Comparing LLM prompting with Cross-lingual transfer performance on Indigenous and Low-resource Brazilian Languages

David Ifeoluwa Adelani
University College London, UK
d.adelani@ucl.ac.uk

A. Seza Doğruöz
Universiteit Gent, LT3, IDLab, Gent, Belgium
as.dogruoz@ugent.be

André Coneglian
Federal University of Minas Gerais, Brazil
coneglia@ufmg.br

Atul Kr. Ojha
Insight SFI Research Centre for Data Analytics, University of Galway, Ireland
atulkumar.ojha@insight-centre.org

Abstract

Large Language Models are transforming NLP for a variety of tasks. However, how LLMs perform NLP tasks for low-resource languages (LRLs) is less explored. In line with the goals of the AmericasNLP workshop, we focus on 12 LRLs from Brazil, 2 LRLs from Africa and 2 high-resource languages (HRLs) (e.g., English and Brazilian Portuguese). Our results indicate that the LLMs perform worse for the part of speech (POS) labeling of LRLs in comparison to HRLs. We explain the reasons behind this failure and provide an error analysis through examples observed in our data set.

1 Introduction

Despite numerous advancements in the NLP research due to Large Language Models (LLMs), available resources mainly cover 20 out of the estimated 7,000 languages (Magueresse et al., 2020). As a result, majority of world languages could still be considered as ”low-resource”.

Being a low-resource language (LRL) encompasses different types of inadequacies with respect to the availability of data for creating language technologies (Gupta, 2022). Focusing on multilingual linguistic scene in South America, we test the performance of LLMs for annotating part-of-speech (POS) tagging for 12 LRLs from Brazil, make a comparison with 2 LRLs from Africa and 2 high resource languages (HRLs) (e.g., English and Brazilian Portuguese) through human evaluation.

The evaluation is challenging for two reasons. First, there is a lack of benchmark datasets for the LRLs in Brazil in general. The ones we were able to find in universal dependencies (UD) data base, [1] do not have the training data to fine-tune multilingual language models. Hence, we can only leverage prompting LLMs or cross-lingual transfer through multilingual language models. Secondly, there is a lack of large monolingual data to benefit from effective multilingual and cross-lingual transfer techniques (Pfeiffer et al., 2020; Ansell et al., 2022; Alabi et al., 2022). We could only find the Bible corpora with less than 35K sentences for 7 out of the 12 languages.

We perform the evaluation on 12 Brazilian LRLs by prompting GPT-4 LLM and cross-lingual transfer individually from English and Brazilian Portuguese leveraging XLM-R. We preferred GPT-4 because the other open multilingual models (e.g., mT0 (Muennighoff et al., 2022), AYA (Ustun et al., 2024)) do not support the LRLs in this study. The results of both methods indicate low performance (less than 34.0% while high-resource languages achieved over 90.0%). However, GPT-4 leads to better results and Brazilian Portuguese performs better than English in zero-shot evaluation. Furthermore, to boost the performance of cross-lingual transfer, we perform language adaptation using XLM-R on each language, before fine-tuning Brazilian Portuguese, and evaluating on that language. This boosts the performance by +3 to +12.0 points on six out of seven languages. Our findings suggest that cross-lingual transfer to these languages is very challenging and having few training examples may further boost the performance. Therefore, there is a need for building NLP resources across different tasks for these LRLs.

1.1 Multilingualism in Brazil

Brazil is the 5th largest country of the world (qua land area) with a population of 203 million[2] and

[1] https://universaldependencies.org/

[2] Instituto Brasileiro de Geografia e Estatística. 2023. https://www.ibge.gov.br/en/cities-and-states.html. Accessed: 2023-12-15

Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 34–41
June 21, 2024 ©2024 Association for Computational Linguistics

it is highly multilingual. Although (Brazilian) Portuguese is the official language, there are approx. 160 native/indigenous as well as sign and immigrant languages.[3]

Following Rodrigues (1986), the two macro-language families among Brazilian native languages are Tupi (8 language families, 52 languages), and Macro-Jê (7 language families, 39 languages). There are also several large language families (e.g., Karib (21 languages), Arawak (20 languages), Arawá (7 languages), Tukano, Maku, and Yanomami), six smaller language families to the south of the Amazon river (e.g., Guaikurú (1 language), Nambikwára (3 languages), Txapakura (3 languages), Pano (13 languages), Múra (2 languages), and Matukína (4 languages)) and approx. 10 languages which are not part of any these families.

