# Adapting Large Language Models for Underrepresented Languages

Eliseo Bao, Anxo Pérez, and Javier Parapar

Information Retrieval Lab (IRLab), Faculty of Computer Science, Universidade da Coruña, 15071 A Coruña, Spain
Centro de Investigación CITIC, Universidade da Coruña, 15071 A Coruña, Spain
Correspondence: eliseo.bao@udc.es

DOI: *https://doi.org/543210/xxxxx1234567890*

*Abstract*: The popularization of Large Language Models (LLMs), especially with the development of conversational systems, makes mandatory to think about facilitating the use of artificial intelligence (AI) to everyone. Most models neglect minority languages, prioritizing widely spoken ones. This exacerbates their underrepresentation in the digital world and negatively affects their speakers. We present two resources aimed at improving natural language processing (NLP) for Galician: (i) a `Llama 3.1` *instruct* model adapted through continuous pre-training on the CorpusNós dataset; and (ii) a Galician version of the Alpaca dataset, used to assess the improvement over the base model. In this evaluation, our model outperformed both the base model and another Galician model in quantitative and qualitative terms.

## Introduction

Large Language Models (LLMs) are artificial intelligence (AI) systems designed to recognize and generate text. Trained on vast amounts of textual data, LLMs learn the patterns, structures, and characteristics of human communication, enabling them to generate coherent and contextually relevant text based on the input they receive. The modern era of language models (LMs) began in 2017 with the introduction of the Transformer architecture (Vaswani et al. 2017), which paved the way for models such as `T5` (Brown et al. 2020) and `GPT-3` (Brown et al. 2020). More recently, LLMs gained attention with the introduction of instruct-oriented variations like `InstructGPT` (Ouyang et al. 2022) in early 2022 and later became ubiquitous thanks to conversational systems like `ChatGPT` (OpenAI 2022). These conversational models are fine-tuned versions of general-purpose LLMs, specifically adapted to manage conversations and answer questions in a more interactive and natural manner.

The quality and diversity of the training data plays a crucial role in defining the abilities of the model. This data typically encompasses a wide range of sources, including books, articles, websites, and various other forms of written content, which can be collectively described as an Internet dump. However, since the majority of web content is in English, LLMs tend to excel in English but struggle with minority languages like Galician, which have a more limited digital presence. This further exacerbates the underrepresentation of these languages and negatively affects their speakers, transferring real world problems to the digital. To address this shortcoming, we propose adapting `Llama 3.1-8B-Instruct` to Galician by continuing the pre-training of the

model with the CorpusNós (de Dios-Flores et al. 2024), a massive Galician corpus made up of 2.1B words encompassing a wide range of genres and sources.

We make both the resulting model[1] and our new evaluation dataset[2] publicly available on the Hugging Face Hub under open licenses. Additionally, we release all source code for replication and further development in other similarly underserved languages[3].

## Related Work

In the context of LLMs, attention to languages other than English has been addressed primarily through multilingual models (Nicholas and Bhatia 2023). These models are pre-trained on datasets containing various languages, enabling cross-language tasks, such as answering in a different language from the one in which a question was asked. Multilingual models also benefit from transfer learning (Bornea et al. 2021), where knowledge of one language enhances understanding and performance in others. Following this approach, early models like `M-BERT` (Devlin et al. 2019), `XLM-R` (Conneau et al. 2020), or `mT5` (Xue et al. 2021) had a great success in resource-rich languages, but they struggled with low-resource languages such as *Galician*, which are poorly represented in standard corpora. More recent models such as `Llama 3.1` explicitly state that they were trained on a broader collection of languages than the officially supported, but also state that usage in languages beyond those explicitly referenced as supported is out-of-scope. Therefore, the performance gap between high-resource languages (such as English and Spanish) and low-resource languages remains substantial.

