A Galician-Portuguese Generative Model

Pablo Gamallo[1], Pablo Rodríguez[1], Daniel Santos[2], Susana Sotelo[1], Nuno Miquelina[2], Silvia Paniagua[1], Daniela Schmidt[2], Iria de-Dios-Flores[3], Paulo Quaresma[2], Daniel Bardanca[1], José Ramom Pichel[1], Vítor Nogueira[2], and Senén Barro[1]

[1] Centro de Investigación en Tecnoloxías Intelixentes (CiTIUS)
Universidade de Santiago de Compostela
{pablo.gamallo,pablorodriguez.fernandez,susana.sotelo.docio,jramon.pichel,
silvia.paniagua.suarez,danielbardanca.outeirino,senen.barro}@usc.gal
[2] VISTA Lab, Centro Algoritmi, Universidade de Évora
{dfsantos,daniela.schmidt,pq,vbn}@uevora.pt, d37384@alunos.uevora.pt
[3] Universitat Pompeu Fabra
iria.dedios@upf.edu

Abstract. Large language models (LLMs) have revolutionized natural language processing, but their predominant focus on English has resulted in biases and performance differences across various languages. This situation is maintained in generative multilingual models, where English continues to be the predominant language. In these models, the presence of European Portuguese is marginal and that of the Galician variety is almost residual. In this work, we describe an open-source Galician-Portuguese generative model, *Carvalho_pt-gl*, focused precisely on these two language variants, which are very close lexically and syntactically. The model was trained using a GPT architecture with 1.3 billion parameters on more than 6B words, balanced between the two varieties. The strategy of continual pertaining was used to adapt a pre-existing LLM that was trained on a trilingual dataset with related languages, thereby overcoming the data limitations that would be faced if the training was started from scratch. Evaluation results involving task-based datasets from standardized benchmarks indicate a promising performance. These findings highlight the critical importance of supporting linguistic diversity in generative models.

Keywords: Large Language Models · Generative Models · Galician · Portuguese · Continual Pretraining.

1 Introduction

The recent emergence of large language models (LLMs) has revolutionized natural language processing (NLP), enabling them to understand and generate human-like text across various languages with remarkable capabilities. However, these models, predominantly trained on vast English corpora, have become biased towards English and exhibit disparities in performance across different languages.

This bias is attributed to the overwhelming majority of training data being in English, with only a small fraction dedicated to other languages. This results in the marginalization and underrepresentation of most languages and their contexts.

As the global landscape evolves towards linguistic diversity, it is essential to address the limitations of current LLMs, particularly regarding their treatment of under-resourced languages like Galician. The lack of adequate representation for minority languages hinders equitable access to NLP technologies and services for diverse linguistic communities, perpetuating the digital language divide and reinforcing linguistic hegemony [10]. The dominance of English in LLM training data has resulted in biases and disparities in performance across different languages. For instance, 92.65% of the training data for GPT-3 was English text.[4]By acknowledging and addressing these biases, we can work towards developing more inclusive and equitable NLP technologies that promote linguistic diversity and bridge the digital language divide.

Within this framework of action, we propose the development of an open-source generative model for the European variants of Portuguese, including the Galician variety, spoken in the northwestern peninsular in the Autonomous Community of Galicia, and in an advanced process of minorization compared to Spanish. Although strongly influenced by Spanish, especially at the lexical level, the syntactic structure of the Galician variety is very similar to that of the standard European Portuguese, which could reinforce mutual learning during the pretraining of the generative model. Crucially, it should be noted that Galician, in spite of belonging to the Galician-Luso-Brazilian diasystem from a typological perspective, is nowadays written using a different spelling system, borrowed from Spanish.

By developing this free LLM specialized in the European varieties of Portuguese, we hope to help companies and third parties interested in technological tools for Galician and Portuguese users to expand their target market by including these two communities of speakers which are so close to each other. Therefore, companies and organizations will be able to integrate this model into their applications, enabling them to adapt their applications to the specific linguistic needs and cultural context of Galician-Portuguese speakers.