These languages share grammatical properties due to family inheritance or areal contact (Aikhenvald, 2002). In terms of morphology, most of these languages are polysynthetic, head-marking, and agglutinating with little fusion (Dixon and Aikhenvald, 1999; Hengeveld et al., 2007). In term of syntax, there is quite some variation in terms of word order among these languages(Campbell, 2012).

2 Literature Overview

In terms of labelled datasets for Brazilian LRLs, we only found datasets from the UD tasks: Gerardi et al. (2022) developed for TUDET UD treebanks covering 8 Tupian languages, other languages covered in UD are Apurina (Hämäläinen et al., 2021), Bororo, Madi-Jarawara, and Xavante (contributed by the TUDET team). For the monolingual data, we found seven Bible corpora on the eBible corpus (Akerman et al., 2023) that are freely available. All languages lack a large monolingual corpus which makes it very challenging for cross-lingual transfer and multilingual pre-training of LLMs.

In terms of evaluation, some studies have already shown the potential of prompting LLMs in multilingual settings (Ahuja et al., 2023a; Lai et al., 2023), including some LRLs (Ojo et al., 2023; Ahuja et al., 2023b). However, evaluations covering Brazilian LRLs are lacking. To the best of our knowledge, our study is the first to fill this gap.

[3]PIB. 2023. https://pib.socioambiental.org/pt/Linguas. Accessed: 2023-12-15

3 Experimental setup

We focus our evaluation of POS tagging (a subtask of universal dependencies (UD)) on Brazilian LRLs due to the simplicity of the task, its popularity, and the availability of the test evaluation datasets in UD [4].

3.1 Evaluation Datasets

We evaluated 12 Brazilian LRLs and 2 African languages for a comparison across other regions with low-resource languages. Finally, we added 2 HRLs (i.e., English and Brazilian Portuguese). Our definition of HRL is based on the size of unlabelled data on the web. The larger their size are, the more likely they are included in pre-training of the LLMs [5] and multilingual pre-trained LMs (Conneau et al., 2020). While UD (Zeman et al., 2023) covers many languages, most LRLs only have a test set because of their limited sizes (less than 10k tokens). The Brazilian LRLs we evaluated on have also less than 13k tokens (except Nheengatu with 12,621 tokens).

Table 1 shows the languages in our evaluation, their language family, availability of monolingual corpus or Bible corpus in that language, UD dataset, and sizes. We collected the Bible corpus from the eBible website and used it for language adaptation. We have two test sets in our evaluation: (1) **Test set A**: the original test set in the UD benchmark (2) **Test set B** the subsample of Test set A where we removed sentences that GPT-4 fails to provide predictions for (mostly due to not properly identifying the language). We added this information for a fair comparison of the methods (i.e. using the same number of sentences in evaluation).

3.2 Models

For the experiments, we consider three approaches that are popular in the zero-shot setting since we lack the training data for the Brazilian languages (see Appendix A for details).

Prompting GPT-4 We prompt GPT-4 using a similar prompt provided by Lai et al. (2023) where the model is provided a task description before the input (seeAppendix B for details).

Cross-lingual transfer We trained a POS tagger individually for English and Portuguese, and per-

[4]https://universaldependencies.org/
[5]https://help.openai.com/en/articles/8357869-chatgpt-language-support-alpha-web

