Advancing Generative AI for Portuguese with Open Decoder Gervásio PT*

Rodrigo Santos, João Silva, Luís Gomes, João Rodrigues, António Branco
University of Lisbon
NLX - Natural Language and Speech Group, Department of Informatics
Faculdade de Ciências, Campo Grande, 1749-016 Lisboa, Portugal
{rsdsantos, jrsilva, luis.gomes, jarodrigues, antonio.branco}@fc.ul.pt

Abstract

To advance the neural decoding of Portuguese, in this paper we present a fully open Transformer-based, instruction-tuned decoder model that sets a new state of the art in this respect. To develop this decoder, which we named Gervásio PT*, a strong LLaMA 2 7B model was used as a starting point, and its further improvement through additional training was done over language resources that include new instruction data sets of Portuguese prepared for this purpose, which are also contributed in this paper. All versions of Gervásio are open source and distributed for free under an open license, including for either research or commercial usage, and can be run on consumer-grade hardware, thus seeking to contribute to the advancement of research and innovation in language technology for Portuguese.

Keywords: Portuguese, large language model, decoder, open source, open license, open distribution

1. Introduction

This paper presents a model that is the first competitive, 7 billion parameter, fully open and fully documented large language model of the decoder family of Transformers that is prepared specifically for the Portuguese language, by means of instruction tuning, for both the European variant, spoken in Portugal (PTPT) and the American variant, spoken in Brazil (PTBR). These variants have enough differences in terms of vocabulary and syntax to warrant the creation of specialized models.

By being fully open, it is open source and openly distributed for free under a free license, including for research and commercial purposes. By being fully documented, the new datasets that were specifically developed for its construction can be reused, its development can be reproduced, and reported performance scores can be independently assessed. By being fully open and documented, its further development and improvement is openly available to the community.

In the last half decade, the neural approach to natural language processing became pervasive, with virtually any language processing task attaining top performance under the Transformer architecture (Vaswani et al., 2017). Initially proposed and explored in an encoder-decoder setup (Raffel et al., 2020), subsequent research has shown the particular strengths of separate encoder-only and decoder-only solutions (Devlin et al., 2019; He et al., 2021; Brown et al., 2020), with decoders becoming specially notable with the availability of ChatGPT to the general public (Ouyang et al., 2022; OpenAI, 2023).

Among the thousands of natural languages spoken in the world, English is the one whose research is, by a huge margin, better funded and thus the one for which more language resources exist, including the gigantic collections of text that are necessary to train top performing large language models. Consequently, the largest and best performing monolingual models have been developed for this particular language (Touvron et al., 2023b; He et al., 2021).

Seeking to build on the strength of such monolingual models, multilingual models have also been developed. Typically, they are trained over datasets where relatively small portions of data from a few other languages are added to the data from English (Devlin et al., 2019; Scao et al., 2022). Interestingly, these models have shown competitive performance in handling tasks in languages other than English, leveraged by the massive volume of data thus made available and outdoing the meager results that would otherwise be obtained if a monolingual model had been trained only in the data available for those languages alone (Pires et al., 2019).

In order to further mitigate the relative data scarceness impacting the non-English languages, further approaches have been undertaken that include the continuation of the self-supervised training with monolingual data from a specific language. This continuation of causal language modelling (CLM) has been experimented with over multilingual models or even monolingual English models. Research has shown that when such training is appropriately continued, the performance of the resulting model for that specific language exceeds the performance of the baseline model on that language, whose training has not been thus continued (Kaplan et al., 2020; Rodrigues et al., 2023).

SIGUL2024 Workshop, pages 16–26
21-22 May, 2024.

