answered on the basis of the results obtained in the different experiments carried out.

[RQ1] *Is it more effective to combine Galician and European Portuguese into a single model, or should separate models be trained for each variety?* Among the three best-performing Carvalho models, results suggest that the model with the two varieties, Carvalho PT-GL, outperforms the best models with just one of the two varieties, Carvalho PT and Carvalho GL, in both Galician and Portuguese. This justifies the need for a dedicated Galician-Portuguese model since models that combine varieties work better than those that separate them.

[RQ2] *Does a large, noisy corpus yield better results than a smaller, high-quality corpus?* The results indicate that the models with small high-quality corpus, such as Carvalho PT-GL, Carvalho GL, and Carvalho PT perform no worse than other models trained with larger noisy corpora, such as Carvalho All, Carballo-Llama, Carballo-Bloom or Carvalho_pt-gl-1.3B in both task resolution and text generation. And if we focus on the best model trained with small high-quality corpus, Carvalho PT-GL, it outperforms all other models. Therefore, the answer to the question seems clear: The use in continued pretraining of a small corpus of high-quality is more useful than a larger but noisy model.

[RQ3] *Should instruction-based training be incorporated during the pretraining phase to enhance performance prior to instruction tuning?* Carvalho PT-GL, Carvalho GL, and Carvalho PT, which include instructions during their continued pretraining, are close to Llama3.1-8B in solving tasks. However, those models pretrained without explicit instructions, such as Carballo-Llama, Carballo-Bloom, or Carvalho_pt-gl-1.3B, tend to have difficulty not only responding appropriately to tasks but even understanding what to do to solve the task. So, including instructions in continued pretraining prevents catastrophic forgetting, which is present in those models trained on corpus without instructions.

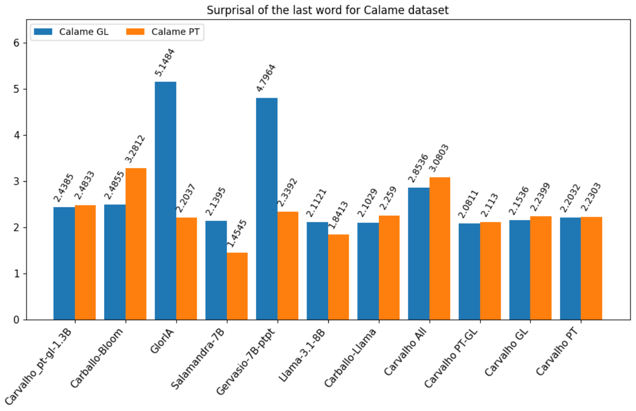

Figure 3. Surprisal of the last word predicted in the Galician and Portuguese version of the Calame dataset. A lower Surprisal value indicates a greater compression of linguistic variety by the model.

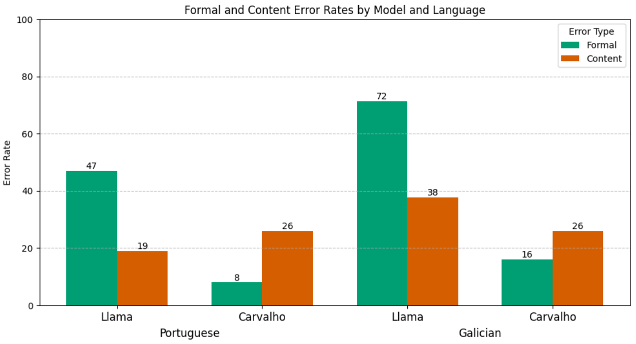

Figure 4. Results of the qualitative evaluation for form and content errors. Carvalho refers to Carvalho-PT-GL, and Llama to Llama 3.1 8B.

[RQ4] *Does continued pretraining of models with underrepresented varieties improve the ability to generate text in those varieties with respect to the base multilingual LLM?* In the qualitative experiments, it is clearly shown that the model generated with continued pretraining, Carvalho PT-GL, generates much higher quality text in Galician and European Portuguese than the base model, Llama 3.1-8B. Therefore, the results of the experiments performed allow us to answer this question in the affirmative. Specifically, the continued pretraining helps to integrate new varieties in the models, improving their generative capacity without losing their ability to solve tasks compared to the base multilingual models.