To train our model, called *Carvalho_pt-gl*, we explore a strategy based on continual pretraining, an efficient technique to build new LLMs [7], instead of adopting the common approach of pretraining from scratch with randomly initialized weights. The volume of data necessary for pretraining from scratch can be enormous, rendering it unattainable. However, by starting from a base pretrained LLM, we can leverage the existing language knowledge that is already encapsulated within it.

This paper is organized as follows. We begin in Section 2 by introducing the main existing LLMs for Portuguese and Galician, as well as for other Iberian languages. We continue in Section 3 by describing the characteristics of the

[4] `https://github.com/openai/gpt-3/blob/master/dataset_statistics/languages_by_word_count.csv`

two corpora used in the training. Section 4 presents how the experiment for the training of the model was carried out, while Section 5 describes in detail the evaluations performed, using two different types of benchmarks. Finally, we highlight the main conclusions and point to future work in the last section.

2 Related Work: LLMs for Portuguese, Galician and other Iberian languages

In recent years, multilingual LLMs have become the primary approach in NLP and AI system development. However, it is important to note that these models are largely focused on English. For example, in the Llama model [22], a vast majority (89.7%) of the training data is in English, with only a small portion allocated to other languages like German (0.17%), French (0.16%), and Chinese (0.13%). Iberian languages such as Galician-Portuguese have very limited or even residual representation in the training data. Given this situation, in the last years, there has been an effort to develop language models for Portuguese, even for Galician, with the goal of countering the dominance of Anglocentrism in LLMs.

In the case of Portuguese language research, some generative models have been recently developed. Namely, contributions come from [12] and [19], who respectively introduced Glória and Gervásio, two autoregressive generative models for Portuguese language. Glória uses the GPTNeo architecture, with 1.3B and 2.7B parameters [9], trained exclusively on Portuguese texts from Portugal (35B tokens), excluding Brazilian content. On the other hand, Gervásio comprises two 7B parameter models, based on continual pretraining from LLaMA 2 [22], one trained on Brazilian data and the other on corpora from Portugal. In the same line, Albertina PT-* [18] presents a 900M parameter DeBERTa encoder language model for Portuguese, developed in both American and European Portuguese versions. Considering only Brazilian Portuguese, [16] proposes Sabiá models, which were pre-trained using the Portuguese subsets from the ClueWeb 2022 dataset using LLaMA's 7 billion and 65 billion parameter architectures.

As for Galician language modeling, two auto-encoding models based on BERT were trained with relatively small corpora, well under 1B tokens: Bertinho [23] and Bert-Galician [6], each available in small (6 layers) and base (12 layers) transformer versions. Recently, a family of generative models for Galician, called Carballo, with 1.3B parameters and trained with part of the corpus used for Carvalho_pt-gl has been released [5].

For the Catalan language, the AINA project has emerged as a significant contributor, presenting pre-trained models such as FLOR-1.3B[5] and FLOR-6.3B[6]. The training corpus, which is constituted by 26B tokens, comprises Spanish and Catalan data distributed evenly, each accounting for approximately 40% of the total, with a smaller portion of English data, constituting around 17%. This allocation of some English text aims to safeguard against catastrophic forgetting

[5] https://huggingface.co/projecte-aina/FLOR-1.3B

[6] https://huggingface.co/projecte-aina/FLOR-6.3B

of the main language in the original models, Bloom-1.7B[7] and Bloom 7.1B[8], from which both FLOR-1.3B and FLOR-6.3B were respectively developed by continual pretraining.

Furthermore, endeavors to construct LLMs for the Basque language have achieved notable progress with the release of Latxa generative models [4]. Developed through continual pretraining from LLaMA 2, these models use a dataset of merely 288 million words, yet encompass parameter ranges from 7 billion to 70 billion, making them the most extensive and proficient LLMs devised for Basque to date. The Latxa models exhibit good performance across various instructional tasks for Basque, improving existing multilingual LLMs by far.

Concerning Spanish language, attention must be drawn to the MarIA family of both auto-encoding and generative models. This family, described in [8], encompasses various models employing different architectures (RoBERTa and GPT2), exhibiting robust performance across diverse NLP tasks. These models were pretrained using an extensive corpus of 135B words sourced from the Spanish Web Archive, crawled by the National Library of Spain between 2009 and 2019.