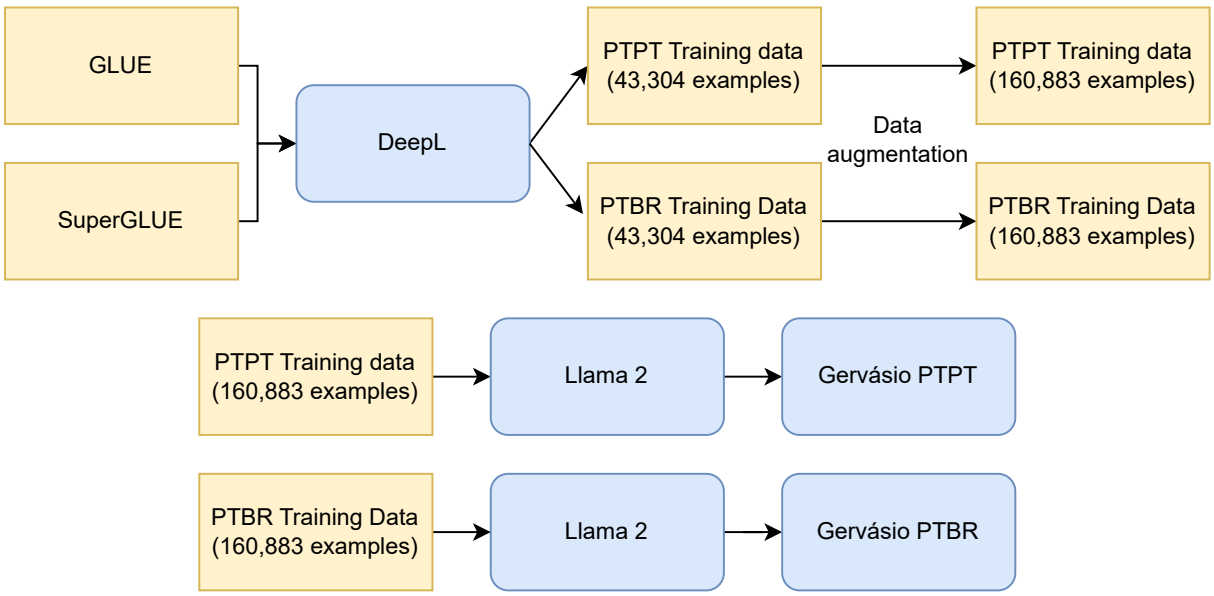

Figure 1: Gervásio PT* Methodology

By exploiting this approach of continuing the training of a previous strong foundation model, we contribute a new model with instruction tuning to foster the technological preparation of the Portuguese language. To the best of our knowledge, this is the first decoder under the Transformer architecture that is both (i) specifically improved for Portuguese, covering two variants of this language, namely PTBR and PTPT, and (ii) fully open, that is it cumulatively complies with all the features of being open source and openly distributed for free under a most permissive license (including for research and for commercial purposes). The model is available at https://huggingface.co/PORTULAN.

To the best of our knowledge and at the time of writing, Gervásio represents the state of the art reported in the literature for open, 7 billion parameter decoders for Portuguese, surpassing the model it is based on as well as other decoders for Portuguese of similar size. The release of Gervásio, alongside the instruction dataset used to train it and which is also a novel contribution of this paper, seeks to contribute to foster research and innovation for the language technology for Portuguese. The methodology employed in this work name be observed in Figure 1.

The remainder of this paper is organized as follows: Related work is covered in the next Section 2; the data used to train and test the model is presented in Section 3; Section 4 describes the decoder for Portuguese created in this study and Section 5 presents and discusses the results of its evaluation. The last Section 6 offers concluding remarks.

2. Related Work

In this section we discuss previous results and resources in the literature that are related to the aim of the present paper. We first address decoders for Portuguese that are publicly reported or publicly distributed, and then we address the available options concerning the base model that can be used to be continued to be trained with Portuguese data.

2.1. Decoders for Portuguese

Looking for decoders specifically developed or improved for Portuguese that are publicly distributed and for which it is possible to find a publicly available report, to the best of our knowledge there can be found only two that, with 7 billion parameters or more, match or surpass the size of Gervásio PT* contributed in the present paper, namely the Sabiá models with 7 and 65 billion parameters (Pires et al., 2023). It is worth noting that: (i) these two models were developed for only one of the variants of Portuguese, PTBR, but not for PTPT; (ii) the 65 billion parameter model is reported in that publication but it is not distributed; and (iii) the 7 billion parameter model is distributed in a non open license, being its reuse restricted to research purposes only.

Other decoders that at the time of writing the present paper can be found of comparable size are not documented, besides being also for only one of the variants of Portuguese, namely PTBR: Boana, Cabra, Cabrita, Canarim.[1]

[1]All on HuggingFace, at lrds-code/boana-7b-instruct, nicolasdec/CabraMistral7b-0.2, 22h/open-cabrita3b, and dominguesm/canarim-7b, respectively.

The other decoders, numbering about a dozen, that can be found for Portuguese have a smaller size, and are also only for PTBR. The largest of these, the 3 billion parameter Cabrita mentioned above, is distributed through Hugging Face (HF) and documented in a non peer-reviewed publication (Larcher et al., 2023). The second largest is Aira,[2] with 1.7 billion parameters and based on Bloom. No evaluation results on benchmarks or downstream tasks for it are reported, it has a residual number of downloads from HF and, being based on Bloom, it inherits the restrictions from Bloom's license and it is thus not fully open as Gervásio.

Common to these decoders other than Sabiá, which are of similar or smaller size, is that while they are publicly distributed, no public detailed presentation of them seems to be provided, be it an implementation report or a paper, either in pre-print or in peer-reviewed versions. This hampers knowing, among other aspects, which datasets were used for their training and thus hampers sensible comparison with other related work and models, which may risk being evaluated in datasets where they were trained.