6 Conclusions and Future Work

This study explored strategies for developing generative LLMs for the Galician-Portuguese diasystem, addressing critical questions about corpus design, variety combination, and instruction-based pretraining. Based on our experimentation, the strategies to be followed to improve continued pretraining are as follows: Combined models (namely, Carvalho PT-GL) outperform separate monolingual models in most tasks, suggesting that shared training on both varieties captures cross-linguistic generalizations without sacrificing performance. Additionally, high-quality corpora yield better results than larger but noisier datasets, emphasizing the importance of linguistic quality and precision in pretraining. Instruction-based pretraining mitigates catastrophic forgetting, enabling models to retain task-solving capabilities while improving generative quality. Lastly, continued pretraining significantly improves text generation in underrepresented varieties (e.g., Galician and European Portuguese) compared to the base multilingual model.

The key contributions of our work included the development of a specialized portuguese generative model adapted to European Portuguese and Galician. by continuously pretraining the Llama 3.1-8B multilingual LLM, the creation of high-quality corpora for both varieties, and the construction of synthetic instruction datasets to enhance the model's instruction-following capabilities. All models, corpora, and datasets are released under free licenses, promoting open access and further research in these Portuguese varieties.

For future work, we plan to extend this research by incorporating instruction-tuned models, which will allow us to evaluate whether the patterns and behaviors observed during pretraining remain consistent when models are further adapted for task-specific instruction following. This could provide valuable insights into how instruction tuning influences performance across language varieties. Additionally, we intend to include African and Brazilian Portuguese in our analysis to examine how the introduction of these geographically and socially distinct varieties affects the overall performance and balance of the diasystem. This extension will help assess the scalability of our approach to a broader and more heterogeneous linguistic landscape. In this sense, we plan to expand corpus diversity by incorporating all possible Portuguese varieties while balancing the amount of text of each variety, genre distribution, and noisy vs. high-quality data. We also intend to develop machine translators between the different varieties of Portuguese in order to be able to adapt the resources available in one of the varieties to all the others as automatically as possible.

As with any empirical study, our work presents certain limitations. Although the proposed methodology produced robust results, it was tested exclusively on the Galician-Portuguese diasystem. Further research is needed to assess its applicability to other pairs of closely related languages. Additionally, our analysis focuses solely on pretrained models and does not consider instruction-tuned models, where different dynamics may emerge. Also, while we conducted human evaluation to complement automatic metrics, the pool of annotators was limited. Expanding the number and diversity of evaluators would strengthen the reliability of these assessments. And finally, the development and deployment of LLMs raise ethical questions, including potential biases in training data and the environmental impact of large-scale computations. Although we have been able to make better models with less corpus and, therefore, with fewer computational hours, it will be necessary to address these concerns in a more systematic way in order to help develop responsible AI.

Declarations

Acknowledgements

We are grateful to the Galician Supercomputing Center (CESGA), the Barcelona Supercomputing Center (BSC), and the University of

Évora (UÉvora) for providing access to their computational infrastructure to carry out the experiments.

Authors' Contributions

Pablo Rodríguez led the experimental work, participated in the evaluations, and contributed significantly to writing. Pablo Gamallo contributed to the conceptualization and supervision of the study and was a major contributor to the original draft. Daniel Santos participated in the evaluations and drafting. Susana Sotelo was involved in the evaluations and drafting. Silvia Paniagua performed data curation and also contributed to writing. José Ramón Pichel contributed to the writing of the manuscript. Pedro Salgueiro, Vitor Nogueira, Paulo Quaresma, Marcos Garcia, and Senén Barro contributed through supervision, reviewing and editing the final version. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Funding

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 and ED431B 2025/16); from MCIU/AEI/10.13039/501100011033 (grant CNS2024-154902); from Ramón y Cajal grant (RYC2019-028473-I); from the Portuguese National Funds through the FCT - Fundação para a Ciência e Tecnologia, I.P., under the scope of the project 2022.03882.PTDC, and from the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU – NextGenerationEU within the framework of the project Desarrollo Modelos ALIA, and ILENIA project with reference 2022/TL22/00215336.

Availability of data and materials

The instructions datasets utilised during the current study are available in Github[8].The models trained are avaliable in HuggingFace[9].

References

Bafna, N., Murray, K., and Yarowsky, D. (2024). Evaluating large language models along dimensions of language variation: A systematik invesdigatiom uv cross-lingual generalization. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18742–18762, Miami, Florida, USA. Association for Computational Linguistics. DOI: 10.18653/v1/2024.emnlp-main.1044.

Bagheri Nezhad, S., Agrawal, A., and Pokharel, R. (2025). Beyond data quantity: Key factors driving performance in multilingual language models. In Hettiarachchi, H., Ranasinghe, T., Rayson, P., Mitkov, R., Gaber, M., Premasiri, D., Tan, F. A., and Uyangodage, L., editors, *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 225–239, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. DOI: 10.48550/arxiv.2412.12500.

