Evaluating Automated Scoring Models on Official ENEM Essays

Laís Nuto Rossman and Igor Cataneo Silveira and Denis Deratani Mauá
Institute of Mathematics and Statistics, University of São Paulo, São Paulo, Brazil
laisnuto@gmail.com, {igorcs,ddm}@ime.usp.br

Abstract

Automated Essay Scoring systems can relieve teachers of this laborious task and allow students to practice more frequently due to faster feedback cycles. In Brazilian Portuguese, there is growing interest in automatic scoring systems for the standardized ENEM exam. However, the only available datasets consist of essays written as practice for the official exam. In the literature, to the best of our knowledge, there is no work that evaluates official ENEM essays using mock-exam datasets. This work fills that gap by presenting a new labeled dataset composed of 157 essays written for the official ENEM exam. The analysis shows that this dataset shares characteristics similar to existing datasets of mock exam essays. The results also indicate that, for small datasets such as this one, the use of LLMs pretrained on mock exams significantly improves the performance of automatic scorers for official ENEM essays, yielding an average gain of 0.27 points in the Quadratic Weighted Kappa metric compared to training solely on official data.

1 Introduction

Automated Essay Scoring (AES) was proposed as a method to unburden teachers from the labor-intensive task of scoring essays (Page, 1966). By making the scoring process faster, it also allows students to practice more frequently. Being able to practice more is especially important in contexts where students need to take a standardized exam that requires writing a long text — usually an essay.

In Brazil, the standardized exam Exame Nacional do Ensino Médio (ENEM) is the main entrance evaluation for universities. This exam is divided into two parts: answering four multiple-choice sections and writing an argumentative essay. Thus, writing a proper essay is crucial for securing a place in higher education.

The vast majority of public datasets for AES in Brazilian Portuguese consist of essays submitted to mock exams administered by websites that simulate the ENEM (Amorim and Veloso, 2017; Marinho et al., 2021; Silveira et al., 2024). Previous work (Silveira et al., 2024) has noted that experienced graders found that the prompts proposed by these websites do not exactly match the characteristics of the official exam. Moreover, test-takers of mock exams have different incentives than test-takers of the official ENEM exam, and essays are written under different (and unknown) conditions. Mock exams are also subject to selection biases, as the process of collecting, grading, and publishing essays on such websites is not disclosed. Thus, the validity and usefulness of existing AES systems as practice tools for the ENEM exams have not yet been established.

This validation gap is complicated by the lack of proper representative datasets of real ENEM essays. Although the prompts of the official exam are made public, only a few perfect-scoring essays are disclosed, making it impossible for third parties to train models using official data. Furthermore, it is also impossible to validate that the models trained on the mock essays are actually assigning scores that would be assigned in the official exam.

The present work fills this gap by presenting a dataset composed of 157 essays written for the official exam, along with their official scores. In order to create this dataset, three steps were taken: first, creating an online form so that students could voluntarily submit their data; second, transforming the essays from images to text; and finally, verifying the OCR output for each essay.

Using this dataset, four research questions are investigated: (1) How similar are the mock essays to the official ENEM essays? (2) Are models trained on mock test datasets able to grade official essays? (3) How do models trained only on the new official dataset perform? and (4) Does pre-training on simulated essays help the model adapt better to real ENEM essays?

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 161–171

To answer the first question, textual features are extracted from the proposed dataset using the NILC-Metrix tool (Leal et al., 2024), and a linear regressor is fitted for each competence. The experiment shows that the most important features in the dataset are the same as those reported for a mock-test dataset (Silveira et al., 2025). This suggests that the essays may share similar linguistic properties.