One approach to addressing this issue is *continued pre-training*, which involves improving a model's linguistic knowledge in a specific domain by further training it on language-specific corpora (Gururangan et al. 2020). This technique has proven effective in improving model performance for languages with limited digital resources. An example of this is `DiscoResearch/Llama3-German-8B`, a model based on `Llama3-8B` that demonstrated significant improvements in German while maintaining strong performance in English (Occiglot 2024). This suggests that this method could also benefit the development of a conversational Galician LLM. Previous efforts in this area, such as the creation of the `Carballo` models by Gamallo et al. (2024), leveraged continued pre-training to adapt two different 1.3B parameter models to Galician. However, these Galician models are only text-generative and were not fine-tuned to follow instructions. Therefore, developing an instruct based and conversational Galician LLM remains a challenge.

## Method

Continued pre-training is the process of continuing to train a pre-trained model using new raw, unlabelled, data. This approach allows to adapt a LLM to enhance its performance for a particular field of knowledge, without discarding the previously acquired knowledge. Therefore, what we propose is to perform continued pre-training

1 *https://huggingface.co/irlab-udc/Llama-3.1-8B-Instruct-Galician*
2 *https://huggingface.co/datasets/irlab-udc/alpaca_data_galician*
3 *https://gitlab.irlab.org/eliseo.bao/xovetic-llms-underrepresented-languages*

(see Training) over `Llama-3.1-8B-Instruct` (Dubey et al. 2024) with a Galician corpus (see CorpusNÓS), in order to adapt this model for the Galician language. We select `Llama 3.1-8B` as it is one of the best performers at the time of this research[4] while keeping a reduced size. We choose the *instruct* version of the model because we are interested in a conversational model oriented to natural language question-answer interaction. A more rigorous approach would imply to start with a non-instruct base model, continue pre-training it, and then fine-tune with instructions. However, performing this whole process, especially the final part of getting the model to follow instruction-response turns, would be very costly and data intensive. Therefore, we decided to continue the pre-training by starting directly from an *instruct* version.

### Data

We used two different data sources for this work: CorpusNÓS, a large collection of Galician texts used for continued pre-training, and Galician Alpaca, a set of Galician instruction-answer pairs, used for model evaluation.

**CorpusNÓS** A comprehensive Galician corpus containing 1.8 billion tokens across 7.9 million documents (de Dios-Flores et al. 2024). We processed the full dataset to prepare it for continued pre-training. Each data instance was split into smaller chunks, with a maximum of 512 tokens per chunk. After this processing step, we obtained approximately 60 million distinct texts, each conforming to the 512-token limit. As previously mentioned, the more training data we use, the better the model's performance could be. However, utilizing larger datasets significantly increases the required training time. Given that our primary objective is to demonstrate the effectiveness of our proposed methodology for adapting LLMs for underrepresented languages rather than to maximize model performance, we opted to reduce the dataset to a random subsample of 2 million instances from the processed dataset.

**Galician Alpaca** The Alpaca dataset consists of 52000 instruction-response pairs generated using OpenAI's `text-davinci-003 model` (Taori et al. 2023). Originally, the dataset is entirely in English, and each data instance contains a task instruction, optional input (included in about 40% of the examples), and a response generated by the model. We generated a new version of this dataset for Galician by automatically translating the original dataset using the Python package `googletranslatepy`[5]. From this newly translated dataset, we created an evaluation set by randomly selecting 500 instances. We acknowledge that automatic translation may introduce errors or biases. Therefore, we advise users to approach the data with caution and to consider developing methods for filtering or improving its quality.

### Training

Pre-training a state-of-the-art LLM requires a lot of computational resources, as even the smallest versions of these models contain billions of parameters. To optimize the pre-training process, we utilize a technique called Low-Rank Adaptation (LoRA) (Hu et al. 2022), which is a type of Parameter-efficient fine-tuning (PEFT) (Ding et al. 2023).

---

[4] *https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard*

[5] *https://suqingdong.github.io/googletranslatepy/*

This lightweight adaptation method modifies pre-trained models, allowing them to be tuned efficiently with minimal computational resources. Instead of updating all the model parameters during training, LoRA introduces low-rank matrices that capture domain-specific adaptations while keeping the majority of the model unchanged.