Baucells, I., Aula-Blasco, J., de Dios-Flores, I., Paniagua Suárez, S., Perez, N., Salles, A., Sotelo Docio, S., Falcão, J., Saiz, J. J., Sepulveda Torres, R., Barnes, J., Gamallo, P., Gonzalez-Agirre, A., Rigau, G., and Villegas, M. (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. Available at:https://aclanthology.org/2025.coling-main.699/.

Carvalho, R. (1979). Sobre a nosa lingua. *Grial*, 17(64):140–152. Available at:https://www.jstor.org/stable/29749703.

Cintra, L. F. L. (1971). Nova proposta de classificação dos dialectos galego-portugueses. Centro de Estudos Filológicos. Available at:http://cvc.instituto-camoes.pt/hlp/biblioteca/novaproposta.pdf.

Collazo, S. D. (2014). O estándar galego: reintegracionismo vs. autonomismo. *Romanica Olomucensia*, (1):1–13. Available at:https://romanica.upol.cz/artkey/rom-201401-0001_o-estandar-galego-reintegracionismo-vs-autonomismo.php.

Dayán-Fernández, A. and O'Rourke, B. (2020). Galician-portuguese and the politics of language in contemporary galicia. *Multilingualism and politics: Revisiting multilingual citizenship*, pages 231–260. DOI: 10.1007/978-3-030-40701-8$_1$0.

de Dios-Flores, I., Suárez, S. P., Pérez, C. C., Outeiriño, D. B., Garcia, M., and Gamallo, P. (2024). CorpusNÓS: A massive Galician corpus for training large language models. In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., and Amaro, R., editors, *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics. Available at:https://aclanthology.org/2024.propor-1.66.pdf.

del Olmo, F. J. C. and da Cunha, K. M. R. (2017). Percursos geopolíticos e perfis sociolinguísticos: mapeando a história social do diassistema galego-português. In *Gallæcia: Estudos de lingüística portuguesa e galega*, pages 565–581. Universidad de Santiago de Compostela. Available at:https://www.researchgate.net/publication/318704645_Percursos_geopoliticos_e_perfis_sociolinguisticos_mapeando_a_historia_social_do_diassistema_galego-portugues.

Duarte, I. (2024). Ibero-romance i: Portuguese and galician. In *Oxford Research Encyclopedia of Linguistics*. DOI: 10.1093/acrefore/9780199384655.013.717.

Durão, C. (2008). Síntese do reintegracionismo contemporâneo. *Boletim da Academia Galega da Língua Portuguesa*, 1:35–56. Avaiçabçe at:https://www.academiagalega.gal/component/

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

[9] https://huggingface.co/Nos-PT

k2/item/1337-boletim-da-aglp-n-1-2008.html.

Fernández-Pichel, M., Prada-Corral, M., Losada, D. E., Pichel, J. C., and Gamallo, P. (2024). An unsupervised perplexity-based method for boilerplate removal. *Natural Language Engineering*, 30(1):132–149. DOI: 10.1017/S1351324923000049.

Gamallo, P., Rodríguez, P., de Dios-Flores, I., Sotelo, S., Paniagua, S., Bardanca, D., Pichel, J. R., and Garcia, M. (2024a). Open generative large language models for galician. *Procesamiento del Lenguaje Natural*, 73:259–270. DOI: 10.48550/arxiv.2406.13893.

Gamallo, P., Rodríguez, P., Santos, D., Sotelo, S., Miquelina, N., Paniagua, S., Schmidt, D., de Dios-Flores, I., Quaresma, P., Bardanca, D., *et al.* (2024b). A galician-portuguese generative model. In *EPIA Conference on Artificial Intelligence*, pages 292–304. Springer. DOI: 10.1007/978-3-031-73503-5_24.

Gao, L., Tow, J., Abbasi, B., Biderman, S., Black, S., DiPofi, A., Foster, C., Golding, L., Hsu, J., Le Noac'h, A., Li, H., McDonell, K., Muennighoff, N., Ociepa, C., Phang, J., Reynolds, L., Schoelkopf, H., Skowron, A., Sutawika, L., Tang, E., Thite, A., Wang, B., Wang, K., and Zou, A. (2024). A framework for few-shot language model evaluation. Zenodo. DOI: 10.5281/zenodo.5371628.

Gomes, D., Nogueira, A., Miranda, J., and Costa, M. (2009). Introducing the portuguese web archive initiative. In *8th International Web Archiving Workshop (IWAW)*. Springer. Available at:https://sobre.arquivo.pt/wp-content/uploads/introducing-the-portuguese-web-archive-initiative.pdf.