Language	Language family	Monolingual data size	UD dataset name	Train	Dev	Test set A	Test set B
high-resource languages							
English (en)	Indo-European/West Germanic	not collected	en_ewt	12,544	2,001	2,007	-
Portuguese (pt)	Indo-European/Romance	not collected	pt_gsd	9,616	1,204	1,200	-
Brazilian languages							
Apurina (apu)	Arawakan	Bible (8,729)	apu_ufpa	-	-	152	134
Akuntsu (aqz)	Tupian	N/A	aqz_tudet	-	-	343	267
Karo (arr)	Tupian	N/A	arr_tudet	-	-	674	172
Bororo (bor)	Macro-Jê	Bible (8,254)	bor_bdt	-	-	371	161
Guajajara (gub)	Tupian	Bible (33,757)	gub_tudet	-	-	1,182	914
Madi-Jarawara (jaa)	Arawan	Bible (8,606)	jaa_jarawara	-	-	20	18
Makurap (mpu)	Tupian	N/A	mpu_tudet	-	-	37	8
Munduruku(myu)	Tupian	Bible (8,430)	myu_tudet	-	-	158	82
Tupinamba (tpn)	Tupian	N/A	tpn_tudet	-	-	581	458
Kaapor (urb)	Tupian	Bible (8,535)	urb_tudet	-	-	83	20
Xavante (xav)	Macro-Jê	Bible (8,213)	xav_xdt	-	-	148	128
Nheengatu (yrl)	Tupian	N/A	yrl_complin	-	-	1239	-
African languages							
Wolof (wo)	Niger-Congo/Senegambian	not collected	wo_wtb	1188	449	470	470
Yoruba (yo)	Niger-Congo/Volta-Niger	not collected	yo_ytb	-	-	318	318

Table 1: **UD-POS datasets in our evaluation:** We provide the training, validation and test splits we used for experiments. Test set A are the original test set in UD, the Test set B is a subset of A where we removed sentences that GPT-4 is not able to run inference for due to non-identification of the language.

form the zero-shot transfer on other languages. We used the XLM-R-large (or simply, XLM-R) (Conneau et al., 2020) for training the models.

Language Adaptive Fine-tuning (LAFT) We leverage LAFT for an effective cross-lingual transfer by first adapting XLM-R-large model to a new language with limited amount of monolingual data (Alabi et al., 2020; Pfeiffer et al., 2020; Chau and Smith, 2021; Alabi et al., 2022). We make use of the Bible data as the fine-tuning corpus since it is the largest one for these languages and we only found 7 (out of 12 Brazilian languages) languages which have a Bible corpus. Similar to Ebrahimi and Kann (2021), we examine the effectiveness of this small pre-training corpus with 8K-34K sentences. According to Pfeiffer et al. (2020), this approach can significantly boost cross-lingual transfer. However, it is not parameter-efficient like the MAD-X they proposed. On the other hand, Ebrahimi and Kann (2021) argued that simple adaptation to a new language is more effective than MAD-X especially when using the Bible corpus for adaptation and we follow this recommendation in our evaluation.

4 Results

Table 2 shows the result of our evaluation on POS tagging with the following key findings:

Zero-shot evaluation results While POS tagging has a performance of 98% (e.g. for English and Portuguese) when training data are available (especially for HRLs), the performance decreases while performing zero-shot transfer to other languages because POS tagging is language-specific. The transfer performance is low for both Brazilian and African languages (probably) because they are not typologically related whereas English and Portuguese are slightly related (i.e., being in the same Indo-European family) and covered by XLM-R, thus achieving an impressive transfer performance ($> +83\%$).

GPT-4 vs. basic cross-lingual transfer GPT-4 performed slightly better than the zero-shot transfer from other languages in our experiments indicating better abilities of LLMs for this task. For English and Portuguese, the performance reaches to 90% (although it is not on par with fully-supervised setting). For African languages, the performance was lower than the HRLs, but it was still decent (64.8-75.4) probably because the LLMs were exposed to some African languages during pre-training. The struggle of GPT-4 for Brazilian LRLs can be explained with the fact that these languages were probably not included during the pre-training. The generation is often not useful for some examples, where GPT-4 declines to give answers like "*As an AI, I'm unable to provide the POS tags for words in languages I'm not programmed to understand.*". Thus, we had to remove such examples from our evaluation. However, this was not the case for African LRLs and the HRLs.