Within this spectrum of models for Iberian languages, there is no model focused on the Galician-Portuguese language spoken throughout the western peninsular. Despite the lexical differences between these two varieties, Galician and European Portuguese maintain a very homogeneous syntactic structure that may allow the learning of one variety to help the learning of the other, and vice versa. To make up for the lack of such a language model and to take advantage of the similarity of the two varieties, we opted to train a Galician-Portuguese generative model.

3 The corpus

The corpus used for training the model is constituted by two partitions, one in Galician and the other in European Portuguese. They have been well balanced in terms of size.

3.1 Galician Corpus

The Galician partition was obtained from two corpora:

CorpusNÓS [2]: the largest collection of openly available Galician texts. This corpus is made up of 13.95GB of text (2.1B words) primarily devised for training large language models (LLMs). The corpus sources are varied and represent a relatively wide range of genres such as literary, journalistic, scientific and administrative. Crucially, the corpus is divided into two subcorpora depending on how the texts were obtained: either via transfer agreement from the text owners or from publicly available sources. A cleaning pipeline was used to filter out odd documents. These pipeline contains several processes, such as document

[7] `https://huggingface.co/bigscience/bloom-1b7`

[8] `https://huggingface.co/bigscience/bloom-7b1`

deduplication, language identifier, and a perplexity-based process to remove those documents whose perplexity is higher than a threshold set to $2,000$. CorpusNÓS, as well as the cleaning pipeline, is made available via a GitHub repository: `https://github.com/proxectonos/corpora`.

BNE-gl: Galician documents from the web crawlings performed by the National Library of Spain (Biblioteca Nacional de España - BNE) from 2009 to 2019. The National Library of Spain crawls all .es domains once a year, capturing not only documents in Spanish, but also in the other official languages of Spain. This corpus was cleaned with the same pipeline as that used for CorpusNós. After cleaning, the size of the Galician corpus extracted from BNE is 843M words.

In total, by adding CorpusNós and the Galician text cleaned from the BNE corpus, the Galician partition is constituted by approximately 3B words.

3.2 European Portuguese Corpus

The European Portuguese corpus was entirely produced from the Arquivo.pt repository [9]. This is a specialized web archive dedicated to capturing, indexing, and storing a wide array of web content related to the Portuguese web and cultural sphere. By performing these tasks, Arquivo.pt represents a cornerstone in the preservation of the Portuguese digital heritage.

The corpus was constructed by processing the repository index files that complied with the following rules:

- HTTP 200 OK: at the moment of retrieving the original content from the crawled website, the engine received an HTTP 200 OK (success).
- MIME type: from the available (indexed) metadata, only MIME types (or Content Type) from Table 1 were chosen.

Mime type	**Document count**
application/msword	24,272
application/pdf	2,094,074
application/rtf	913
text/html	76,862,145
text/plain	152,279
text/rtf	452

Table 1: Document count for each mime type of the Portuguese corpus

All documents with perplexity over $2,000$ were excluded. Moreover, there is no specific subject, gender, or restriction applied to the content that composes the corpus. This processing generated 6M unique documents with 3.8B tokens. Finally, this corpus arose within the scope of an ongoing project, AiBERTa: pretrAined BERT language model for European Portuguese based on Arquivo.pt [13].

[9] https://arquivo.pt/

4 Training process

4.1 Continual pretraining

Continual pretraining offers a more efficient approach to adapting language models to new languages, improving their performance and reducing the need for extensive retraining from scratch. The general method to perform continual pretraining is to use a base model that was trained using data from languages that are similar to Galician-Portuguese so that the final model can benefit from a non-random initialization of its weights, hence, requiring fewer new tokens. The initial stage of a successful language adaptation involves substituting the model's tokenizer. This step is pivotal because employing the original model tokenizer would result in a low proportion of token splits. Therefore, a new Byte Pair Encoding (BPE) tokenizer was trained using Galician-Portuguese text. Subsequently, the embedding layer undergoes modification by retaining solely the weights corresponding to shared tokens (those found in both the old and new tokenizer), while replacing the remaining ones with the overall mean value [3]. Once the model is suitably initialized, standard pretraining procedures can start using our Galician-Portuguese corpus.