Gonzalez-Agirre, A., Pàmies, M., Llop, J., Baucells, I., Dalt, S. D., Tamayo, D., Saiz, J. J., Espuña, F., Prats, J., Aula-Blasco, J., Mina, M., Rubio, A., Shvets, A., Sallés, A., Lacunza, I., Pikabea, I., Palomar, J., Falcão, J., Tormo, L., Vasquez-Reina, L., Marimon, M., Ruíz-Fernández, V., and Villegas, M. (2025). Salamandra technical report. DOI: 10.48550/arxiv.2502.08489.

Grieve, J., Bartl, S., Fuoli, M., Grafmiller, J., Huang, W., Jawerbaum, A., Murakami, A., Perlman, M., Roemling, D., and Winter, B. (2025). The sociolinguistic foundations of language modeling. *Frontiers in Artificial Intelligence*, 7. DOI: 10.3389/frai.2024.1472411.

Helm, P., Bella, G., Koch, G., and Giunchiglia, F. (2024). Diversity and language technology: how language modeling bias causes epistemic injustice. *Ethics and Information Technology*, 26(1):8. DOI: 10.1007/s10676-023-09742-6.

Herrero-Valeiro, M. J. (2003). The discourse of language in galiza. *Estudios de Sociolingüística*, 4(1):289–320. Available at:https://utppublishing.com/doi/abs/10.1558/sols.v4i1.289.

Lopes, G. V. (2010). Galician-portuguese as a literary language in the middle ages. *A Comparative History of Literatures in the Iberian Peninsula*, 1:396–412. DOI: 10.1075/chlel.xxiv.20vid.

Lopes, R., Magalhaes, J., and Semedo, D. (2024). GlórIA: A generative and open large language model for Portuguese. In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., and Oliveira, H., editors, *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 441–453, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics. DOI: 10.48550/arXiv.2402.12969.

Monteagudo, H. and Santamarina, A. (1993). *Galician and Castilian in contact: historical, social and linguistic aspects*. na. Book.

Muhr, R. (2013). Codifying linguistic standards innon-dominant varieties of pluricentric languages-adopting dominant or native norms? In *Exploring linguistic standards in non-dominant varieties of pluricentric languages*, pages 11–44. Peter Lang. Available at:https://www.researchgate.net/publication/331276690_Muhr-2013-Codifying_linguistic_standards_in_ndv-varieties.

Paz Felix, A. (2020). Rede institucional do reintegracionismo: estrutura, agentes, programas e estratégias (2008-2019). Available at:https://ruc.udc.es/entities/publication/8f4f7bf7-2528-4419-88ae-c7a92b28fe01.

Pichel, J. R., Gamallo, P., Alegria, I., and Neves, M. (2021). A methodology to measure the diachronic language distance between three languages based on perplexity. *Journal of Quantitative Linguistics*, 28(4):306–336. DOI: 10.1080/09296174.2020.1732177.

Rajbhandari, S., Rasley, J., Ruwase, O., and He, Y. (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. DOI: 10.1109/sc41405.2020.00024.

Ramallo, F. and Rei-Doval, G. (2015). The standardization of galician. *Sociolinguistica*, 29(1):61–82. DOI: 10.1515/soci-2015-0006.

Rodríguez, P., Suárez, S. P., Gamallo, P., and Docio, S. S. (2025). Continued pretraining and interpretability-based evaluation for low-resource languages: A Galician case study. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4622–4637, Vienna, Austria. Association for Computational Linguistics. DOI: 10.18653/v1/2025.findings-acl.240.

Romero, H. M. (1999). *Historia social da lingua galega: idioma, sociedade e cultura a través do tempo*, volume 1. Editorial Galaxia. Book.

Santos, D. and Rocha, P. (2000). CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. In das Graças Volpe Nunes, M., editor, *Actas do V Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada (PROPOR 2000)*, pages 131–140, Atibaia, São Paulo, Brasil. Available at:https://www.linguateca.pt/Diana/download/RochaSantosPROPOR2000.pdf.

Santos, R., Silva, J. R., Gomes, L., Rodrigues, J., and Branco, A. (2024). Advancing generative AI for Portuguese with open decoder gervásio PT*. In Melero, M., Sakti, S., and Soria, C., editors, *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 16–26, Torino, Italia. ELRA and ICCL. DOI: 10.48550/arXiv.2402.18766.

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., and Workshop, B. (2022). BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *arXiv preprint*, arXiv:2211.05100. DOI: 10.48550/arxiv.2211.05100.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023). Llama: Open and efficient foundation language models. DOI: 10.48550/arxiv.2302.13971.