Exploring Hybrid Pre-training for Automatic Essay Scoring

Miguel M. Carpi[1ǂ], **Igor C. Silveira**[1ǂ], **Denis D. Mauá**[1], **Marcelo Finger**[1]

[1]Department of Computer Science
Institute of Mathematics, Statistics and Computer Science (IME)
University of São Paulo (USP) – São Paulo, SP – Brazil

{miguel,igorcs,ddm,mfinger}@ime.usp.br

Abstract. *Alternatives to Large Language Models have been proposed to develop smaller models that require substantially less training data. In this paper, we propose a monolingual (Portuguese) Hybrid Transformer model trained with only 260M words, whose size is comparable to that of small Encoder-based models. After pre-training, we fine-tune our model and compare it against 11 existing models on the AES-ENEM dataset — an Automatic Essay Scoring benchmark in which models are required to evaluate five distinct textual dimensions. Our experiments demonstrate that our model is always competitive with the (bigger) best available model, despite its smaller scale.*

1. Introduction

Since the introduction of BERT [Devlin et al. 2019], Large Language Models (LLMs) based on the Transformer architecture [Vaswani et al. 2017] have been at the center of Natural Language Processing for most, if not all, languages. This applies to Portuguese, which has at least one (monolingual) model from each of the Transformer families: Encoder [Souza et al. 2020, Rodrigues et al. 2023, Zago and Pedotti 2024], Decoder [Pires et al. 2023, Corrêa et al. 2025a], and Encoder-Decoder [Carmo et al. 2020].

Such models for Portuguese are influenced by (at least) two different sources. The first source involves architectural innovations, which mostly come from English models, such as the changes from BERT to RoBERTa [Liu et al. 2019] and DeBERTa [He et al. 2020]. The second source is the release of Portuguese corpora, such as the GigaVerbo [Corrêa et al. 2025b] and *Corpus* Carolina [Finger et al. 2025]. In general, developers of models for Portuguese tend to follow the neural scaling law, where bigger models — requiring more parameters, data and training time — are assumed to be better.

Although recent advances have largely favored increasingly larger models, smaller language models remain an attractive alternative for specialized tasks or when computational costs are constrained, and they have been argued to represent an important direction for the future of AI systems [Belcak et al. 2025]. Reducing computational costs — memory usage, inference time, and training time — is not a new problem and predates the LLM era [Menghani 2023]. Therefore, a vast literature has been devoted to efficient pre-training and fine-tuning using different techniques, such as knowledge distillation [Hinton et al. 2015], low rank adaptation [Hu et al. 2021], and quantization [Lee et al. 2024]. These and related techniques have been successful in creating small BERT models like nanoBERT [Maity et al. 2024] and Tiny BERT [Jiao et al. 2020].

ǂ Equal contribution

Other approaches showed that BERT-like models require less pre-training data. For example, recent work showed that 100 million words can yield state-of-the-art performance on some NLP benchmarks, although presenting some limitations in more complex tasks [Micheli et al. 2020, Zhang et al. 2021, Samuel et al. 2023]. Moreover, in-context learning was demonstrated in smaller decoders and in encoder-only models [Samuel 2024, Schick and Schütze 2021]. Finally, there is ongoing research in unifying encoder and decoder models in a single architecture to exploit the advantages of each other [Charpentier and Samuel 2024, Yu et al. 2024].

In this work, we propose novel Hybrid Transformer-based models extending previous works [Charpentier and Samuel 2024, Yu et al. 2024, Carpi and Finger 2026]. Our model is approximately the size of a traditional Encoder-only model (150M parameters) but is trained on only 260 million words and has a wider context window than similar Encoders (1024 vs. 512). In order to train this model, a masked-causal hybrid training is employed: the training regime alternates between batches of Masked Language Modeling (Encoder pre-training, through a bidirectional attention mask) and Causal Language Modeling (Decoder pre-training, through a causal attention mask). This creates a model that is capable of behaving like both an Encoder and a Decoder.

To assess the effectiveness of the proposed pre-trained model, we fine-tune and evaluate it on the AES-ENEM dataset [Silveira et al. 2024]. This Automatic Essay Scoring (AES) dataset comprises argumentative essays written for a mock Brazilian national university entrance exam, the ENEM. We selected this dataset for two main reasons. First, the task requires the model to adapt to five different aspects of written language, posing a significant and comprehensive challenge. Second, 15 different LLMs were benchmarked in this dataset [Barbosa et al. 2025], allowing for direct comparison against many models with only one experiment. When comparing our best performance against the benchmarked results, we see that: I) in one textual dimension (argumentative quality) it outperformed the others — but it was still statistically similar to the previous best; II) in another dimension (persuasiveness) it had an inferior performance, but the confidence interval still overlaps with the best models; III) in the remaining three dimensions, our model exhibited a performance similar to the best model, losing by only a small margin — remaining inside the confidence interval.

As a by-product of our different combinations of training regimes, we can create a spectrum that goes from Decoders to Encoders. Thus, we are capable of evaluating the influence of the training regime on different traits, which has never been assessed before. Additionally, we contribute to AES with a new baseline performance: the performance of a Transformer model without pre-training.