Language adaptation for cross-lingual transfer performance We performed LAFT training on the Bible corpus individually for the *apu*, *bor*, *gub*, *jaa*, *myu*, *urb*, and *xav*. Our results indicate an im-

Language	XLM-R Test set A Full-sup.	XLM-R (zero-shot cross-lingual transfer) Test set A en→ xx	 pt→ xx	 LAFT + pt→ xx	 Test set B en→ xx	 pt→ xx	 LAFT + pt→ xx	GPT-4 Test set B 0-shot
high-resource languages								
en_ewt	**98.0**	**98.0**	83.6			91.9		
pt_gsd	**97.8**	90.0	**97.8**			92.4		
Brazilian languages								
apu_ufpa	-	37.5	40.6	**44.9**	36.8	40.2	**44.7**	42.6
aqz_tudet	-	31.9	37.8		31.3	36.8		**49.5**
arr_tudet	-	3.9	14.9		6.3	19.8		**27.7**
bor_bdt	-	19.0	23.5	27.3	18.4	23.0	26.4	**41.3**
gub_tudet	-	26.5	30.2	36.0	27.8	32.1	**37.1**	36.2
jaa_jarawara	-	28.2	28.4	**34.5**	27.2	27.9	33.6	**33.0**
mpu_tudet	-	4.9	**9.0**		0.0	0.8		0.0
myu_tudet	-	21.2	27.1	**30.3**	10.8	14.8	16.5	18.2
tpn_tudet	-	39.1	41.9		38.9	41.8		**47.2**
urb_tudet	-	7.8	11.8	21.2	9.2	9.5	21.6	**32.3**
xav_xdt	-	26.5	29.0	28.2	27.3	29.9	29.3	**36.5**
yrl_complin	-	28.9	31.5		29.0	31.7		**41.2**
African languages								
wo_wtb	**87.6**	29.3	35.6					64.8
yo_ytb	-	22.5	31.5					**75.4**
Average (Brazilian languages)	-	23.0	27.1		21.9	25.7		33.8

Table 2: **POS accuracy results for Brazilian languages:** We compare the accuracy of GPT-4 to zero-shot cross-lingual transfer from English language and Portuguese leveraging XLM-R-large multilingual pre-trained language model. Test set A is the original test set found on UD while Test set B are the ones GPT-4 could automatically detect their language to run inference.

provement in accuracy on 6 out of the 7 languages, except for *xav*. The performance improvement is quite large for *urb* (+7.2 on test A, and 12.1 on Test B), and moderate improvement of +3 to +6 for other languages. This experimental result shows that with sufficient monolingual texts, we can increase the performance of the cross-lingual transfer results. However, for the LRLs, such data is scarce. A more effective approach is perhaps to annotate few examples (e.g. 10 or 100 sentences) for training POS taggers to boost the performance (cf. (Lauscher et al., 2020; Hedderich et al., 2020) for a larger boost in performance for token classification tasks in this few-short setting). Regardless, there is a need for better methods to leverage small monolingual data sets.

5 Error analysis

In this section, we provide examples from 2 Brazilian languages (Karo and Guajajara) where the LLMs made errors with the POS tagging. The first line refers to the original sentence, the second line refers to the gold-standard UD POS tag; the third line refers to the GPT-4 POS tag.)

In example (1), the auxiliary verb (in Karo) has the same orthographic form as the English interjection *okay*. In example (2), the Guajajara verb has (partially) the same orthographic form as the English interjection (*oh*). Due to these similarities, GPT-4 seems to tag the POS for these words according to English instead of the POS tagged in UD for Karo and Guajajara.

1. awero toba **okay**
NOUN VERB **AUX**
NOUN NOUN **INTJ**
2. **Oho** kaapii rehe .
VERB NOUN ADP PUNCT
INT VERB ADV PUNCT

6 Discussion & Conclusions

In our study, we explored how LLMs perform the NLP task of POS tagging for 12 LRLs in Brazil and compared this performance with 2 LRLs in Africa and 2 HRLs (English, Brazilian Portuguese). POS is a well established NLP task and it provides insights about the linguistic structures of the different languages especially when only limited data is available, such linguistic annotations have been shown to improve language understanding and generation for endangered languages (Zhang et al., 2024). Our results indicate that the LLMs (GPT-4) perform worse for LRLs on this task in general but older approaches like language adaptive fine-tuning that leverage multilingual encoder models provides some improvements. However, with the lack of available data, any improvements across methods are limited. Although we focused on 12 Brazilian LRLs, there are many other LRLs which we were not able to cover. Future work can expand this evaluation to more tasks and to other LRLs not only from Brazil but from other regions around the world as well.