The training configuration was as follows: Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1 \times 10^{-8}$, with the initial learning rate set to $1 \times 10^{-4}$. With the LoRA technique, we reduced the trainable parameters to 20M out of a total of 8B. Training was performed over 1 epoch with a total training batch size of 256. The model was trained for 5352 steps, starting with an initial loss of 2.6347, which steadily decreased to 1.9329 by the final step. The entire training process spanned 16 hours.

## Experiments

We designed an evaluation experiment to assess the performance of `Llama-3.1-8B-Instruct-Galician` in comparison to the base model, `Llama-3.1-8B-Instruct`, and also `Carballo-cerebras-1.3B`, the best-performing model of those proposed by Gamallo et al.. We randomly selected a sample of 500 instances from Galician Alpaca, obtained the responses from all models to these instructions, and evaluated their performance according to our Evaluation framework. This allowed us to obtain the Results of the experiment.

### Evaluation

**Quantitative** We used two quantitative metrics: BLEU (Papineni et al. 2002) and ROUGE (Lin 2004). Both metrics compare the generated responses to a given set of instructions with the expected answers, though they emphasize different aspects: BLEU assesses fluency and adequacy, focusing on alignment with reference texts, while ROUGE measures content overlap, ensuring the generated output captures key information from the references. Notably, these metrics allow for some degree of paraphrasing, making them useful for evaluating the quality of responses that may differ slightly from the ground truth but still convey the same intended meaning.

**Qualitative** Evaluating LLM-based chat assistants presents significant challenges due to their wide range of capabilities and the limitations of existing benchmarks. To address these issues, Zheng et al. investigated the use of strong LLMs as evaluators for LLM-based chat assistants. Building on this research, we developed an LLM-based judge using the `Llama-3.1-70B-Instruct` model, prompted according to the template shown in Appendix A. The prompt is designed to guide the LLM in evaluating and comparing the responses of two AI assistants to a specific user query, with the goal of determining which assistant provides a better answer in terms of Galician usage. The model is also instructed to remain objective, impartial, and methodical throughout the evaluation.

### Results

For space constraints, in Tables 1 and 2 we refer to `Llama-3.1-8B-Instruct-Galician`, `Llama-3.1-8B-Instruct`, and `Carballo-cerebras-1.3B` as *Llama GL*, *Llama*, and *Carballo*, respectively. In Table 1 we report the results for the quantitative evaluation.

Across all considered metrics, *Llama GL* outperformed *Llama* and *Carballo*. Our model achieved higher BLEU-4 and ROUGE scores, indicating a greater n-gram overlap and content similarity to reference responses compared to the other models.

Table 1: Quantitative evaluation of responses to 500 Galician Alpaca instructions. Highest scores for each metric are marked in bold.

| | BLEU-4 | ROUGE-1 | ROUGE-2 | ROUGE-L |
|---|---|---|---|---|
| *Llama* | 20.97 | 28.77 | 11.37 | 20.43 |
| *Llama GL* | **23.13** | **30.42** | **11.95** | **21.84** |
| *Carballo* | 2.60 | 19.41 | 4.93 | 2.94 |

Table 2: Qualitative evaluation of responses to 500 Galician Alpaca instructions. The table presents the percentage of instances where the judge preferred each model, including ties and malformed outputs. Best value for each pair is marked in bold.

| | *Llama GL* | *Llama* | *Carballo* | Tie | Malform. |
|---|---|---|---|---|---|
| *Llama GL* Vs. *Llama* | **41.6**% | 37.6% | – | 17.0% | 3.8% |
| *Llama GL* Vs. *Carballo* | **72.4**% | – | 14.0% | 11.6% | 2.0% |
| *Carballo* Vs. *Llama* | – | **43.8**% | 33.6% | 20.0% | 2.6% |

Table 2 shows the results of the qualitative. The judge LLM preferred responses from the *Llama GL* model in 41.6% of the 500 Galician Alpaca instructions evaluated, outperforming the *Llama* model, which was preferred in 37.6% of cases. In 17% of instances, the judge found the responses to be tied, while 3.8% of cases resulted in malformed outputs. In the second comparison between *Llama GL* and *Carballo*, the judge significantly favored the *Llama GL* model in 72.4% of cases, while the *Carballo* model was preferred in only 14% of the cases. In the final comparison between *Carballo* and *Llama*, the judge preferred the *Llama* model in 43.8% of the cases, while the *Carballo* model was preferred in 33.6% of cases.