To answer the following two questions, three encoder models previously trained on mock-exam essays by Barbosa et al. (2025) are employed: BERTimbau-base (Souza et al., 2020), BERTuguês (Mazza Zago and Agnoletti dos Santos Pedotti, 2024), and mBERT (Devlin et al., 2019). The models are evaluated using the Quadratic Weighted Kappa (QWK) and F1 metrics. The results show that the models present a reasonable level of generalization. To address the second question, the models are fine-tuned on the official dataset and their performance is compared with the previous scenario. The results suggest that the models usually improve after fine-tuning, although the gain varies according to the model and the competence. The largest improvement is observed with BERTimbau in Competence 5 (0.316 QWK), while the largest decrease occurs when fine-tuning mBERT in Competence 1, reducing performance by 0.04 QWK.

Finally, an ablation study compares the performance of the fine-tuned models with versions trained exclusively on the official dataset in order to answer the fourth question. On average, pretrained models achieve a 0.27 higher QWK score than models trained without pretraining.

In summary, the contributions of this work are as follows.

- A dataset of official ENEM essays along with their scores is constructed and made available;[1]
- The dataset is shown to share similar characteristics with previous mock essay datasets;
- Pretrained models are shown to exhibit good generalization to official exams;
- Pretraining is shown to have a significant impact on model performance.

The remainder of the work is divided as follows: related work, creation of the dataset, methodology, results, discussion, and conclusion.

[1]Available at: HuggingFace.

2 Related Work

Previous datasets related to ENEM essays were created by scraping essays along with their grades from websites (Amorim and Veloso, 2017; Marinho et al., 2021; Silveira et al., 2024). These websites present a monthly prompt. Students, then, write an essay about this prompt and send it to the website, which then grades the essay and make it available online. These online essays are then scraped and either used with the grade assigned by the website (Amorim and Veloso, 2017; Marinho et al., 2021) or regraded by experts (Silveira et al., 2024). Such datasets provide a large number of samples, for instance approximately 1,840 essays in Amorim and Veloso (2017) and 4,570 essays in Essay-BR. However, this collection strategy introduces an important trade-off between scale and alignment with the official ENEM exam. According to Silveira et al. (2024), experts observed that the prompts proposed by these platforms often differ from those used in the official exam, which may influence linguistic properties such as text length, lexical richness, and grade distribution. Although these datasets represent valuable contributions, the ultimate goal of AES systems for the ENEM is not to evaluate mock essays, but to grade essays written under the conditions of the official exam. Our work addresses this limitation by presenting a dataset composed exclusively of essays written for the official ENEM exam, along with their official grades.

Other datasets for AES available in Brazilian Portuguese include the Narrative Dataset (Mello et al., 2024; Oliveira et al., 2025) and Diplomatrix (Cavalcanti et al., 2025). The first consists of narrative essays written by students from the 5th to the 9th grade and is not associated with a standardized exam. The second contains essays written for the Brazilian diplomat exam and presents a bias toward high grades, as it was composed of essays from candidates who were approved in the exam. A similar bias is observed in the dataset introduced in this work, since the essays were submitted by students who are currently enrolled in higher education.

Several models for AES were tested to score mock-test ENEM essays by Barbosa et al. (2025). These models range from feature-based to modern reasoning-based Large Language Models. The authors found that although there is no model that is always the best across all competencies, Encoder-based models were a good compromise between