Turning to Sabiá, while there is a paper with its reporting (Pires et al., 2023), this model was developed by a commercial company and the variant with 7 billion parameters is not openly distributed, with its license restricting its use only for research, a restriction inherited from the license of LLaMA 1 (Touvron et al., 2023a), which was taken as its base model. The variant with 65 billion parameter, in turn, does not appear to be publicly distributed. Sabiá is reported to have been obtained by continuing the training of LLaMA 1 both in its 7 billion and 65 billion parameter versions. A third version of Sabiá was trained over GPT-J (Wang, 2021), with 6 billion parameters. All of these were trained for the PTBR variant of Portuguese only.

Looking into the collection of tasks reported to have been used to evaluate Sabiá, one finds a few that are common with the evaluation of Gervásio, such as BoolQ, which were also machine translated into PTBR to evaluate Sabiá. Additionally, Sabiá's authors present its performance scores in a few other downstream tasks whose datasets did not result from machine translation from English ones, but were developed originally in PTBR.

The performance scores from Sabiá's publication are repeated in Section 5, side by side with related scores of the Gervásio PTBR, for American Portuguese. Against this background, and as it will be discussed at length in that Section, at the time of this writing and to the best of our knowledge, Gervásio offers the state of the art in terms of **fully open** decoders specifically improved for Portuguese in both PTPT and PTBR variants, and it is the first 7 billion parameter decoder specifically developed and distributed for the PTPT variant.

[2]On HF at nicholasKluge/Aira-2-portuguese-1B7.

2.2. Base Models

In this connection, it is worth noting also that not only Gervásio happens to be the top performing 7 billion open decoder for Portuguese, but also that it adopted the best possible setup and codebase available at the time of its development given the goals and requirements assumed for its construction.

There are a number of multilingual decoders reported in the literature, such as mBART (Liu et al., 2020), XLM-R (Conneau et al., 2020), mT5 (Xue et al., 2021), ByT5 (Xue et al., 2022), XGLM (Lin et al., 2022), mGPT (Shliazhko et al., 2023), Bloom (Scao et al., 2022), and LLaMA (Touvron et al., 2023b), to which the promising English open models Mistral (Jiang et al., 2023) and Pythia (Biderman et al., 2023) were added in our considerations of the options available. From these possibilities, many had to be excluded given their non-open license, leaving only those from the Mistral, Bloom, Pythia, and LLaMA families as viable bases on which to build Gervásio.

From these, we decided to leave out Mistral given that, unlike the others, it is indicated to have been developed with no guardrails or other possible state-of-the-art preventive measures available that could help mitigate possible ethical issues.

From the remainder three models left, Bloom is distributed under a RAIL license,[3] which hampers its use in some important application domains, such as law and healthcare, and thus it was left aside.

Finally, as LLaMA models appear to generally deliver better performance than similarly sized Pythia models in the Hugging Face's Open LLM Leaderboard,[4] we adopted LLaMA for our base model. In this leaderboard, LLaMA appears as superior to all the other models mentioned above, except possibly to Mistral, for which it is a matching or close alternative option, with the important advantage over Mistral though of safeguarding ethical aspects to the extent possible given the current status of knowledge concerning foundation models.

3. Data

In this section we present the datasets we developed or reused to train and evaluate Gervásio.

[3]https://huggingface.co/spaces/bigscience/license

[4]https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

3.1. Developed Datasets

To benefit from the advantages of instruction tuning over standard supervised fine-tuning (Wei et al., 2022), and to keep some alignment with mainstream benchmarks for English, we resorted to tasks and respective datasets in the GLUE (Wang et al., 2018) and the SuperGLUE (Wang et al., 2019) collections.

Task selection We selected those datasets where the outcome of their machine translation into Portuguese could preserve, in the target language, the linguistic properties at stake and thus be acceptable for the purposes of this paper.

For instance, the COLA dataset from the GLUE benchmark contains examples of grammatical and non-grammatical expressions from English. This dataset had to be put aside given that an automatic machine translator typically delivers grammatical expressions in the target language, even if the source expression is not grammatical, defeating the purpose of the benchmark.

From GLUE, we resorted to the following four tasks: (i) MRPC (paraphrase detection), (ii) RTE (recognizing textual entailment), (iii) STS-B (semantic textual similarity), and (iv) WNLI (coreference and natural language inference). And from SuperGLUE, we included these other four tasks: (i) BoolQ (yes/no question answering), (ii) CB (inference with 3 labels), (iii) COPA (reasoning), and (iv) MultiRC (question answering).

Task translation To machine translate into European Portuguese and into American Portuguese, we resorted to DeepL,[5] which to our knowledge is the only online service that translates to both of these variants.

Task templates Instruction templates have been manually crafted for each task. These take the various fields in the dataset and arrange them into a prompt by, for instance, appending "Frase 1:" (Eng. "Sentence 1:") before the first sentence of an example in the RTE dataset. A more detailed example is provided below in the Annex A.