To summarize, our main contributions are:

- presenting a family of Hybrid models trained with limited data;[1]
- showing that our best models are statistically comparable to the SOTA;
- filling the gap in AES regarding the importance of the pre-training regime;
- presenting a Transformer baseline performance for the AES-ENEM benchmark.

The remainder of the document is divided as follows: related work, presentation of our hybrid models, methodology, results, and conclusion.

[1]Available at: `https://github.com/mmcarpi/flexqwen`

Table 1. Encoder-based and Decoder (Small LLMs) benchmarked in [Barbosa et al. 2025] together with their context and model size.

Model Name	Architecture	Context	#Params
BERTimbau Base	Encoder	512	109M
BERTuguês	Encoder	512	110M
mBERT	Encoder	512	178M
BERTimbau Large	Encoder	512	334M
Albertina	Encoder	512	1.5B
Phi3	Decoder	4k	2B
Tucano	Decoder	4k	2.4B
Llama3	Decoder	128k	8B
Phi4	Decoder	16k	14.7B

2. Related Work

In this section, we discuss the two topics related to our research. We started by situating our model in the LLM landscape. Then, we review models for AES in Portuguese.

Traditionally, Transformer-based models can be categorized according to their architecture — Encoder, Decoder, or Encoder-Decoder — and pre-training regime — Masked Language Modeling or Causal Language Modeling. Pure-Encoder models have an Encoder architecture and are pre-trained using Masked Language Modeling. Such models usually excel at language understanding and are used for tasks such as text classification. One advantage of this group is that they are relatively small, the most popular ones used for Portuguese — BERTimbau [Souza et al. 2020], BERTuguês [Zago and Pedotti 2024], and Albertina [Rodrigues et al. 2023] — range from 109M to 1.5B parameters. Their disadvantage is their limited context window of 512 tokens.

Pure-Decoder models are those that use Decoder architecture and are pre-trained using Causal Language Modeling. This class can be further divided. The first division is API Decoders, as they are usually accessed through APIs due to their huge size and are often proprietary, making them unable to be fine-tuned on everyday desktops. Examples of these models are Sabiá3 [Pires et al. 2023], GPTo1 [OpenAI et al. 2024], and DeepSeek-R1 [DeepSeek-AI et al. 2025]. The second group is often called Small Large Language Models because they are smaller than the previous group, but still orders of magnitude bigger than Encoder models. Examples of these models are some variants of Llama3 [Touvron et al. 2023], Qwen3 [Yang et al. 2025], Phi [Abdin et al. 2024a, Abdin et al. 2024b], and Tucano [Corrêa et al. 2025a]. These models can be fine-tuned, but they usually require high-end equipment that is not commonly available.

Finally, Encoder-Decoder models have the architecture of the same name and are pre-trained using both of the pre-training regimes. Although they are available in Portuguese through the PTT5 models [Carmo et al. 2020, Piau et al. 2024], they are not as widely adopted as the previous groups.

Our Hybrid model differs from all the previous ones as it uses a single architecture and is trained with both pre-training regimes, thus being able to behave like an encoder- or decoder-only model. Additionally, it has a smaller size, as an Encoder, it is trained on a small portion of data and has an increased context of 1024 tokens that can be further

extended thanks to the employment of positional encoding through RoPE [Su et al. 2024].

AES of ENEM-like essays has a tradition of employing all the stages of Natural Language Processing technologies. In the early days, approaches were based on word occurrence [Bazelato and Amorim 2013] and features that were proxies for textual quality [Amorim and Veloso 2017]. In the neural network era, feature extraction was delegated through fine-tuning, first to fixed dense representations and LSTMs [Fonseca et al. 2018, Marinho et al. 2022], then to Encoder models [Silveira et al. 2024], and later to Small LLMs [Silveira et al. 2025a]. Finally, modern conversational agents with reasoning were employed through zero-shot prompting [Barbosa et al. 2025]. Noteworthy, due to their smaller size, Encoder Transformers are widely employed for Portuguese AES, either directly [Ribeiro et al. 2024, de Sousa et al. 2024, Rossman et al. 2026] or indirectly through majority voting [de Lima et al. 2024], Two-Stage Learning [Silveira et al. 2025b], combined representations [Matos and Feltrim 2026], or neuro-symbolic methods [Silveira and Mauá 2026].

Several models, ranging from feature-based to modern reasoning-based API Decoders, have been benchmarked [Barbosa et al. 2025] on the same ENEM-like AES dataset [Silveira et al. 2024]. This allows any new model to be directly compared to them. The models that are comparable to ours and that were benchmarked in the aforementioned work are summarized in Table 1. According to the authors of the benchmarking, there was no model that was consistently better than the others, highlighting that each model may excel in a different dimension of language. Alternatively, the impact of different representations used by AES systems was investigated in [Chalegre et al. 2026]. The authors of the previous study concluded that although contextual representations given by Encoders have advantages in mechanical traits, TF-IDF-based representations could be competitive in thematic traits, which are usually dominated by Decoders in the benchmark.