7 Limitations

Due to limited space, we only focused on POS tagging for this paper but there is a need to explore how LLMs perform other NLP tasks for LRLs. We only evaluated ChatGPT in the zero-shot learning setting but we do not have comparisons with other recent multilingual LLMs, e.g., BLOOM (Scao et al., 2022), and Gemini, in various other learning scenarios. While some of these models are currently less accessible for large-scale evaluations, our plan is to include more models and learning settings along the way to strengthen our evaluations and comparisons in the future. Finally, the current work only evaluates ChatGPT in terms of performance over NLP tasks in different languages. To better characterize ChatGPT and LLMs, other evaluation metrics should also be investigated to report more complete perspectives for multilingual learning, including but not limited to adversarial robustness, biases, toxic/harmful content, hallucination, accessibility, development costs, and interoperability.

8 Ethics Issues

Since we used publicly available data sets, we do not foresee any major issues in terms of ethical concerns.

Acknowledgements

Atul Kr. Ojha would like to acknowledge the support of the Science Foundation Ireland (SFI) as part of Grant Number SFI/12/RC/2289_P2 Insight_2, Insight SFI Centre for Data Analytics and CA21167 COST Action UniDive (by COST (European Cooperation in Science and Technology). David Adelani acknowledges the support of DeepMind Academic Fellowship programme.

References

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023a. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Maxamed Axmed, Kalika Bali, et al. 2023b. Megaverse: Benchmarking large language models across languages, modalities, models and tasks. *arXiv preprint arXiv:2311.07463*.

Alexandra Y. Aikhenvald. 2002. *Language contact in Amazonia*. Oxford University Press.

Vesa Akerman, David Baines, Damien Daspit, Ulf Hermjakob, Tae Young Jang, Colin Leong, Michael Martin, Joel Mathew, Jonathan Robie, and Marcus Schwarting. 2023. The ebible corpus: Data and model benchmarks for bible translation for low-resource languages. *ArXiv*, abs/2304.09919.

Jesujoba Alabi, Kwabena Amponsah-Kaakyire, David Adelani, and Cristina España-Bonet. 2020. Massive vs. curated embeddings for low-resourced languages: the case of Yorùbá and Twi. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2754–2762, Marseille, France. European Language Resources Association.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Alan Ansell, Edoardo Ponti, Anna Korhonen, and Ivan Vulić. 2022. Composable sparse fine-tuning for cross-lingual transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1778–1796, Dublin, Ireland. Association for Computational Linguistics.

Lyle Campbell. 2012. Typological characteristics of south american indigenous languages. *The indigenous languages of South America: A comprehensive guide*, pages 259–330.

Ethan C. Chau and Noah A. Smith. 2021. Specializing multilingual language models: An empirical study. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 51–61, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

R. M. W. Dixon and Alexandra Y. Aikhenvald. 1999. *The Amazonian Languages*. Cambridge University Press.

Abteen Ebrahimi and Katharina Kann. 2021. How to adapt your pretrained multilingual model to 1600 languages. 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)*, pages 4555–4567, Online. Association for Computational Linguistics.

Fabrício Ferraz Gerardi, Stanislav Reichert, Carolina Aragon, Lorena Martín-Rodríguez, Gustavo Godoy, and Tatiana Merzhevich. 2022. Tudet: Tupían dependency treebank.

Akshat Gupta. 2022. On building spoken language understanding systems for low resourced languages. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 1–11, Seattle, Washington. Association for Computational Linguistics.

Mika Hämäläinen, University of Helsinki, Niko Partanen, and Khalid Alnajjar, editors. 2021. *Multilingual Facilitation*. University of Helsinki.

Michael A. Hedderich, David Adelani, Dawei Zhu, Jesujoba Alabi, Udia Markus, and Dietrich Klakow. 2020. Transfer learning and distant supervision for multilingual transformer models: A study on African languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2580–2591, Online. Association for Computational Linguistics.

Kees Hengeveld et al. 2007. Parts-of-speech systems and morphological types. *ACLC Working Papers Volume 2, issue*, page 31.

Viet Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir R. Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2022. Crosslingual generalization through multitask finetuning. *ArXiv*, abs/2211.01786.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462, Online. Association for Computational Linguistics.

Jessica Ojo, Kelechi Ogueji, Pontus Stenetorp, and David I Adelani. 2023. How good are large language models on african languages? *arXiv preprint arXiv:2311.07978*.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Aryon Dall'lgna Rodrigues. 1986. *Línguas Brasileiras: para o conhecimento das línguas indígenas*. São Paulo: Edições Loyola.