Training data For continuing causal language modelling (CLM) with Portuguese data, we used the datasets STS-B and WNLI, from GLUE, and BoolQ, CB and MultiRC, from SuperGLUE, machine translated into Portuguese twice, once for PTPT, and another time for PTBR.

For CLM, each training instance includes the task instruction followed by one or more examples taken from the training partition of that task (including the respective gold answers).

[5] https://www.deepl.com

task	#exs.tra	#exs.aug	total
STS-B	5749	5749	11498
WNLI	635	1270	1905
BoolQ	9427	28281	37708
CB	250	500	750
MultiRC	27243	81729	108972
Total #exs	43304	117529	160833
Total #tok pt	17.9M	50.1M	68.0M
Total #tok br	17.8M	50.6M	68.4M

Table 1: Size of translated (tra) and augmented (aug) training datasets, in number of examples (#exs). The number of examples is identical for both variants, since they are translated from EN to PTPT and PTBR. Token counts (#tok) concern examples only and do not include the instruction or the context examples in few-shot mode

Every instance from the training partitions is seen twice during CLM: once where it is the only example in the respective training instance (that is, it is not preceded by other examples — zero-shot mode); and once where it is preceded by other, 1 to n randomly selected examples (few-shot mode), where n is the largest number possible given the sequence length in CLM.[6] Instances, examples, modes and values for n are shuffled.

Statistics on the training datasets are in Table 1. Taking into account the instructions, the examples in few-shot mode and the two subsets, one for zero-shot mode and the other for few-shot mode, altogether, the CLM resorted to a 83 million token dataset (83.1M for PTPT and 83.6 for PTBR) when we trained our model.

Testing data For testing, we reserved the translated datasets MRPC (similarity) and RTE (inference), from GLUE, and COPA (reasoning/qa), from SuperGLUE, which were taken as representatives of three major types of tasks, and were not seen during training in CLM.

Each testing prompt includes the task instruction followed by an instance from the validation partition (without the gold label). This instance may be preceded by zero (in zero-shot prompting) or by a few examples (in few-shot prompting) taken from the training partition (these examples include the respective gold labels).

Augmented datasets Following (Iyer et al., 2023), we employ data augmentation techniques to enhance the size and diversity of our dataset.

[6] Exceptions were BoolQ and MultiRC, which given the size of their examples and the maximum sequence length of the model, allowed zero-shot mode only.

translated tasks	#exs
MRPC	408
RTE	277
COPA	100
subtotal	785

reused tasks	#exs
ASSIN2 RTE	2448
ASSIN2 STS	2448
BLUEX	178
ENEM 2022	118
FaQuAD	63
subtotal	5255

Table 2: Size of translated and reused testing datasets, in number of examples (#exs). The number of examples is identical for both variants. Reused tasks are pt-br only

This involves repurposing the tasks in various ways, such as generation of answers from MultiRC, question generation from BoolQ, and other relevant modifications. These are presented in the Annex B. Table 1 summarizes the number of examples in the augmented datasets we arrived at. We did not perform data augmentation for any dataset reserved for testing.

3.2. Reused Datasets

For further testing our decoder, in addition to the testing data described above, we also reused some of the datasets that had been resorted to by (Pires et al., 2023) for American Portuguese to test the Sabiá model and that were originally developed with materials from Portuguese: ASSIN 2 RTE (entailment) and ASSIN 2 STS (similarity) (Real et al., 2020), BLUEX (question answering) (Almeida et al., 2023), ENEM 2022 (question answering) (Nunes et al., 2023) and FaQuAD (extractive question-answering) (Sayama et al., 2019). To secure comparability with that model, we filtered out these datasets and prepared their test instances as indicated in the Annex of the Sabiá paper.[7]

Statistics on the testing datasets are show in Table 2.

4. Models

The Gervásio models are based on the LLaMA 2 (Touvron et al., 2023b) model with 7 billion parameters. LLaMA 2 is a open-sourced decoder-based Transformer that has achieved state-of-the-art results in various natural language processing tasks in the English language. In comparison with previous decoder-based models, such as LLaMA 1 (Touvron et al., 2023a), the main reasons for its superiority are the use of a larger context length of 4096 tokens and the extensive volume of data it was trained on, a volume that is currently lacking for the Portuguese language. More specifically, the LLaMA 2 model is pretrained using 2 *trillion* tokens from publicly available sources. The Gervásio models aim to advance generative AI capacity to handle the Portuguese language by further pretraining it on the data we have curated for Portuguese language variants.

Regarding the details of the decoder architecture, the model has a hidden size of 4096 units, an intermediate size of 11,008 units, 32 attention heads, 32 hidden layers, and a tokenizer obtained using the Byte-Pair Encoding (BPE) algorithm implemented with SentencePiece (Kudo and Richardson, 2018), featuring a vocabulary size of 32,000.