These results highlight the effectiveness of a continued pre-training approach for adapting a LLM to a new language, in this case Galician. Even with a relatively small training corpus and limited training time due to computational constraints, the resulting model demonstrates improvements over the base model across all evaluated metrics and scenarios.

## Conclusions

In this work, we introduced `Llama-3.1-8B-Instruct-Galician`, a Large Language Model (LLM) adapted for an underrepresented language, Galician, through continued pre-training on a specialized textual corpus. Additionally, we created a new instruction dataset, *Galician Alpaca*, which we used to assess the model's performance. `Llama-3.1-8B-Instruct-Galician` demonstrated greater content alignment with reference responses and showed improved fluency during evaluation, outperforming both its base model, `Llama-3.1-8B-Instruct`, and another Galician model, `Carballo-cerebras-1.3B`, in both quantitative and qualitative evaluations. As we move forward, we encourage further research and collaboration in this field to continue ad-

vancing the capabilities of LLMs for diverse linguistic communities. It is mandatory to ensure that technology serves as a bridge rather than a barrier, empowering speakers of all languages in the digital age.

## $CO_2$ Emission

Training and experiments were conducted using a private infrastructure with an estimated carbon efficiency of 0.432 $kgCO_2eq/kWh$. A total of 60 hours of computation was performed on 4x NVIDIA A100 SXM4 80GB hardware. The estimated total emissions were 10.37 $kgCO_2eq$. The estimations were made using the MachineLearning Impact calculator[6] presented by Lacoste et al. (2019).

## Acknowledgements

This work has received support from projects: PLEC2021-007662 (MCIN/AEI/10.1303 9/501100011033 Ministerio de Ciencia e Innovación, European Union NextGe neration EU/PRTR) and PID2022-137061OB-C21 (MCIN/AEI/10.13039/501100011033, Ministerio de Ciencia e Innovación); Consellería de Educación, Universidade e Formación Profesional, Spain (grant number ED481A-2024-079 and accreditations 2019–2022 ED4 31G/01 and GPC ED431B 2022/33) and the European Regional Development Fund, which acknowledges the CITIC Research Center.

## Bibliography

BORNEA, MIHAELA, LIN PAN, SARA ROSENTHAL, RADU FLORIAN, and AVIRUP SIL, May 2021. "Multilingual transfer learning for qa using translation as data augmentation." *Proceedings of the AAAI Conference on Artificial Intelligence* 35(14), pages 12,583–12,591.

BROWN, TOM, BENJAMIN MANN, NICK RYDER, MELANIE SUBBIAH, and JARED D KAPLAN, 2020. "Language models are few-shot learners." In *Advances in Neural Information Processing Systems*, volume 33. Curran Associates, Inc., volume 33, pages 1877–1901.

CONNEAU, ALEXIS, KARTIKAY KHANDELWAL, NAMAN GOYAL, VISHRAV CHAUDHARY, and GUILLAUME WENZEK, July 2020. "Unsupervised cross-lingual representation learning at scale." In *Proceedings of the 58th Annual Meeting of the ACL*. ACL, pages 8440–8451.

DEVLIN, JACOB, MING-WEI CHANG, KENTON LEE, and KRISTINA TOUTANOVA, June 2019. "BERT: Pre-training of deep bidirectional transformers for language understanding." In *Proceedings of the 2019 Conference of the North American Chapter of the ACL: Human Language Technologies, Volume 1 (Long and Short Papers)*. ACL, pages 4171–4186.

DING, NING, YUJIA QIN, GUANG YANG, FUCHAO WEI, and ZONGHAN YANG, Mar 2023. "Parameter-efficient fine-tuning of large-scale pre-trained language models." *Nature Machine Intelligence* 5(3), pages 220–235. ISSN 2522-5839.