The present work intersects with AES literature in two different points. The first is by evaluating a novel family of Hybrid Transformer models, which have the same number of parameters as Encoder models but a larger context, in the same benchmark. The second is by evaluating the same LLM with different pre-training regimes, which is still not covered in the literature.

3. Hybrid Model

For the model configuration, a model is instantiated from scratch based on the architectural specifications of Qwen 3 [Yang et al. 2025], rather than from a pre-trained checkpoint. To increase neural network expressivity, we additionally incorporate layer dropout and layer weighting [Charpentier and Samuel 2023]. To support our hybrid pre-training objective, the model contains a total of 152,981,916 trainable parameters, featuring a vocabulary size of 64,000, a hidden dimension of 768, an intermediate feed-forward dimension of 3,072, 12 transformer blocks, 12 attention heads with a head dimension of 64, and 4 key-value groups. Although we could have used the original Qwen 3 tokenizer, or another existing multilingual or Portuguese-specific tokenizer, we trained a custom byte-pair encoding tokenizer [Sennrich et al. 2016] with a vocabulary size of 64,000 on a reserved split of our *corpus* (described below) to ensure tighter alignment between the vocabulary and our pre-training data. Before chunking, text is normalized using Unicode Normalization Form C so that accented characters (e.g., *ã*, *é*) map to unified code points.

Table 2. Model performance on test set by hybrid strategy and objective function: Causal Language Modeling (CLM) and Masked Language Modeling (MLM). Models were evaluated in sequences of 1024 tokens. The corruption for MLM was the same as the one used for training (30%).

Objective	Model	Perplexity	Loss
CLM	causal	**15.34**	**2.73**
	constant	17.94	2.88
	linear	20.42	3.01
	cosine	20.55	3.02
	shift	21.52	3.06
MLM	masked	17.23	2.84
	constant	17.90	2.88
	linear	17.29	2.85
	cosine	17.22	2.84
	shift	**17.04**	**2.83**

This study exclusively utilizes the *Corpus* Carolina v2.0 [Finger et al. 2025], a curated Brazilian Portuguese web dataset, for language model pre-training. We implemented a two-stage filtering pipeline following established data-curation practices for pre-training large language models [Rae et al. 2021]. First, we applied typology-aware length thresholds — excluding the top and bottom 5% of texts — and removed low-entropy documents based on unique word ratios, punctuation density, and extreme word lengths, thus removing texts that are too short, or that contains an excessive number of URLs or markup text. Second, to eliminate near-duplicate documents that can lead to memorization and degraded generalization [Lee et al. 2022, Leveling et al. 2024], we applied MinHash-based Jaccard similarity. This pipeline reduced collectively the *corpus* from approximately 2.07 million to 1.78 million documents. From this refined dataset, a final sample of 1.5 million documents, roughly 320 million words, was extracted and partitioned into 260 million words for model training, 50 million for tokenizer development, and 5 million each for validation and testing.

Furthermore, we investigate a family of hybrid Transformer models capable of functioning as either encoder-only or decoder-only architectures purely through attention mask modifications. Building on prior research in hybrid causal-masked pre-training [Charpentier and Samuel 2024, Yu et al. 2024, Carpi and Finger 2026], we evaluate four objective-mixing strategies that dictate the probability of applying a Causal Language Modeling (CLM) versus a Masked Language Modeling (MLM) objective during training: Shift (a sequential transition from CLM to MLM), Constant (a fixed CLM probability throughout training), Linear (a linear reduction of the CLM probability over time), and Cosine (a cosine decay applied to the CLM probability schedule). These strategies were influenced by past works. The Shift strategy originates from [Yu et al. 2024] which studies taking a causal model and further training it on the masked next-token prediction to achieve an encoder model. The Linear strategy is adapted from [Carpi and Finger 2026] which proposes alternating batches of MLM and CLM optimization. The Constant strategy is an adaptation from [Charpentier and Samuel 2024], where instead of optimizing both CLM and MLM objectives in every batch, we choose with some probability p if a particular batch should be used to optimize the CLM or the MLM objective. To the best

Table 3. Variation space for fine-tuning. The models masked, causal and non-pre-trained correspond to our baselines. Note that we do not fine-tune the causal model with a bidirectional mask nor the masked model with a causal mask.

Parameter	Values
Model	masked, shift, constant, linear, causal, cosine, non-pretrained
Input	essay, full
Pooling	last token, average
Pre-training	bidirectional, causal

of our knowledge, while the cosine decay is widely used as a learning rate scheduler, its application to balancing CLM and MLM objective-mixing ratios has not yet appeared in the literature.