4.2 The experiment

Carvalho_pt-gl was pre-trained using as base model a Catalan/Spanish/English version of Cerebras-GPT-1.3B,[10] which follows the GPT-3 architecture and was initially pretrained solely in English[11]. This decoder architecture consists of 16 attention heads and 24 layers, where the hidden layers have 2048 dimensions. It should be noted that, although the base model does not contain Galician-Portuguese, the most representative languages are Catalan and Spanish, and not English. Unlike most multilingual models, the base model is thus oriented to Romance languages, with a structure and vocabulary close to the target Galician-Portuguese. As a result, we obtained a Galician-Portuguese LLM with the same architecture as the base model, and freely available[12].

To adapt the tokenizer model to Galician-Portuguese, a new Byte Pair Encoding (BPE) tokenizer was trained on our corpus giving rise to a Galician-Portuguese vocabulary with $50,257$ tokens by making use of the adaptation described in the previous subsection.

To pre-train the model, we set the hyperparameters as follows. For optimization, we employ Adam [11] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$, coupled with a weight decay of 0.1. The learning rate commences at 5×10^{-5} and decays linearly. The sequence length is fixed at 2048 tokens, mirroring the Cerebras base model. Training was performed with BF-16 mixed-precision.

We leverage HuggingFace Transformers library [25] for executing the Causal Language Modeling pre-training objective, while DeepSpeed [17] with ZeRO

[10] This model was built within the Aina project: https://projecteaina.cat/

[11] `https://huggingface.co/cerebras/Cerebras-GPT-1.3B`

[12] `https://huggingface.co/Nos-PT/Carvalho_pt-gl-1.3B`

stage 2 optimizations is used to accelerate training. All the experiments were conducted at the Galician Supercomputing Center (CESGA) utilizing 4 nodes equipped with NVIDIA A100 40GB GPUs.

Figures 1 and 2 illustrate the evolution of the loss function in the training and validation sets, respectively. The dual coloration is a consequence of the limitations of the cluster employed, necessitating the training of the initial epoch and the initiation of a new process for the second epoch. It is important to mention that the appropriate parameters were used to continue the training, not to restart it. On the other hand, the change in the trend of the graph, which drops more steeply in the first run, can be explained by the fact that in the second run, the model is seeing the same texts again, which constrains its learning capacity.

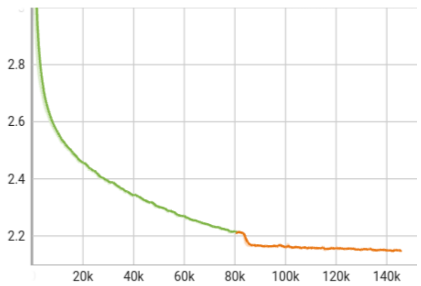

Fig. 1. Training loss graph, constructed from loss values tracked every 500 steps.

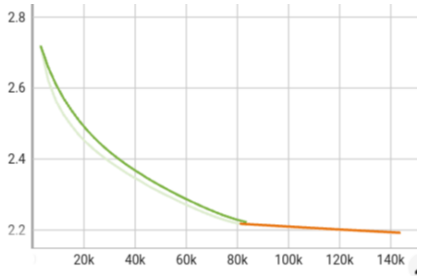

Fig. 2. Validation loss graph, constructed from loss values tracked every 4000 steps.

5 Evaluation

The evaluation of our model and its comparison with other analogous models has been done on two different types of datasets, one focused on generative abilities (CALAME), and the other focused on non-generative and classification tasks (ExtraGLUE). The first one entails zero-shot learning based on prompting an LLM without any examples, while the second one is fine-tuning.