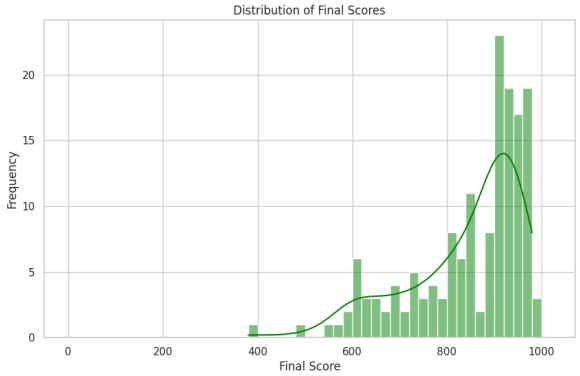

Figure 1: Distribution of Scores

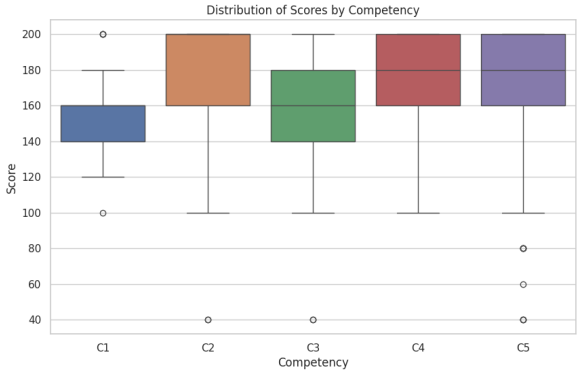

Figure 2: Distribution of Scores by Competence

size and performance. In the present work, the models made available by the authors are employed to conduct the experiments.

NILC-Metrix (Leal et al., 2024) is a tool for calculating textual complexity in Portuguese. It extracts different features, such as the percentage of each grammatical class in the text, complexity indices, and length related scores. These features were employed in previous work to verify the differences between Diplomatrix and Essay-br (Cavalcanti et al., 2025). In this study, NILC-Metrix is used to compare the linguistic properties of the official ENEM essay dataset with those of a mock-test dataset.

Silveira et al. (2025) employed linear regression based on features from NILC-Metrix. They noticed that the most important features for each competence were usually the same. In this study, the same analysis is replicated to verify whether the most predictive features for each competence remain consistent with those previously reported.

3 Official ENEM Essays Dataset

The ENEM essay is evaluated on a scale from 0 to 200 points in increments of 40 for each of the five official competencies: (C1) mastery of formal written Portuguese; (C2) comprehension and development of the proposed topic; (C3) organization of information and argumentative consistency; (C4) textual cohesion; (C5) elaboration of a socially responsible intervention proposal.

Each essay is scored independently by at least two certified raters, and the essay could be reviewed by a third one if the first two scores differ by more than 80 points in any competence or if the sum of the scores in the 5 competencies differs by more than 100 points. The final grade is the arithmetic mean of the two closest scores. Only this final averaged score is disclosed to the candidate.

One of the main challenges in automatic essay scoring for the ENEM is the lack of publicly available official essays. To address this limitation, a new dataset was built using official essays from the ENEM. The essays were collected through an online form distributed across university communities, preparatory courses, and personal networks. Participants voluntarily submitted digital copies of their official essays together with their scoring sheets and explicitly authorized the use of their data for academic research through the consent form. After the text digitization process, all personal information was removed, and the dataset retained only the essay text, the exam year, and the corresponding competence scores. The dataset complies with Brazilian data protection regulations (LGPD) and contains no personally identifiable information.

A total of 173 essays were collected from real participants covering eight ENEM editions (2016, 2018, 2019, 2020, 2021, 2022, 2023, and 2024). The handwritten texts were then extracted and transcribed using Large Language Models (LLMs) combined with prompt engineering techniques to partially automate the digitization and textual normalization process. Each text was manually reviewed to ensure fidelity to the original writing.

After cleaning and verifying the consistency of both texts and scores, incomplete or illegible essays were excluded, resulting in a final dataset of 157 complete samples. Each record includes the essay text, the year of the exam, the official scores for all five competencies, the essay prompt, and the supporting texts provided in the exam.

As the form was widely shared within the university community, many of the participants who submitted their essays had achieved relatively high ENEM scores. Consequently, the score distribu-

Metric	Official ENEM	Mock Exams
Characters per essay	2158.3	1653.1
Words per essay	335.4	263.8
Sentences per essay	11.9	9.7
Average word length	5.28	5.13
Type-token ratio	0.581	0.619
Hapax ratio	0.431	0.472
Commas per sentence	2.53	2.14
Connectives per 100 words	7.94	8.29
Sentence length (words)	29.0	29.4

Table 1: Comparison of linguistic characteristics between the official ENEM essays introduced in this work and the mock exam dataset. All metrics reported are the average of that metric.

tions differ from that of the overall ENEM scores, presenting a larger concentration of essays receiving scores above the national average, while lower-scoring essays are underrepresented (see Figures 1 and 2). At the same time, the score distribution also differs significantly from that of mock exams (Silveira et al., 2024), which mitigates some of the biases introduced by the data collection strategy.