We pre-trained six models — a CLM baseline, an MLM baseline, and the four aforementioned hybrid variants — for four epochs with a batch size of 524,288 tokens. Optimization relied on Muon [Keller Jordan et al. 2024] (learning rate 0.02) for 2D matrices and AdamW [Loshchilov and Hutter 2018] (learning rate 3×10^{-3}) for the embedding layer, applying a 0.95 momentum, 0.02 weight decay, zero dropout, and a cosine-decay scheduler with a 5% warmup. Training featured a progressive sequence length warmup (128, 512, and 1024 tokens across the first three epochs), while validation strictly evaluated 1024-token sequences. For masked and hybrid architectures, inputs were corrupted at a 30% rate using span masking (lengths 3 to 10) [Joshi et al. 2020]. Hybrid objective ratios were governed by either maintaining a fixed 0.55 CLM probability (constant strategy) or dynamically decaying the CLM probability from 0.9 to 0.2 over the first 55% of the training steps (cosine, linear, and shift strategies).

The pre-training results in Table 2 show that our hybrid models performed worse than the decoder-only baseline (causal) in the CLM objective, while in the MLM objective the shift strategy beat the encoder-only (masked). Additionally, the constant strategy resulted in a model that performs roughly the same in both CLM and MLM, and not so distant from the baselines. The other hybrid models present a higher loss in the CLM objective in comparison with their loss in MLM, which is expected since they were trained for longer in the MLM objective.

4. Methodology

To fine-tune and assess the models, we employ the same split used in the benchmarking of the other models [Barbosa et al. 2025] — a filtered version of AES-ENEM [Silveira et al. 2024]. In this filtered version, 385 mock essays were evaluated by two experienced graders according to the five traits of the exam: C1) proper usage of Portuguese, C2) thematic coherence, C3) argumentation quality, C4) cohesion, C5) persuasiveness of conclusion. Each trait is evaluated on a scale from 0 to 200, in steps of 40. Noteworthy, each essay appears twice, as it was graded by two graders.

The models are evaluated according to the Quadratic Weighted Kappa (QWK) metric, which is widely used for AES [Doewes et al. 2023]. This metric ranges from 1 (perfect agreement) to -1 (perfect disagreement), where 0 indicates random agreement. The expected random agreement is calculated based on the marginal distributions of the

Table 4. Comparison of model performance using Bidirectional mode for essay-only input. In bold, the highest scoring model inside each trait and each pooling strategy

Mode: Bidirectional												
model	C1	C2	C3	C4	C5	avg.	C1	C2	C3	C4	C5	avg.
Pool. Strategy:	*average*						*last token*					
constant	**.64**	.22	.33	.47	.40	.41	.44	.13	.16	.36	**.47**	.31
cosine	.54	**.28**	.28	.40	.39	.37	**.59**	.16	.26	**.56**	.31	.37
linear	.49	.02	.22	.44	.04	.24	.51	**.23**	.28	.39	.37	.35
masked	.46	-.05	**.36**	.36	.35	.29	.44	.15	**.31**	.40	.28	.31
shift	.52	-.09	.35	.37	**.41**	.31	.50	.13	.25	.45	.44	.35
non-pretrained	.16	.16	.16	**.48**	.03	.19	.22	.16	.25	.36	.11	.22

labels assigned by the annotators being evaluated, and disagreement is penalized quadratically based on the distance between the ordinal labels. The reported performance of the model is the average of QWK with each grader. Finally, we followed the same approach from the benchmark and calculated the 95% confidence interval ($N = 1000$) of each model, so that we could assert whether the differences in performance are significant.

Our models were fine-tuned using the AdamW optimizer (learning rate of 5×10^{-5}, weight decay of 0.05 excluding biases and RMSNorm, and dropout of 0.05) alongside a cosine-decay scheduler with a 10% warmup, training for up to 20 epochs with a batch size of 32 and a four-epoch early stopping patience based on validation QWK scores. Our experimental framework evaluates a *non-pretrained* baseline against several architectural and input variations: pooling strategy (last token versus average representation), training mode (causal versus bidirectional masking), and input strategy (*essay-only* versus a *full* context exceeding the 1024-token pre-training limit to assess length generalization). We add a non-pretrained model to assess the impact of pre-training for the different traits. Table 3 summarizes our variation space.

5. Results

As our models have a large variation space, we structure our analysis first by input strategy. Then, we segment training mode and pooling strategies across all five traits. Finally, we establish the highest-performing configurations and compare our findings with the previous literature.

Starting with essay-only input, we present in Table 4 the results of training with a bidirectional mask. As with previous research, there is no model that consistently outperforms all the others. The only case where dominance occurs is in the average pooling, between the constant and linear models. When comparing the model with the highest average performance against the non-pretrained baseline, we see that pre-training was advantageous. However, its importance changes according to the model — importantly, coherence (C4) in the average pooling was better assessed without pre-training. When comparing the pooling strategies, we can see that averaging the representations caused the models to have a higher variance — their average performances ranges from .24 to .41, while using the last token ranges from .31 to .37.