DE DIOS-FLORES, IRIA, SILVIA PANIAGUA SUÁREZ, CRISTINA CARBAJAL PÉREZ, DANIEL BARDANCA OUTEIRIÑO, MARCOS GARCIA, and PABLO GAMALLO, March 2024. "CorpusNÓS:

[6] *https://mlco2.github.io/impact*

A massive Galician corpus for training large language models." In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*. ACL, pages 593–599.

Dubey, Abhimanyu, Abhinav Jauhri, and Abhinav Pandey, 2024. "The llama 3 herd of models."

Gamallo, Pablo, Pablo Rodríguez, Iria de Dios-Flores, Susana Sotelo, and Silvia Paniagua, 2024. "Open generative large language models for galician."

Gururangan, Suchin, Ana Marasović, Swabha Swayamdipta, Kyle Lo, and Iz Beltagy, July 2020. "Don't stop pretraining: Adapt language models to domains and tasks." In *Proceedings of the 58th Annual Meeting of the ACL*. ACL, pages 8342–8360.

Hu, Edward J, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, and Yuanzhi Li, 2022. "LoRA: Low-rank adaptation of large language models." In *International Conference on Learning Representations*.

Lacoste, Alexandre, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres, 2019. "Quantifying the carbon emissions of machine learning."

Lin, Chin-Yew, July 2004. "ROUGE: A package for automatic evaluation of summaries." In *Text Summarization Branches Out*. ACL, pages 74–81.

Nicholas, Gabriel and Aliya Bhatia, 2023. "Lost in translation: Large language models in non-english content analysis."

Occiglot, 2024. "New set of german language models."

OpenAI, 2022. "Introducing chatgpt."

Ouyang, Long, Jeffrey Wu, Xu Jiang, Diogo Almeida, and Carroll Wainwright, 2022. "Training language models to follow instructions with human feedback." In *Advances in Neural Information Processing Systems*, volume 35. Curran Associates, Inc., volume 35, pages 27,730–27,744.

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

Taori, Rohan, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, and Xuechen Li, 2023. "Stanford alpaca: An instruction-following llama model."

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, and Llion Jones, 2017. "Attention is all you need." In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., volume 30.

Xue, Linting, Noah Constant, Adam Roberts, Mihir Kale, and Rami Al-Rfou, June 2021. "mT5: A massively multilingual pre-trained text-to-text transformer." In *Proceedings of the 2021 Conference of the North American Chapter of the ACL: Human Language Technologies*. ACL, pages 483–498.

ZHENG, LIANMIN, WEI-LIN CHIANG, YING SHENG, SIYUAN ZHUANG, and ZHANGHAO WU, 2024. "Judging llm-as-a-judge with mt-bench and chatbot arena." In *Proceedings of the 37th International Conference on Neural Information Processing Systems*. Curran Associates Inc., NIPS '23.

## Appendix A

Below is the prompt template we use for the *LLM-as-a-judge* evaluation:

```
[Sistema]
Debes actuar como un xuíz imparcial e avaliar a calidade
lingüística das respostas proporcionadas por dous asistentes de
IA en galego á pregunta do usuario.  A túa tarefa é determinar cal
dos asistentes fai un mellor uso da lingua galega en termos de cor-
rección gramatical, precisión léxica e riqueza de vocabulario.  Un
aspecto moi importante é que o asistente faga uso do galego e que
non responda en portugués nin en castelán.  Comeza comparando am-
bas respostas de forma detallada e imparcial, evitando prexuízos
derivados da orde en que se presenten.  Lembra que a lonxitude non
debe influír no teu xuízo, nin tampouco debes favorecer un asis-
tente por cuestións de estilo ou nome.  Limítate a avaliar a cali-
dade do uso do galego.  Despois de proporcionar a túa explicación,
mostra o teu veredicto final seguindo estrictamente este formato:
"[[A]]" se o asistente A é mellor, "[[B]]" se o asistente B é mel-
lor, e "[[C]]" se hai un empate.

[Pregunta do Usuario]
{pregunta}

[Comezo da Resposta do Asistente A]
{resposta_a}
[Fin da Resposta do Asistente A]

[Comezo da Resposta do Asistente B]
{resposta_b}
[Fin da Resposta do Asistente B]
```