We adopted the LLaMA 2 implementation provided by Hugging Face (Wolf and et al., 2020) as our codebase. For this purpose, we employed the Transformers library in conjunction with Accelerate (Gugger et al., 2022), Flash Attention (Dao et al., 2022) and DeepSpeed (Rasley et al., 2020).

Fine-tuning In accordance with the previously described architecture and pre-trained model, we applied supervised fine-tuning for each variant of Portuguese, PTPT and PTBR. The training objective was causal language modeling (CLM) using the training data specified in Section 3.

It is noteworthy that we implemented the zero-out technique during the fine-tuning process, as outlined in (Touvron et al., 2023b). Specifically, while the entire prompt received attention during fine-tuning, only the response tokens were subjected to back-propagation.

In terms of hyper-parameters, we aimed to closely match those utilized in (Touvron et al., 2023b). Consequently, both models were trained with a learning rate of 2×10^{-5}, a weight decay of 0.1, a two-epoch training regime without warm-up, and to ensure the same number of tokens back-propagated per step, we employed an input sequence of 512 tokens with a batch size of 16 and 16 accumulation steps.

Due to hardware limitations that imposed a shorter sequence length (512) compared to the base model (4096), instead of the typical practice of concatenating all training examples and then dividing them into batches with the same input sequence length, we separated each example individually. In other words, each example occupies

[7] We did not reuse TweetSentBR because its distribution is discontinued; ENEM Challenge because it is very similar to ENEM 2022, which was already on board; and FaQuAD because its domain is very narrow (viz. higher education institutions).

the full input sequence length.

To achieve this, we adapted the tokenizer of the base model to accept padding to allow grouping examples with different size into batches while preserving the original input sequence length.

Considering the substantial discrepancy in dataset sizes between the training set and the pretraining corpus used for the base model, with the latter being orders of magnitude larger, and given the language shift from English to Portuguese, we were uncertain about the expected loss behavior. We observed that both models exhibited convergence, featuring in the training steps an initial acceleration in terms of loss decay followed by a deceleration. This behavior suggests the inherent ability of the base model to adapt its focus to a new language, especially considering that the tokenizer was not retrained for Portuguese.

For the model training process, we resorted to an a2-megagpu-16gb Google Cloud A2 VM, equipped with 16 GPUs, 96 vCPUs, and 1.360 GB of RAM. The training of each model took approximately two hours.

5. Evaluation and Discussion

To assess Gervásio models, we resorted to the test sets introduced above in Section 3. For every task under evaluation, we use the respective evaluation metrics commonly found in the literature, typically the F1 score or the Pearson correlation coefficient, as indicated below.

In this connection, it is worth noting that in a text generation task where the generated text is evaluated against a gold label, various responses may arise in which the generated text does not match any of the predefined classes. In such cases, the response was considered different from the correct label and thus incorrect. To maintain the integrity of the generated text, which corresponds to the final label, in tasks where the answer is a word, like "sim" ou "não" (Eng. "yes" or "no"), we only considered the first word provided as the response, after trimming any leading whitespace. In tasks where the outcome involve classes consisting of single digit numeric value, only the first digit is accepted as the response.

Regarding the hyper-parameters relevant in inference time for the decoder to generate responses to the test tasks, we employed a temperature setting of 1.0, greedy decoding, a beam search value of 1, and applied top-k filtering with a threshold of 50.

Each performance score reported below is the average of the outcome of three independent runs using different seeds.

Tasks from GLUE and SuperGLUE Each language variant of Gervásio was evaluated with the respective translated version of the test tasks selected from GLUE and SuperGLUE. The evaluation scores are displayed in Table 3.

Model	MRPC	RTE	COPA
Gervásio ptbr	**0.7822**	**0.8321**	0.2134
LLaMA 2	0.0369	0.0516	0.4867
LLaMA 2 Chat	0.5432	0.3807	**0.5493**
Gervásio ptpt	**0.7273**	**0.8291**	**0.5459**
LLaMA 2	0.0328	0.0482	0.3844
LLaMA 2 Chat	0.5703	0.4697	0.4737

Table 3: F1 scores for ptbr and ptpt tasks translated from GLUE and SuperGLUE, not seen during training. Best scores for each task are in bold

The LLaMA 2 and LLaMA 2 Chat models were evaluated by us over the Portuguese data for both variants by following also the same approach used for Gervásio, described above.

Other downstream tasks Gervásio PTBR was also evaluated in the downstream tasks whose data sets were not translated from English but originally developed for Portuguese. The evaluation scores are displayed in Table 4. For Sabiá, the results presented therte are those reported in the respective publication (Pires et al., 2023).