In Table 5, we present the results when the training regime was causal. Here, there

Table 5. Comparison of model performance using Causal mode for essay-only input. In bold, the highest scoring model inside each trait and each pooling strategy

Mode: Causal												
model	C1	C2	C3	C4	C5	avg.	C1	C2	C3	C4	C5	avg.
Pool. Strategy:	*average*						*last token*					
causal	**.47**	.01	.26	.35	-.01	.21	**.57**	.19	.32	.37	.35	.36
constant	.41	.00	**.27**	**.47**	.08	.24	.55	.21	.22	.41	.44	.36
cosine	.42	.00	.24	.43	.02	.22	.46	**.30**	.29	.47	.40	.38
linear	.37	**.02**	.20	.33	.08	.20	.47	.23	**.42**	.38	.38	.37
shift	**.47**	-.25	.13	.38	**.28**	.20	.46	.18	.12	**.48**	**.49**	.34
non-pretrained	.41	.12	.24	.20	-.02	.19	.11	.13	.15	.45	.17	.20

Table 6. Lower and upper bound of the best hybrid model by trait

trait	lower	upper	QWK	model	classifier head	mode
C1	.54	.74	.64	constant	reduce mean	Bidirectional
C2	.13	.48	.30	cosine	last token	Causal
C3	.24	.56	.42	linear	last token	Causal
C4	.42	.69	.56	cosine	last token	Bidirectional
C5	.31	.63	.49	shift	last token	Causal

are no cases of dominance. We see that the models with average pooling did not benefit from pre-training. The last token pooling, on the other hand, did benefit from it. Here, both pooling strategies yielded models that are very similar on average: they present only .04 points of amplitude in both cases.

When analyzing based on the best performance in each trait, we see an interesting result: the average pooling appears as the best strategy in only one case (C1). The bidirectional mode was the best in two traits, C1 and C4, both previously identified as traits better evaluated by Encoders [Barbosa et al. 2025] — which are bidirectional. Lastly, the pure models (causal model with causal mode and masked model with bidirectional mode) are never the best performing. The best ones are, respectively: constant, cosine, linear, cosine, and shift — all of them are hybrid models. In Table 6, we present the best scoring configurations for each trait, together with their lower and upper bounds of performance according to the bootstrap method.

Next, we investigate the usage of the full context to grade the essays. This step is relevant because C2 and C3 require relating the essay to the prompt and the supporting text given in the context. Ideally, adding the context should not hurt the performance in the other three traits, but it has been noted that this is not the case [Barbosa et al. 2025]. We present the best performances in the first division of Table 7, where we refer to our models as FlexQwen.

In general, we see that the performance either remained the same (C5) or decreased when adding the full context. One explanation for this observation is that the model did not receive input sequences with length longer than 1024 during training. Thus we conducted an experiment whre the model was pre-trained for one more epoch with

Table 7. Performance of best performing models: ours and previously benchmarked.

	C1	C2	C3	C4	C5
Best FlexQwen (essay-only)	.64	.30	.42	.56	.49
Best FlexQwen (full context)	.50	.24	.26	.46	.49
Best Encoder [Barbosa et al. 2025]	.68	.32	.29	.60	.63
Best Small LLM [Barbosa et al. 2025]	.66	.36	.37	.49	.59

sentences of length 2048 and then fine-tuned for C2 and C3, but the results were not significantly higher.

In the second division of Table 7, we present the best model performance in each trait as presented in the benchmark. Importantly, these performances come from the experiments where the models only had access to the essay text. The best Encoder are: a draw between Albertina and BERTimbau-large for C1, BERTimbau-large for C2, BERTuguês for C3, BERTimbau-large for C4, and BERTimbau-base for C5. While the best Decoders are, respectively: Tucano, Llama3, Llama3, Tucano, Phi4.

Three cases emerge when we compare our best essay-only models against the selected models from the benchmark. The first case occurs in C1 and C2: the SOTA is not that far from our model's performance (.06) and is within the confidence interval. The second case occurs in C3 and C4, where our model is better than at least one of the SOTA — when our model has better performance, the other model is within its confidence interval. The last case comprises C5, where the SOTA is the upper bound of our model. With this comparison, we see that our model, created with limited data, is comparable to models with many more parameters and trained for longer time on more data.

As a final comparison, we extend the baseline performances presented in the benchmark. There, the authors used a Linear Regressor and Random Forest over textual features calculated by NILC-Metrix [Leal et al. 2024]. With our work, it is possible to select the best non-pretrained model and present it as the baseline performance of the Transformer architecture. In addition, in the benchmark, the confidence interval of the feature-based models was not presented.[2] We calculate for our models and present the comparison in Table 8.

It is interesting to note that in the mechanical traits (C1 and C4), the hybrid models had the same performance as the feature-based, that use syntax- and grammar-related features. This further supports the hypothesis that Transformer-based models are approximating these features when fine-tuned on this dataset [Silveira et al. 2025a]. For the two main thematic traits (C2 and C3), our models performed worse than the feature-based ones, suggesting that the models were not capable of identifying which features would be important for these cases. Finally, for C5, our model is slightly better, probably because in this trait, some formulations are important — which are better identified through the words rather than proxies — but they also require some world knowledge, that is why it was not able to achieve a good performance. Given that, for the mechanical traits, the non-pretrained model achieves nearly .4 QWK and the pre-trained reach nearly .6, it is

[2]Their confidence intervals are presented at: `https://huggingface.co/datasets/kamel-usp/jbcs2025_experiments_report`

Table 8. Lower and upper bound of best non-pretrained hybrid model by trait and best performance from Feature-based models in the benchmark.