5.1 CALAME-PT-GL

CALAME-PT, or Context-Aware Language Modeling Evaluation for Portuguese, is a benchmark designed by the group that trained GlórIA [12], with the aim of evaluating the generative ability of LLMs with regard to the Portuguese language. It consists of 2076 short texts (contexts) and their corresponding final words, covering a wide range of domains and topics. The goal of the contexts is to provide sufficient information for a model to predict the final word accurately, avoiding excessive specificity or ambiguity. This benchmark is therefore intended to evaluate the quality of the LLM generation in an autocomplete task by using a zero-shot setting.

We have automatically translated the dataset from Portuguese to Galician using a rule-based translator, Apertium [1], and then transliterated all unknown words (i.e., words not found in the Apertium dictionaries) with the software *port2gal*,[13] as was done in [14]. The result of this process is a new dataset adapted for Galician: CALAME-GL. For this purpose, we chose not to use a Portuguese-Galician neural translator, since this task requires preserving the order of the constituents to maintain the word order of the last incomplete sentence. Since neural machine translation can modify in many cases the order of the last sentence by placing the phrase to be completed before the end, we considered that it is not the best strategy to perform the translation. We therefore decided to use, for this benchmark, a quasi-literal translation/transliteration strategy. It should be noted that, being two variants of the same language, literal translation is an optimal choice.

Models	**CALAME-GL**	**CALAME-PT**
Carvalho_pt-gl 1.3B	0.397	0.455
GlorIA 1.3B	0.219	0.488
Gervasio-PTPT 7B	0.265	0.434
mGPT 1.3B	0.264	0.425
Bloom-1b1	0.262	0.456
Cerebras-GPT 1.3B	0.177	0.167

Table 2: Exact match results in CALAME-GL and CALAME-PT

The results obtained with these two benchmarks are shown in Table 2. Our bilingual model, Carvalho_pt-gl, is compared to:

- Generative LLMs of Portuguese: GlórIA 1.3B and Gervásio-PTPT 7B. These models were described in Section 2.
- Multilingual and generative LLMs of analogous size: mGPT [21] and Bloom-1b1 [20]. mGPT is a 1.3B GPT-3 architecture trained on 61 natural languages from 25 language families, while Bloom-1b1 is a 1.1B GPT-3 architecture trained on 46 natural languages and 13 programming languages.
- The trilingual base LLM on which continual pretraining was carried out: Cerebras-GPT 1.3B

All these models were evaluated on both CALAME-PT and CALAMET-GL datasets, in a zero-shot setting. The evaluation is carried out by calculating the percentage of exact matches achieved by each model. In the Galician scenario, Carvalho_pt-gl outperforms the other models by a relevant margin. Carvalho_pt-gl is second, a few points behind GlórIA, in the CALAME-PT dataset. Note

[13] https://github.com/gamallo/port2gal

that GlórIA was trained with more Portuguese text than our model, up to eight times more. The results show a very good performance of Carvalho_pt-gl in both datasets and, consequently, in both language varieties. As this is a typically generative task for decoder-only LLMs, we have not included encoder-only models in the evaluation.

5.2 GLUE Datasets

The GLUE benchmark, introduced by Wang et al. [24], is a collection of nine diverse tasks designed to evaluate and compare the performance of NLP models on tasks such as sentiment analysis, textual entailment, and similarity scoring. In this new evaluation, we assess Carvalho_pt-gl in comparison to both the same generative LLMs evaluated previously and the leading European Portuguese encoder models. The evaluation is performed on discriminative and non-generative tasks by fine-tuning the base models through supervised learning. It is important to point out that encoder-only models generally excel at these tasks due to their bidirectional nature, frequently surpassing decoder-only models, as shown by BERT's leading position on the GLUE leaderboard.[14] This analysis seeks to determine how generative LLMs such as Carvalho_pt-gl measure up against other encoder models, even though encoder-only models naturally have an advantage in such tasks.

GLUE tasks challenge the model's understanding of language, testing both the breadth and depth of linguistic features. The selected tasks for this work are: Microsoft Research Paraphrase Corpus (MRPC), which is aimed at identifying whether two sentences are paraphrases of each other, essentially evaluating a model's ability to detect semantic equivalence; Recognizing Textual Entailment (RTE), which determines whether a given hypothesis can logically be inferred from a premise sentence; Semantic Textual Similarity Benchmark (STS-B), which requires models to assign a similarity score to pairs of sentences on a continuous scale from 1 (not similar at all) to 5 (semantically equivalent); and Winograd Schema Challenge (WNLI), which tests a model's ability to handle coreference resolution within the framework of the Winograd Schema Challenge.