To further contextualize these differences, we compared corpus-level linguistic characteristics between the official ENEM essays and the mock exam dataset used by Barbosa et al. (2025). Table 1 shows that official essays are generally longer, while both datasets exhibit similar sentence-level structures.

Lexical diversity is slightly higher in the mock dataset, which is consistent with its shorter essays. Other discourse-level indicators, such as connective usage and punctuation density, are broadly comparable. These findings suggest that although mock essays are somewhat shorter, the two datasets present similar structural characteristics.

4 Methodology

The methodology adopted in this work aimed to address four main questions: (a) How similar are the mock essays to the official ENEM essays? (b) Are models trained on mock test datasets able to grade official essays? (c) How do models trained only on the new official dataset perform? and (d) Does pre-training on simulated essays help the model adapt better to real ENEM essays?

To answer these questions, a textual complexity analysis was conducted comparing simulated and official essays, followed by training and evaluation experiments using different BERT-family encoders. In addition to these encoder experiments, an extra evaluation was performed using GPT-4o, as it has been previously tested on ENEM-like mock exams and, unlike the encoders, does not undergo task-specific fine-tuning.

Textual Complexity Analysis The textual analysis aimed to investigate whether the most important linguistic features that predict ENEM essay scores in non-official datasets are also the most important in the official dataset. To this end, 72 textual metrics were extracted from NILC-Metrix for 157 official ENEM essays, including syntactic and cohesion aspects such as word and sentence counts, part-of-speech ratios, and discourse connectives.

Independent linear regression models were then trained for each competence (C1–C5). All features were standardized using z-score normalization, allowing direct interpretation of coefficients as the expected change in score for a one-standard-deviation increase in a given metric.

Encoder Models Three base models from the BERT family were selected: BERTimbau, BERTuguês, and mBERT. Each model was evaluated under three scenarios: (1) zero-shot performance of models pretrained only on non-official essays; (2) performance of models exclusively fine-tuned on the dataset introduced in this work, without any prior training on essay grading; and (3) models pretrained on non-official essays and subsequently fine-tuned on the dataset introduced in this work.

To assess the models on the dataset, two metrics were employed. First, the traditional Quadratic Weighted Kappa (QWK), the standard metric used for AES (Doewes et al., 2023; Fonseca et al., 2018; Marinho et al., 2022a,b). This metric ranges from -1 to 1, indicating perfect disagreement in the negative case and perfect agreement in the positive case — zero indicates random agreement. QWK penalizes quadratically the distance between assigned labels and also uses the label distribution to determine an expected random agreement between them, which is used to weight the observed agreement. The second metric is the weighted F1, which is also commonly used in AES (Mello et al., 2024; de Sousa et al., 2024; Ribeiro et al., 2024; Barbosa et al., 2025). Complementary to the previous, the weighted F1 assigns high values when frequent classes are correctly identified.

The pretrained versions of the models are the ones made available in (Barbosa et al., 2025). In order to fine-tune the models, the dataset was di-

	mBERT						BERTuguês						BERTimbau						GPT4o					
	C1	C2	C3	C4	C5	avg.	C1	C2	C3	C4	C5	avg.	C1	C2	C3	C4	C5	avg.	C1	C2	C3	C4	C5	avg.
Mock Exams	.520	.220	.350	.500	.000	.318	.620	.330	.290	.540	.360	.428	.600	.360	.350	.550	.630	.498	.510	.200	.380	.510	.550	.430
Zero-shot	.501	.349	.447	.626	.030	.391	.382	.375	.290	.619	.271	.387	.378	.253	.387	.628	.069	.343	.327	.362	.365	.330	.268	.330
Pretrained FT	.456	.464	.519	.727	.000	.433	.467	.449	.286	.685	.489	.475	.393	.355	.453	.658	.385	.449	–	–	–	–	–	–
Exclusive FT	.000	.348	.000	.272	.359	.196	.073	.313	.150	.315	.020	.174	.000	-.023	.000	.452	.346	.155	–	–	–	–	–	–

Table 2: QWK across experiments, with the original performance for reference. Columns are grouped by model; within each model, values are reported per competence (C1–C5) followed by their mean.

vided into training and test subsets, following the distribution of essay years to reduce the impact of essay prompt on model performance. Essays from 2016, 2018, 2022, and 2023 formed the test set (43 essays), while essays from 2019, 2020, 2021, and 2024 constitute the training set (114 essays).