Discussion The first important result worth underling is that Gervásio largely outperforms its baseline LLaMA 2 in all tasks by both models, as reported in Table 3, except for the PTBR model on the COPA task.

This demonstrates that it was rewarding to continue the causal language modeling of LLaMA 2 with the Portuguese data, even though LLaMA 2 had been pre-trained over a overwhelming majority of English data, and also despite the Portuguese dataset used to continue its pre-training being tiny (1.8 billion tokens) when compared to the one used for LLaMA 2 (2 trillion tokens).[8]

[8]To further examine the outlier score of COPA in ptbr, we proceeded with cross evaluation. The PTPT model shown quite similar scores for both the PTPT and PTBR datasets, which seems to indicate that the possible cause for the outlier value did not occur with the construction of the PTBR dataset. The PTBR model, in turn, run over the PTPT testset, shown again an outlier score, similar to the outlier score obtained for PTBR, which may indicate the root of the difference occurs with the training sets. In fact, the base model LLaMA was trained on 2.8 Billion tokens of Portuguese (0.09% of the total 2 Trillion tokens used for its training in English), where PTBR texts were most probably in much superior number than PTPT ones, given the respective demographics. This indicates in

Model	ENEM 2022	BLUEX	RTE	STS
Gervásio ptbr	0.1977	0.2640	**0.7469**	**0.2136**
LLaMA 2	0.2458	0.2903	0.0913	0.1034
LLaMA 2 Chat	0.2232	0.2959	0.5546	0.1750
Sabiá-7B	**0.6017**	**0.7743**	0.6487	0.1363

Table 4: Evaluation (F1 for RTE, Accuracy for ENEM 2022 and BLUEX, Pearson for STS) in data sets originally developed for American Portuguese, not seen during training. Best scores in bold

Another result from the values in Table 3 is aligned with similar results that had been found in (Rodrigues et al., 2023). The different performance scores of Gervásio for each of the language variants reinforce that it is relevant to have a specific version of the model for each language variant.

Turning to Table 4, one finds the results obtained with datasets originally developed for PTBR, thus not having been obtained by machine translation. For two of the tasks, namely RTE and STS, the performance scores obtained here repeat the same contrast obtained with the other test datasets translated into Portuguese whereby Gervásio PTBR greatly outperforms its baseline LLaMA 2.

For the two other tasks, ENEM 2022 and BLUEX, in turn, Gervásio does not show clear advantage over its starting model. This difference in performance seems to be justified by the different type of tasks in each group. Gervásio seems to cope better with tasks concerned with comparing sentences (RTE, with binary decision, and STS, with 6-way decision), rather than with tasks concerned with question answering (ENEM2022, with 5-way, and BLUEX, with 4-way), likely less exercised in the training set.

The scores of Sabiá in Table 4 invite to contrast them with Gervásio's but such comparison needs to be taken with some caution.

First, these are a repetition of the scores presented in the respective paper (Pires et al., 2023), which only provide results for a single run of each task, while scores of Gervásio are the average of three runs, with different seeds.

Second, the evaluation methods adopted by Sabiá are *sui generis*, and different from the one's adopted for Gervásio. Following Gervásio's decoder nature as a generative model, our scores are obtained by matching the output generated by Gervásio against the ground labels. Sabiá, in turn, followed a convoluted approach away from its intrinsic generative nature, by "calculating the likelihood of each candidate answer string based on the input text and subsequently selecting the class with the highest probability" (Pires et al., 2023, p.231), which forces the answer to be one of the possible classes and likely facilitates higher performance scores than Gervásio's, whose answers are generated without constraints.

Third, to evaluate Sabiá, the examples included in the few-shot prompt are hand picked, and identical for every test instance in each task (Pires et al., 2023, p.4). To evaluate Gervásio, the examples were randomly selected to be included in the prompts.

Even taking these considerations into account, it is noticeable that the results in Table 4 indicate performance scores for Gervásio that are clearly better than for Sabiá, over the same two test tasks where it also excels over its starting model.

Given that Gervásio, in addition, is distributed as an fully open model, and Sabiá is publicly available for research only, all these circumstances seems to speak for Gervásio's advantage in terms of its usage for research and commercial purposes.

Limitations and Potential Negative Impact Large language models come with their own set of limitations and potential for negative impacts. One notable limitation is their dependency on the data they were trained on, which can embed biases into their outputs, potentially perpetuating stereotypes and discriminatory practices.

In this work we make use of curated data, namely the GLUE and SuperGLUE, which mitigates the propagation of the aforementioned issues. Nonetheless, we inherit all the bias and limitations of the Llama 2 model which is the base to the Gervásio model.