The models were benchmarked using a variant of the GLUE dataset called ExtraGLUE, which is adapted to Portuguese through automatic translation of selected tasks from the GLUE and SuperGLUE benchmarks [15]. This translation ensures the models are evaluated in a language-specific context, thereby providing more accurate and relevant results for Portuguese.

To optimize the fine-tuning of the evaluated models, particularly given the high computational overheads, we utilized a batch size of 16 per device and 2 gradient accumulation steps. For Gervásio, due to its extensive parameter count (7 billion), we adjusted the batch size down to 4 per device to accommodate the model's demands within the available GPU memory. This adjustment required the use of all eight GPUs to handle the increased computational load and to maintain a reasonable training throughput. Training epochs were set to five.

[14] https://gluebenchmark.com/leaderboard

The models underwent fine-tuning over a range of hyperparameters: learning rates of 1e-4 and 1e-5, the use of either a linear or constant scheduler to manage learning rate adjustments, and fixed seeds for training reproducibility (41, 42, 43). These hyperparameters were selected to optimize performance while ensuring that the models could adequately learn from the translated tasks.

The evaluation results are shown in Table 3, where both encoder-only and decoder-only models were evaluated in the four tasks described above. The results show that, as expected, the largest encoder-only transformers (Albertina-PTPT 900m/1.5B) are the best at these tasks, along with Gervásio 7B, the largest decoder-only. Our model is among the best generative LLMs of its size.

Models	**RTE**	**MRPC**	**STS-B**	**WNLI**
	Acc	F1	Pearson	Acc
Encoder-only				
AiBERTa Base	55.3	83.2	80.2	58.9
Albertina-PTPT 100m	55.4	87.6	84.5	65.1
Albertina-PTPT 900m	80.6	89.8	88.7	65.1
Albertina-PTPT 1.5B	82.9	90.3	88.7	59.6
Decoder-only				
Carvalho_pt-gl 1.3B	68.0	86.0	82.6	65.1
Gloria 1.3B	63.8	85.2	82.0	65.1
Gervásio 7B	83.2	90.5	87.9	64.4
Bloom 1.1B	71.5	87.7	85.1	63.7
mGPT 1.3B	58.9	85.5	78.3	65.1

Table 3: Evaluation results on the Portuguese ExtraGLUE tasks.

6 Conclusions

The evaluation involving the generative task on CALAME-PT-GL in Section 5 showed that our Galician-Portuguese model outperforms all other generative LLMs in Galician and is at the highest level of performance in Portuguese. In relation to the evaluation with non-generative datasets in Portuguese (see Table 3 ExtraGLUE), Carvalho_pt-gl is among the best decoder-only models of size 1.1 or 1.3B, but does not reach the results of the best encoder-only models nor those of the decoder-only with the largest architecture, Gervásio. These results confirm that our model, trained with varieties of the same language, has a very acceptable performance in the two varieties separately.

In future work, we will concentrate on three primary aspects: firstly, our goal is to develop foundation models with scaled-up architectures, while avoiding unnecessary complexity, and train them using larger text corpora. To increase the corpus size, we will focus not only on Galician-Portuguese, but also on other neighboring languages, aiming to facilitate multilingualism among closely related languages. By leveraging multilingual modeling predominantly comprising typologically similar languages, we will evaluate whether the text generated in both Galician and Portuguese is improved within this particular multilingual context.

Secondly, we also intend to go deeper into the inclusion of the third variety of Portuguese, which is also the most widespread and widely spoken: Brazilian. We will develop a variety of identifiers between Brazilian and European Portuguese to add a third, uniquely Brazilian, partition to our corpus. We will evaluate the generation ability of models trained with the two varieties perfectly separated (as if they were different languages) and with them mixed in the training corpus. If we take into account the data used in the training of most multilingual models, the Portuguese variants have not been previously separated before training. The results of these tests, differentiating separate and mixed variants, may be useful for many other situations in which there are similar cases of close varieties and similar languages.