All models underwent the same fine-tuning process. A grid search was conducted to identify the best hyperparameter configurations, using the QWK as the optimization metric. We tested learning rates of 10^{-5}, batch sizes of 16 and 32, and numbers of epochs equal to 8, 12, and 16.

Because the dataset is relatively small, a 4-fold cross-validation scheme was adopted to mitigate overfitting and improve the robustness of model selection. In this setup, each fold corresponded to a different set of years in the training data, ensuring that essays from the same year never appeared simultaneously in training and validation. Hyperparameters were selected based on the average QWK across the four folds, providing a more stable estimate of performance than a single validation split.

Each official ENEM essay is evaluated by two graders in five competencies, receiving individual scores between 0 and 200 in increments of 40. The final score for each competence is the average of the two graders, resulting in values in increments of 20. This creates a structural mismatch: the pretrained models only output scores in multiples of 40, while human-granted scores can have intermediate values (such as 20, 60, etc) from averaging. To ensure compatible training labels and avoid mismatches with the model's output space, any score that was not a multiple of 40 was rounded to the nearest lower multiple of 40 before fine-tuning.

There is also some uncertainty in understanding the model's accuracy. Even if a final score is a multiple of 40, that does not mean the number is straightforward. For example, a final score of 80 could mean that both graders assigned 80 or that they were far apart (for example, 40 and 120), since it still averages to 80. This illustrates the challenge with ENEM scoring: matching on a 40-point scale does not always indicate that graders truly agreed, and since only the average is shared, it complicates matters further.

Furthermore, the QWK metric itself depends on how predictions and actual scores are compared on these incompatible scales. Different rounding strategies for actual scores generate different metrics (Doewes et al., 2023). For this reason, several evaluation protocols with alternative rounding rules were tested. However, since they all led to the same conclusions, the main analysis of this work focuses exclusively on the method that compares predictions and actual scores without rounding. The complete results for the 5 rounding methods can be found in the Appendix A. In other words, the rounding from 20-point scale to 40-point scale was only done for training, while in testing the comparison is between a 20-point scale (actual score) and a 40-point scale (model prediction).

GPT-4o A complementary experiment was conducted using GPT-4o to compare encoder-based models with a modern decoder model. The evaluation followed the interaction pattern presented in Barbosa et al. (2025). Three choices were necessary: selecting the decoder model, deciding whether to include an explicit prompt, and choosing a grading guideline. GPT-4o was selected because it showed strong performance in previous studies. The extended prompt was not used, since it mainly improves the results for C2 and C3 and is not comparable to the encoder setup. Among the available guideline options, the "Student" guideline provided the most stable results and was therefore adopted. Under this configuration, GPT-4o works in a zero-shot setting, generating scores based only on its pretrained abilities, without any fine-tuning.

Statistical Significance Tests Since the official ENEM dataset is relatively small, we performed statistical significance tests to verify whether the observed performance differences are robust. We applied bootstrap resampling on the test set to compute confidence intervals for the QWK metric. Differences are considered statistically significant when the confidence intervals do not overlap.

	mBERT						BERTuguês						BERTimbau					
	C1	C2	C3	C4	C5	avg	C1	C2	C3	C4	C5	avg	C1	C2	C3	C4	C5	avg
Zero-shot	.373	.571	.143	.334	.002	.285	.296	.565	.221	.369	.272	.345	.302	.517	.193	.396	.082	.298
Pretrained FT	.375	.630	.209	.354	.180	.350	.372	.625	.195	.376	.373	.388	.360	.591	.183	.391	.368	.379
Exclusive FT	.271	.608	.088	.260	.386	.323	.289	.595	.161	.280	.187	.302	.271	.502	.088	.308	.382	.310