	Hybrid			Best Feature-based		
Trait	**Lower**	**Upper**	**QWK**	**Lower**	**Upper**	**QWK**
C1	.26	.53	.41	—	—	.41
C2	.00	.32	.16	—	—	.32
C3	.05	.44	.25	—	—	.35
C4	.35	.62	.48	—	—	.48
C5	-.01	.35	.17	—	—	.12

possible that better models can be created for these traits by using fewer Transformer blocks or by pre-training for less time.

6. Conclusion

In this work, we presented contributions in two different areas. The first, and main, area is the proposal of a family of Hybrid Transformer-based models trained with a relatively small amount of data. Then, through Automatic Essay Scoring, we validated that this model is competitive with much larger models, such as Albertina and Phi4, in evaluating five different textual dimensions — ranging from proper usage of Portuguese to argumentative quality. This suggests that it is possible to create better models for Portuguese without relying on increasing their scale.

The second area to which we contribute is Automatic Essay Scoring. Previous baseline performances were based on feature-based Linear Regressors and Random Forest. Our non-pretrained model fills the role of a Transformers-based baseline model, as it presents the performance of this architecture without pre-training. Previous research has shown that mechanical traits (C1 and C4) were better graded by Encoders. In our experiments, we showed that the bidirectional mode of training — typically used to pre-train Encoders — yielded the best performing models for these traits. Also, in accordance with previous research, thematic competencies were better evaluated by models trained with causal mask. One other contribution of our work was presenting that the best scoring model although following the trend of previous research, was not a pure masked model for mechanical traits nor a pure causal model for the thematic ones. This suggests that combining both pre-training regimes may lead to better models. Finally, the analysis of our non-pretrained model may suggest that for mechanical traits, it may be advantageous to design models with fewer Transformer blocks, making the model even more manageable.

Limitations

For our experiments, we employed the same dataset split and methods used in [Barbosa et al. 2025] so that we could compare against the previously evaluated models. One downside of this approach is that the bootstrap method with a relatively small dataset causes the confidence intervals to be large, which induces more false positives. In other words, some models may seem like equivalents without actually being so.

The capacity of our Hybrid models was not exhausted. It is possible to further improve its context window by presenting sentences of 2048 tokens or longer during

training. Finally, it is also possible to use it as a Decoder and prompt it to assess the essays. We focused on evaluating it as a traditional Encoder, as it reduces the number of experiments and makes the output simpler.

Ethics Statement

This research utilizes the *Corpus* Carolina v2.0 for pre-training and a previously established, anonymized dataset of mock ENEM essays for fine-tuning and evaluation. All data used in this study is publicly available and properly licensed for research purposes. We recognize that language models can inadvertently encode or amplify linguistic biases present in their training data, potentially disproportionately affecting marginalized dialects or non-standard varieties of Portuguese. While our hybrid model demonstrates competitive performance, we advocate that such automatic essay scoring systems should be seen as research objects. If employed as systems, they should be used as assistive tools for educators rather than autonomous grading authorities, ensuring human oversight in high-stakes educational assessments.

Disclosure of Generative AI Usage

During the preparation of this manuscript, the authors utilized generative AI to assist with English grammar revision and structural proofreading. AI tools were also used for finding bugs in experimental code. All core research, conceptualization, data analysis, and final content decisions were solely performed by the human authors.

Aknowledgements

This work was carried out at the Center for Artificial Intelligence (C4AI-USP), with support by the University of São Paulo, the São Paulo Research Foundation (FAPESP) (grant 2019/07665-4). Marcelo Finger was partly supported by the São Paulo Research Foundation (FAPESP) (grants 2023/00488-5 (SPIRA-BM), 2022/11254-2 (EMU)); and the National Council for Scientific and Technological Development (CNPq) (grant PQ1 302963/2022-7). Denis Deratani Mauá was partly supported by FAPESP (grant 2022/02937-9) and CNPq (grant 305136/2022-4). Igor Silveira was supported by FAPESP (grant 2026/15768-1). This study was supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) — Finance Code 001.

References

Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A. A., Bach, N., Bahree, A., Bakhtiari, A., Bao, J., Behl, H., Benhaim, A., Bilenko, M., Bjorck, J., Bubeck, S., , et al. (2024a). Phi-3 technical report: A highly capable language model locally on your phone.

Abdin, M., Aneja, J., Behl, H., Bubeck, S., Eldan, R., Gunasekar, S., Harrison, M., Hewett, R. J., Javaheripi, M., Kauffmann, P., et al. (2024b). Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Amorim, E. C. F. and Veloso, A. (2017). A multi-aspect analysis of automatic essay scoring for Brazilian Portuguese. In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 94–102. Association for Computational Linguistics.

Barbosa, A., Silveira, I. C., and Mauá, D. D. (2025). An empirical analysis of large language models for automated cross-prompt essay trait scoring in brazilian portuguese. *Journal of the Brazilian Computer Society*, 31(1):857–870.