A. Ustun, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas

Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. *ArXiv*, abs/2402.07827.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Salih Furkan Akkurt, Gabrielė Aleksandravičiūtė, Ika Alfina, Avner Algom, Khalid Alnajjar, Chiara Alzetta, Erik Andersen, Lene Antonsen, Tatsuya Aoyama, Katya Aplonova, Angelina Aquino, Carolina Aragon, Glyd Aranes, Maria Jesus Aranzabe, and et al. 2023. Universal dependencies 2.13. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Kexun Zhang, Yee Man Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024. Hire a linguist!: Learning endangered languages with in-context linguistic descriptions. *ArXiv*, abs/2402.18025.

A Models

For the experiments, we consider three approaches that are popular in the zero-shot setting since we lack training data for the Brazilian languages.

Prompting GPT-4 GPT-4 [6] is a large language model developed by pre-training on a large amount of texts and code from the web, followed by instruction prompt tuning based on human feedback. We prompt GPT-4 using a similar prompt provided by Lai et al. (2023) where the model is provided a task description before the input. We provide the details in Appendix B.

Cross-lingual transfer We trained a POS tagger individually for English and Portuguese, and perform zero-shot transfer on other languages. We make use of the XLM-R-large (or simply, XLM-R) (Conneau et al., 2020) for training the models. XLM-R has been pre-trained on 100 languages of the world with over 2TB pre-training corpus size but this corpus does not include any indigenous Brazilian languages.

Language Adaptive Fine-tuning (LAFT) We leverage LAFT for an effective cross-lingual transfer by first adapting XLM-R-large model to a new language with limited amount of monolingual data (Alabi et al., 2020; Pfeiffer et al., 2020; Chau and Smith, 2021). This method was proven to be very effective for low-resource languages (Adelani et al., 2021; Muller et al., 2021). We make use of the Bible data as the fine-tuning corpus since it is the largest we found for these languages. We only found 7 (out of 12 Brazilian languages) languages with the Bible corpus. Similar to Ebrahimi and Kann (2021), we examine the effectiveness of this small pre-training corpus with 8K-34K sentences. Pfeiffer et al. (2020) showed that this approach can significantly boost cross-lingual transfer. However, it is not parameter-efficient like the MAD-X they proposed. On the other hand, Ebrahimi and Kann (2021) argued that simple adaptation to a new language is more effective than MAD-X especially when using the Bible corpus for adaptation and we follow this recommendation in our evaluation.

Hyper-parameter of experiments For the cross-lingual and LAFT experiments, we used HuggingFace transformers (Wolf et al., 2020) and A100 Nvidia GPU for fine-tuning the models. For the LAFT, we train for 3 epochs on one GPU while for cross-lingual, we fine-tune English and Portuguese individually using a batch size of 64, with gradient accumulation of 2, and a training epoch of 10.

B Prompt Template

Table 3 provides the prompt template we used for GPT-4 evaluation.

[6] https://chat.openai.com/

	Prompt
Task Description	Please provide the POS tags for each word in the input sentence. The input will be a list of words in the sentence. The output format should be a list of tuples, where each tuple consists of a word from the input text and its corresponding POS tag label from the tag label set: [”ADJ”, ”ADP”, ”ADV”, ”AUX”,”CCONJ”, ”DET”, ”INTJ”, ”NOUN”, ”NUM”,”PART”, ”PRON”, ”PROPN”, ”PUNCT”,”SCONJ”, ”SYM”, ”VERB”, ”X”].
Note	Your response should include only a list of tuples, in the order that the words appear in the input sentence, with each tuple containing the corresponding POS tag label for a word.
Input	[”What”, ”if”, ”Google”, ”Morphed”, ”Into”, ”GoogleOS”, ”?”]
Output	[(”What”, ”PRON”), (”if”, ”SCONJ”), (”Google”, ”PROPN”), (”Morphed”, ”VERB”), (”Into”, ”ADP”), (”GoogleOS”, ”PROPN”), (”?”, ”PUNCT”)]

Table 3: **Prompt template used for POS tagging** based on Lai et al. (2023). An example prediction by GPT-4