Thirdly, we aspire to develop instructed Portuguese models by using instruction datasets prepared in all varieties. We will carry out these experiments by translating (as manually as possible) the existing datasets into European Portuguese and then adapting them to Galician and Brazilian with automatic techniques such as transliteration. Alternatively, we will translate the datasets into Galician and then adapt them to the other two varieties.

We consider that the modeling of the Portuguese language must be carried out by taking into account all its complexity and heterogeneity, both linguistic and cultural, as it is an international language spoken on several continents with substantial differences. In short, the language modeling strategy we propose should, on the one hand, take advantage of the unity of the language and, on the other hand, respect the richness of diversity.

Acknowledgments

This publication was produced within the framework of the 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, and by the Xunta de Galicia through the collaboration agreements signed in with the University of Santiago de Compostela in 2021 and 2022.

Additionally, the authors of this article received funding from MCIN/AEI/10.13039/501100011033 and the European Union Next Generation EU/PRTR (TED2021-130295B-C33), the Galician Government (ERDF 2014-2020: Call ED431G 2019/04, and ED431F 2021/01), by MCIN/AEI/10.13039/501100011033 (grants with references PID2021-128811OA-I00, PLEC2021-007662, and TED2021-130295B-C33, the latter also funded by the European Union Next Generation EU/PRTR), a Ramón y Cajal grant (RYC2019-028473-I), and a Juan de la Cierva Grant (JDC2022-049433-I) funded by MCIN/AEI/10.13039/501100011033 and the European Union Next Generation EU/PRTR. This work also is funded by Portuguese National Funds through the FCT - Fundação para a Ciência e Tecnologia, I.P., under the scope of the project 2022.03882.PTDC.

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

References

1. Corbí-Bellot, A.M., Forcada, M.L., Ortiz-Rojas, S., Pérez-Ortiz, J.A., Ramírez-Sánchez, G., Sánchez-Martínez, F., Alegria, I., Mayor, A., Sarasola, K.: An open-source shallow-transfer machine translation engine for the Romance languages of Spain. In: European Association for Machine Translation, 10th Annual Conference. pp. 79–86. Budapest (2005)
2. de Dios-Flores, I., Suárez, S.P., Pérez, C.C., Outeiriño, D.B., Garcia, M., Gamallo, P.: CorpusNÓS: A massive Galician corpus for training large language models. In: Proceedings of the 16th International Conference on Computational Processing of Portuguese. pp. 593–599 (2024)
3. Downey, C., Blevins, T., Goldfine, N., Steinert-Threlkeld, S.: Embedding structure matters: Comparing methods to adapt multilingual vocabularies to new languages. In: Ataman, D. (ed.) Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL). pp. 268–281. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.mrl-1.20, https://aclanthology.org/2023.mrl-1.20
4. Etxaniz, J., Sainz, O., Perez, N., Aldabe, I., Rigau, G., Agirre, E., Ormazabal, A., Artetxe, M., Soroa, A.: Latxa: An open language model and evaluation suite for basque (2024)
5. Gamallo, P., Rodríguez, P., de Dios-Flores, I., Sotelo, S., Paniagua, S., Bardanca, D., Pichel, J.R., Garcia, M.: Open generative large language models for galician. Procesamiento del Lenguaje Natural **to appear**(arXiv) (2024)
6. Garcia, M.: Exploring the representation of word meanings in context: A case study on homonymy and synonymy. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 3625–3640. ACL (2021). https://doi.org/10.18653/v1/2021.acl-long.281, https://aclanthology.org/2021.acl-long.281
7. Gupta, K., Thérien, B., Ibrahim, A., Richter, M.L., Anthony, Q., Belilovsky, E., Rish, I., Lesort, T.: Continual pre-training of large language models: How to (re)warm your model? In: arXiv (2023)
8. Gutiérrez-Fandiño, Armengol-Estapé, A., Pàmies, J., Llop-Palao, M., Silveira-Ocampo, J., Carrino, J., Armentano-Oller, C., Rodriguez-Penagos, C., Gonzalez-Agirre, A., Villegas, M.: MarIA: Spanish Language Models. Procesamiento del Lenguaje Natural **68**, 39–60 (2022)
9. Hendrycks, D., Basart, S., Kadavath, S., Mazeika, M., Arora, A., Guo, E., Burns, C., Puranik, S., He, H., Song, D., Steinhardt, J.: Measuring coding challenge competence with apps (2021)
10. Khanuja, S., Ruder, S., Talukdar, P.: Evaluating the diversity, equity, and inclusion of NLP technology: A case study for Indian languages. In: Vlachos, A., Augenstein, I. (eds.) Findings of EACL 2023. pp. 1763–1777. ACL, Dubrovnik, Croatia (May 2023). https://doi.org/10.18653/v1/2023.findings-eacl.131, https://aclanthology.org/2023.findings-eacl.131