Bazelato, B. S. and Amorim, E. C. F. (2013). A bayesian classifier to automatic correction of portuguese essays. In *Conferência Internacional sobre Informática na Educação (TISE)*, volume 18, pages 779–782.

Belcak, P., Heinrich, G., Diao, S., Fu, Y., Dong, X., Muralidharan, S., Lin, Y. C., and Molchanov, P. (2025). Small Language Models are the Future of Agentic AI.

Carmo, D., Piau, M., Campiotti, I., Nogueira, R., and Lotufo, R. (2020). Ptt5: Pretraining and validating the t5 model on brazilian portuguese data. *arXiv preprint arXiv:2008.09144*.

Carpi, M. d. M. and Finger, M. (2026). FlexQwen: Exploring Hybrid Objectives and Text Originality for Portuguese. In Souza, M., de-Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 1079–1084, Salvador, Brazil. Association for Computational Linguistics.

Chalegre, P. C., Machado, V. d. R., and Feltrim, V. D. (2026). Avaliação automática de redações do enem: Uma análise comparativa entre engenharia de características e transformers. In Souza, M., de Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 738–748, Salvador, Brazil. Association for Computational Linguistics.

Charpentier, L. G. G. and Samuel, D. (2023). Not all layers are equally as important: Every Layer Counts BERT. In Warstadt, A., Mueller, A., Choshen, L., Wilcox, E., Zhuang, C., Ciro, J., Mosquera, R., Paranjabe, B., Williams, A., Linzen, T., and Cotterell, R., editors, *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 238–252, Singapore. Association for Computational Linguistics.

Charpentier, L. G. G. and Samuel, D. (2024). GPT or BERT: Why not both? In Hu, M. Y., Mueller, A., Ross, C., Williams, A., Linzen, T., Zhuang, C., Choshen, L., Cotterell, R., Warstadt, A., and Wilcox, E. G., editors, *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 262–283, Miami, FL, USA. Association for Computational Linguistics.

Corrêa, N. K., Sen, A., Falk, S., and Fatimah, S. (2025a). Tucano: Advancing Neural Text Generation for Portuguese. *Patterns*.

Corrêa, N. K., Sen, A., Falk, S., and Fatimah, S. (2025b). Tucano: Advancing neural text generation for Portuguese. *Patterns*, 6(11):101325.

de Lima, T. B., Freitas, E., and Macario, V. (2024). Aesvoting: Automatic essay scoring with bert and voting classifiers. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 2*, pages 6–9.

de Sousa, R. F., Marinho, J. C., Neto, F. A., Anchiêta, R., and Moura, R. S. (2024). PiLN at PROPOR: A BERT-Based Strategy for Grading Narrative Essays. In *Proceedings of*

the 16th International Conference on Computational Processing of Portuguese-Vol. 2, pages 10–13.

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., and et al. (2025). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Doewes, A., Kurdhi, N. A., and Saxena, A. (2023). Evaluating quadratic weighted kappa as the standard performance metric for automated essay scoring. In Feng, M., KÃ¤ser, T., and Talukdar, P., editors, *Proceedings of the 16th International Conference on Educational Data Mining*, pages 103–113, Bengaluru, India. International Educational Data Mining Society.

Finger, M., de Sousa, M. C. P., Namiuti, C., do Monte, V. M., Costa, A. S., Serras, F. R., Sturzeneker, M. L., Carpi, M. d. M., Palma, M. F., and Lachi, G. A. (2025). Building Carolina: Metadata for Provenance and Typology in a Corpus of Contemporary Brazilian Portuguese. *Cadernos de Linguística*, 6(4):e812–e812.

Fonseca, E. R., Medeiros, I., Kamikawachi, D., and Bokan, A. (2018). Automatically grading brazilian student essays. In *Computational Processing of the Portuguese Language. PROPOR 2018.*, pages 170–179.

He, P., Liu, X., Gao, J., and Chen, W. (2020). DEBERTA: DECODING-ENHANCED BERT WITH DISENTANGLED ATTENTION. In *International Conference on Learning Representations*.

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the Knowledge in a Neural Network.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., and Liu, Q. (2020). TinyBERT: Distilling BERT for Natural Language Understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online. Association for Computational Linguistics.

Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., and Levy, O. (2020). SpanBERT: Improving Pre-training by Representing and Predicting Spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein (2024). Muon: An optimizer for hidden layers in neural networks | Keller Jordan blog. https://kellerjordan.github.io/posts/muon/.

Leal, S. E., Duran, M. S., Scarton, C. E., Hartmann, N. S., and Aluísio, S. M. (2024). Nilc-metrix: assessing the complexity of written and spoken language in brazilian portuguese. *Language Resources and Evaluation*, 58(1):73–110.

Lee, C., Jin, J.-g., Cho, Y., and Park, E. (2024). QEFT: Quantization for Efficient Fine-Tuning of LLMs. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Findings*

of the Association for Computational Linguistics: EMNLP 2024, pages 13823–13837, Miami, Florida, USA. Association for Computational Linguistics.

Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., and Carlini, N. (2022). Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445.

Leveling, J., Helmer, L., Stein, B. J., Wegener, D., Sheikh, Z., Fernandes, E., and Abdelwahab, H. (2024). Evaluation of document deduplication algorithms for large text corpora. In *International Conference on Machine Learning, Optimization, and Data Science*, pages 390–404. Springer.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Loshchilov, I. and Hutter, F. (2018). Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Maity, K., Chaulwar, A. T., Vala, V., and Guntur, R. S. (2024). NanoBERT: An Extremely Compact Language Model. In *Proceedings of the 7th Joint International Conference on Data Science & Management of Data (11th ACM IKDD CODS and 29th COMAD)*, CODS-COMAD '24, pages 342–349, New York, NY, USA. Association for Computing Machinery.

Marinho, J. C., Cordeiro, F., Anchiêta, R. T., and Moura, R. S. (2022). Automated essay scoring: An approach based on enem competencies. In *Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional*, pages 49–60. SBC.

Matos, G. G. d. and Feltrim, V. D. (2026). Avaliação automática de redações do enem: Um estudo empírico sobre representações linguísticas e contextuais. In Souza, M., de Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 488–497, Salvador, Brazil. Association for Computational Linguistics.

Menghani, G. (2023). Efficient Deep Learning: A Survey on Making Deep Learning Models Smaller, Faster, and Better. *ACM Computing Surveys*, 55(12):1–37.

Micheli, V., d'Hoffschmidt, M., and Fleuret, F. (2020). On the importance of pre-training data volume for compact language models. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7853–7858, Online. Association for Computational Linguistics.

OpenAI, :, Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., and et al. (2024). Openai o1 system card.

Piau, M., Lotufo, R., and Nogueira, R. (2024). ptt5-v2: A closer look at continued pretraining of t5 models for the portuguese language. In *Brazilian Conference on Intelligent Systems*, pages 324–338. Springer.

Pires, R., Abonizio, H., Almeida, T. S., and Nogueira, R. (2023). *Sabiá: Portuguese Large Language Models*, page 226–240. Springer Nature Switzerland.

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., et al. (2021). Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Ribeiro, E., Mamede, N., and Baptista, J. (2024). Exploring the automated scoring of narrative essays in brazilian portuguese using transformer models. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 2*, pages 14–17.

Rodrigues, J., Gomes, L., Silva, J., Branco, A., Santos, R., Cardoso, H. L., and Osório, T. (2023). Advancing neural encoding of portuguese with transformer albertina pt-*.

Rossman, L. N., Silveira, I. C., and Mauá, D. D. (2026). Evaluating automated scoring models on official ENEM essays. In Souza, M., de Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 161–171, Salvador, Brazil. Association for Computational Linguistics.

Samuel, D. (2024). Berts are generative in-context learners. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C., editors, *Advances in Neural Information Processing Systems*, volume 37, pages 2558–2589. Curran Associates, Inc.

Samuel, D., Kutuzov, A., Øvrelid, L., and Velldal, E. (2023). Trained on 100 million words and still in shape: BERT meets British National Corpus. In Vlachos, A. and Augenstein, I., editors, *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1954–1974, Dubrovnik, Croatia. Association for Computational Linguistics.

Schick, T. and Schütze, H. (2021). It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. In Erk, K. and Smith, N. A., editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Silveira, I. C., Barbosa, A., da Costa, D. S. L., and Mauá, D. D. (2025a). Investigating universal adversarial attacks against transformers-based automatic essay scoring systems. In Paes, A. and Verri, F. A. N., editors, *Intelligent Systems*, pages 169–183, Cham. Springer Nature Switzerland.

Silveira, I. C., Barbosa, A., and Mauá, D. D. (2024). A new benchmark for automatic essay scoring in Portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 228–237.

Silveira, I. C. and Mauá, D. D. (2026). Neuro-symbolic approaches for rubric-based automatic essay evaluation of ENEM essays. In Souza, M., de Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 790–799, Salvador, Brazil. Association for Computational Linguistics.

Silveira, I. C., Ribeiro, E., Mamede, N., and Baptista, J. (2025b). Aprendizado por transferência para correçao automática de redaçao. *Linguamática*, 17(2):99–116.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., and Liu, Y. (2024). RoFormer: Enhanced transformer with Rotary Position Embedding. *Neurocomputing*, 568:127063.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. (2025). Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yu, X., Guo, B., Luo, S., Wang, J., Ji, T., and Wu, Y. (2024). AntLM: Bridging Causal and Masked Language Models. https://arxiv.org/abs/2412.03275v1.

Zago, R. M. and Pedotti, L. A. d. S. (2024). Bertugues: A novel bert transformer model pre-trained for brazilian portuguese. *Semina: Ciências Exatas e Tecnológicas*, 45:e50630.

Zhang, Y., Warstadt, A., Li, X., and Bowman, S. R. (2021). When Do You Need Billions of Words of Pretraining Data? In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *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 1112–1125, Online. Association for Computational Linguistics.