6. Conclusion

This paper contributes new, instruction-tuned large language models of the decoder family of Transformers specifically developed for the Portuguese language, as well as the instruction datasets used to train and evaluate them.

which measure the two training conditions for PTBR and PTPT may differ. Nevertheless, if this larger exposure to PTBR data, by the starting model LLaMA, was the cause for the outlier value with COPA, then it will remain to expalin why the score for MRPC and RTE are in line for both PTBR and PTPT. We leave this for future research.

The models are openly available for free and with no registration required under an MIT license at https://huggingface.co/PORTULAN, where the respective datasets are also openly available for free and with no registration required.

With a 7 billion parameter, these models have an unique set of features for their size. They are fully open: they are open source; and they are openly distributed, under an open license, thus including for either research or commercial purposes. They are the most encompassing models for the Portuguese language: they cover both the European variant, spoken in Portugal, and the American variant, spoken in Brazil; and the model for the European variant it is the first of its class, known in the literature. They show a competitive performance: they outperform other models of similar size publicly reported, thus representing the state of the art. They are fully documented: the new datasets that were specifically developed for its construction can be reused and its development can be reproduced; and reported performance scores can be independently assessed.

By being fully open and fully documented, its further development and improvement is openly available to the community.

Also, given their size, these models can still be run on consumer-grade hardware with technological solutions currently available, thus being a contribution to the advancement of research and innovation in language technology for Portuguese.

Future work will include taking these models as the inaugural members of a future family of fully open decoders for Portuguese with a range of other sizes, and characteristics and for other variants of Portuguese.

Acknowledgements

This research was partially supported by: PORTULAN CLARIN—Research Infrastructure for the Science and Technology of Language, funded by Lisboa 2020, Alentejo 2020 and FCT (PINFRA/22117/2016); ACCELERAT.AI—Multilingual Intelligent Contact Centers, funded by IAPMEI (C625734525-00462629); and GPTPT—Transformer-based Decoder for the Portuguese Language, funded by FCT (CPCA-IAC/AV/478395/2022).

7. Annex A: Template Example

As an example, here we describe the template used for the RTE task in PTPT. In this task, two sentences are given and the task consists in determining whether the first sentence entails the second. Each instance in the dataset contains the fields premise, hypothesis and labels. The template describes how to handle these fields, usually by prepending some string to their contents, as well as defining the initial instruction.

instruction "Nesta tarefa vais receber duas frases. Indica se a primeira frase implica claramente a segunda frase. Ou seja, indica se se conclui que a segunda frase é verdadeira desde que a primeira frase seja verdadeira. Deves responder 'sim' se a primeira frase implica a segunda frase ou deves responder 'não' no caso contrário." (Eng. "In this task you'll receive two sentences. Indicate whether the first sentence clearly entails the second sentence. That is, indicate whether one can conclude that the second sentence is true as long as the first sentence is true. You should answer 'yes' if the first sentence entails the second sentence or 'no' otherwise.")
This is the instruction that is given at the beginning of the input.

premise "Frase 1:" (Eng. "Sentence 1:")
This is placed before the contents of the 'premise' field of the RTE instance.

hypothesis "Frase 2:" (Eng. "Sentence 2:")
This is placed before the contents of the 'hypothesis' field of the RTE instance.

pre-label "Resposta:" (Eng. "Answer:")
This is placed before the answer.

labels "0" $\rightarrow$ "sim", "1" $\rightarrow$ "não"
This is a mapping from the 0/1 labels used in the RTE dataset to the yes/no labels that are asked for in the instructions for the task.

Applying the template above to an instance gives something like what is shown below.

Nesta tarefa vais receber duas frases. Indica se a primeira frase implica claramente a segunda frase. Ou seja, indica se se conclui que a segunda frase é verdadeira desde que a primeira frase seja verdadeira. Deves responder 'sim' se a primeira frase implica a segunda frase ou deves responder 'não' no caso contrário
Frase 1: Em 1969, redigiu o relatório que propunha a expulsão do partido do grupo Manifesto. Em 1984, após a morte de Berlinguer, Natta foi eleito secretário do partido.
Frase 2: A Natta apoiou o grupo do Manifesto.
Resposta: não

In addition, a separator string formed by 3 to 5 consecutive '=' (equals) symbols is inserted between each instance in the training data. And, during few-shot inference, each instance is headed by "Exemplo n" (Eng. "Example n"), with increasing

n, and within each instance its few-shot examples are delimited by a separator string formed by 3 or 4 consecutive '-' (hyphen) or '*' (asterisk) symbols.