Table 3: Weighted F1 experiments, including the mean across competencies for each model.

	mBERT	BERTuguês	BERTimbau	avg
ΔC1	-.045	.085	.015	.018
ΔC2	.115	.074	.102	.097
ΔC3	.072	-.004	.066	.044
ΔC4	.101	.066	.030	.065
ΔC5	-.030	.218	.316	.168
avg.	.042	.087	.105	.078

	mBERT	BERTuguês	BERTimbau	avg
ΔC1	.456	.394	.393	.414
ΔC2	.116	.136	.332	.194
ΔC3	.519	.136	.453	.369
ΔC4	.455	.370	.206	.343
ΔC5	-.359	.469	.039	.049
avg.	.237	.301	.284	.274

Table 4: Left: Variation in QWK comparing fine-tuned vs. zero-shot (negative means a drop after fine-tuning). Right: Variation in QWK comparing Pretrained FT vs. Exclusive FT (negative means pretraining was harmful).

5 Results

We present the results in two main sections: textual metrics analysis and performance comparison.

Textual Metrics When the linear regression model was trained using the essays in the dataset introduced in this work, the same linguistic variables were observed to dominate the top coefficients for every competence: (1) the proportion of nouns, (2) the number of content words, and counts of (3) verbs, (4) adjectives, and (5) adverbs. These features were the most influential across all five competencies, with the coefficient signs varying according to the writing dimension being assessed.

This pattern strongly mirrors the findings of Silveira et al. (2025), where similar NILC-Metrix features were identified as the main predictors in non-official essays. Such overlap suggests that the linguistic structure and stylistic features of the proposed dataset closely resemble those found in mock essays.

Performance Comparison First, Tables 2 and 3 present the results measured using QWK and weighted F1, respectively. In these tables, the first row, *Mock Exams*, reports the performance of the pretrained models in their original study (Barbosa et al., 2025). The second row, *Zero-shot*, shows the performance of these pretrained models when evaluated on the official ENEM essay dataset. The third row, *Pretrained FT*, reports the performance obtained after fine-tuning the pretrained models using official ENEM essays. Finally, the last row, *Exclusive FT*, presents the results of models trained exclusively on the official ENEM essays dataset, without prior exposure to mock-exam datasets.

It is important to note that the results reported in the *Mock Exams* row were computed under a different annotation setting. In the dataset used by Barbosa et al. (2025), each essay contains the individual scores from two graders, and model predictions are compared separately with each score, with the final metric obtained by averaging the two results. In the official ENEM dataset introduced in this work, however, only the final averaged score is available. Therefore, the values in the *Mock Exams* row should be interpreted only as a reference and are not directly comparable to the results obtained on the proposed dataset, particularly for metrics such as QWK and weighted F1 that depend on the label distribution.

To better understand how model performance changes across experimental conditions, a comparison between different experiments was devised. First, the QWK of each model is compared when moving from Zero-shot to Pretrained FT, showing whether there are gains from conducting fine-tuning. Then, the difference in performance between Pretrained FT and Exclusive FT is evaluated, indicating whether pretraining plays an important role. These comparisons are presented on the left and right sides of Table 4, respectively. Finally, to account for the small evaluation set, statistical significance tests were performed using bootstrap resampling on the QWK metric. The results are shown in Table 5.

	mBERT					BERTuguês					BERTimbau				
	C1	C2	C3	C4	C5	C1	C2	C3	C4	C5	C1	C2	C3	C4	C5
Zero-shot vs. Pretrained FT	No	No	No	No	Yes	No	No	No	No	No	No	No	No	No	No
Exclusive FT vs. Pretrained FT	Yes	No	Yes	Yes	Yes	No	No	No	No	Yes	Yes	Yes	Yes	No	No

Table 5: Statistical significance of the differences between experimental settings based on bootstrap confidence intervals for QWK. Green cells indicate statistically significant differences.