11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2017)
12. Lopes, R., Magalhaes, J., Semedo, D.: GlórIA: A generative and open large language model for Portuguese. In: Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H.G., Amaro, R. (eds.) Proceedings of PROPOR-2024. pp. 441–453 (Mar 2024), https://aclanthology.org/2024.propor-1.45
13. Miquelina, N., Quaresma, P., Nogueira, V.B.: Generating a european portuguese bert based model using content from arquivo.pt archive. In: Yin, H., Camacho, D., Tino, P. (eds.) Intelligent Data Engineering and Automated Learning – IDEAL 2022. pp. 280–288. Springer International Publishing, Cham (2022)
14. Ortega, J., de Dios-Flores, I., Pichel, J.R., Gamallo, P.: A neural machine translation system for galician from transliterated portuguese text. In: SEPLN-PD 2022. Conference of the Spanish Association for Natural Language Processing 2022: Projects and Demonstrations, . pp. 92–95 (September 2022)
15. Osório, T., Leite, B., Cardoso, H.L., Gomes, L., Rodrigues, J., Santos, R., Branco, A.: Portulan extraglue datasets and models: Kick-starting a benchmark for the neural processing of portuguese (2024)
16. Pires, R., Abonizio, H., Almeida, T.S., Nogueira, R.: Sabiá: Portuguese Large Language Models, p. 226–240. Springer Nature Switzerland (2023). https://doi.org/10.1007/978-3-031-45392-2_15, http://dx.doi.org/10.1007/978-3-031-45392-2_15
17. Rajbhandari, S., Rasley, J., Ruwase, O., He, Y.: Zero: memory optimizations toward training trillion parameter models. In: International Conference for High Performance Computing, Networking, Storage and Analysis. IEEE Press (2020)
18. Rodrigues, J., Gomes, L., Silva, J., Branco, A., Santos, R., Cardoso, H.L., Osório, T.: Advancing neural encoding of portuguese with transformer albertina pt-* (2023)
19. Santos, R., Silva, J., Gomes, L., Rodrigues, J., Branco, A.: Advancing Generative AI for Portuguese with Open Decoder Gervásio PT. In: arXiv (2024)
20. Scao, T.L., et al.: Bloom: A 176b-parameter open-access multilingual language model (2023)
21. Shliazhko, O., Fenogenova, A., Tikhonova, M., Kozlova, A., Mikhailov, V., Shavrina, T.: mGPT: Few-Shot Learners Go Multilingual. Transactions of the Association for Computational Linguistics **12**, 58–79 (01 2024). https://doi.org/10.1162/tacl_a_00633, https://doi.org/10.1162/tacl_a_00633
22. Touvron, H., et al.: Llama 2: Open foundation and fine-tuned chat models. In: arXiv (2023)
23. Vilares, D., Garcia, M., Gómez-Rodríguez, C.: Bertinho: Galician BERT Representations. Procesamiento del Lenguaje Natural **66**, 13–26 (2021)
24. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: Glue: A multi-task benchmark and analysis platform for natural language understanding (2019)
25. Wolf, T., et al.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 38–45 (Oct 2020), https://www.aclweb.org/anthology/2020.emnlp-demos.6