8. Annex B: Instruct Training Tasks

The base tasks and their augmented counterparts that together form the training data are:

STS-B for semantic textual similarity, with augmented **STS-B Aug1** for generation of a sentence with a STS score of 0/1/2/3/4/5

WNLI for coreference and natural language inference, with augmented **WNLI Aug1** for generating an hypothesis with Positive/Negative inference, and **WNLI Aug2** for generating a premise with Positive/Negative inference

BoolQ for Yes/No question answering, with augmented **BoolQ Aug1** for question generation with Yes/No answer based on an excerpt, and **BoolQ Aug2** for excerpt generation with Yes/No answer to a question

CB for inference with labels Entailment (E), Contradiction (C) and Neutral (N), with augmented **CB Aug1** for generating an hypothesis with label E/C/N, and **CB Aug2** for generating a premise with label E/C/N

MultiRC for question answering, with augmented **MultiRC Aug1** for question generation, **MultiRC Aug2** for excerpt generation, and **MultiRC Aug3** for answer generation

9. Bibliographical References

Thales Sales Almeida, Thiago Laitz, Giovana K. Bonás, and Rodrigo Nogueira. 2023. BLUEX: A benchmark based on Brazilian leading universities entrance exams. *arXiv preprint arXiv:2307.05410*.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. *arXiv preprint arXiv:2304.01373*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

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.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Sylvain Gugger, Lysandre Debut, Thomas Wolf, Philipp Schmid, Zachary Mueller, and Sourab Mangrulkar. 2022. Accelerate: Training and inference at scale made simple, efficient and adaptable. https://github.com/huggingface/accelerate.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*.

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, et al. 2023. OPT-IML: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212.12017*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent

subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

Celio Larcher, Marcos Piau, Paulo Finardi, Pedro Gengo, Piero Esposito, and Vinicius Caridá. 2023. Cabrita: closing the gap for foreign languages.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, et al. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Desnes Nunes, Ricardo Primi, Ramon Pires, Roberto Lotufo, and Rodrigo Nogueira. 2023. Evaluating GPT-3.5 and GPT-4 models on Brazilian university admission exams. *arXiv preprint arXiv:2303.17003*.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Ramon Pires, Hugo Abonizio, Thales Sales Almeida, and Rodrigo Nogueira. 2023. Sabiá: Portuguese large language models. In *Intelligent Systems*, pages 226–240, Cham. Springer Nature Switzerland.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? *arXiv preprint arXiv:1906.01502*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21:5485–5551.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506.

Livy Real, Erick Fonseca, and Hugo Gonçalo Oliveira. 2020. The ASSIN 2 shared task: A quick overview. In *Computational Processing of the Portuguese Language*, pages 406–412, Cham. Springer International Publishing.

João Rodrigues, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso, and Tomás Osório. 2023. Advancing neural encoding of Portuguese with Transformer Albertina PT-*. In *Progress in Artificial Intelligence*.

Hélio Fonseca Sayama, Anderson Viçoso Araujo, and Eraldo Rezende Fernandes. 2019. FaQuAD: Reading comprehension dataset in the domain of Brazilian higher education. In *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 443–448.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. 2023. mGPT: Few-shot learners go multilingual. *arXiv preprint arXiv:2204.07580*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Ben Wang. 2021. Mesh-Transformer-JAX: Model-parallel implementation of transformer language model with JAX. https://github.com/kingoflolz/mesh-transformer-jax.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Thomas Wolf and et al. 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.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

10. Language Resource References

Thales Sales Almeida and Thiago Laitz and Giovana K. Bonás and Rodrigo Nogueira. 2023. *BLUEX: A benchmark based on Brazilian Leading Universities Entrance eXams*. HuggingFace.

Desnes Nunes and Ricardo Primi and Ramon Pires and Roberto Lotufo and Rodrigo Nogueira. 2023. *ENEM 2022 (Evaluating GPT-3.5 and GPT-4 Models on Brazilian University Admission Exams)*. GitHub.

Real, Livy and Fonseca, Erick and Gonçalo Oliveira, Hugo. 2020. *ASSIN 2 (The ASSIN 2 Shared Task: A Quick Overview)*. HuggingFace.

Sayama, Hélio Fonseca and Araujo, Anderson Viçoso and Fernandes, Eraldo Rezende. 2019. *FaQuAD: Reading Comprehension Dataset in the Domain of Brazilian Higher Education*. HuggingFace.

Touvron, Hugo and Martin, Louis and Stone, Kevin and Albert, Peter and Almahairi, Amjad and Babaei, Yasmine and Bashlykov, Nikolay and Batra, Soumya and Bhargava, Prajjwal and Bhosale, Shruti and others. 2023. *Llama 2 7B (Llama 2: Open foundation and fine-tuned chat models)*. HuggingFace.

Wang, Alex and Pruksachatkun, Yada and Nangia, Nikita and Singh, Amanpreet and Michael, Julian and Hill, Felix and Levy, Omer and Bowman, Samuel. 2019. *Superglue: A stickier benchmark for general-purpose language understanding systems*. HuggingFace.

Wang, Alex and Singh, Amanpreet and Michael, Julian and Hill, Felix and Levy, Omer and Bowman, Samuel. 2018. *GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding*. HuggingFace.