6 Discussion

The results show three main findings. First, the zero-shot models already reached a reasonable level of performance on the official ENEM essays. Even without any adaptation, the models pretrained on non-official essays were able to generalize to the new dataset. This suggests that the linguistic patterns present in simulated essays are similar enough to those in real ENEM texts to support direct transfer.

Second, when the zero-shot models are compared with the pretrained fine-tuned ones, consistent improvements are observed. Out of the 15 comparisons (five competencies across three models), fine-tuning improved performance in 12 cases, and the average QWK increased for all models and all competencies. This indicates that, even with a small amount of official data, the pretrained models were able to adapt and refine their predictions, performing significantly better than in the zero-shot setting.

The comparison between pretrained fine-tuning and exclusive fine-tuning is even more striking. In 14 out of 15 cases, the pretrained and then fine-tuned model outperformed the model trained only on the official dataset, often with much larger differences. This shows that, with such a small dataset, BERT-based models cannot learn the specific scoring patterns of ENEM essays from scratch. The pretraining provides essential information that helps the model deal with the nuances of the ENEM scoring system.

These patterns are also reflected in the statistical significance tests (Table 5). Although Pretrained FT generally achieves higher scores than Zero-shot, most of these differences are not statistically significant, indicating that the zero-shot models already achieve performance close to the fine-tuned ones. In contrast, several competencies show significant improvements when comparing Pretrained FT with Exclusive FT, suggesting that pretraining on mock essays provides consistent benefits rather than improvements due to random variation.

However, it is important to interpret the QWK results with caution. As noted in the dataset analysis, the official essays in the test set come from students who generally performed well on the exam. As a consequence, the scores show low variance, and many essays fall into only a few classes. QWK penalizes this type of distribution: if a model predicts the same class for most examples, the QWK will be close to zero, even if those predictions match the dominant real score. This can be seen in the confusion matrices for the exclusive fine-tuning experiments. For example, in Figure 3, which shows the confusion matrix for mBERT on competence C1, the exclusive fine-tuning model predicts mostly 160 points, which drives the QWK to zero even though 18 of the true scores are also 160. A similar pattern appears for BERTimbau in competence C2, as shown in the confusion matrix in Figure 4. In the exclusive fine-tuning setting, the model collapses almost all predictions to the score 200. This leads to a negative QWK, which would typically suggest performance worse than chance. However, the confusion matrix shows a more nuanced picture: although the model predicts 200 for every essay, 27 essays in the real data also have this score. So, the model is capturing the dominant pattern of an unbalanced dataset rather than failing entirely. The negative QWK arises not because the predictions are unreasonable, but because the metric strongly penalizes models that assign the same score to most examples, even when that score matches the majority of true labels. This highlights how, in this dataset with low score variance, QWK may underestimate the practical quality of the predictions.

For this reason, the weighted F1-score was also examined. Although QWK and F1 cannot be directly compared, the exclusive fine-tuning models score noticeably better under F1 because the weighted F1 metric rewards models that correctly identify the majority class in an imbalanced dataset. Since these models tend to predict the most frequent score most of the time, they achieve a reasonable F1 by correctly classifying many majority-class instances. Even so, the pretrained and fine-

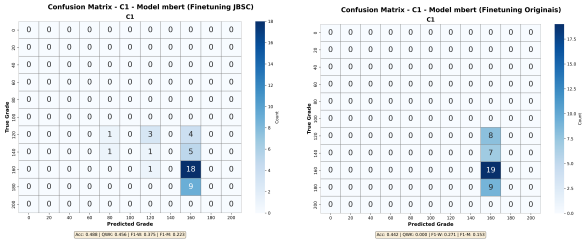

Figure 3: mBERT — Confusion matrices for C1 under (left) Pretrained Fine-tuning and (right) Exclusive Fine-tuning.

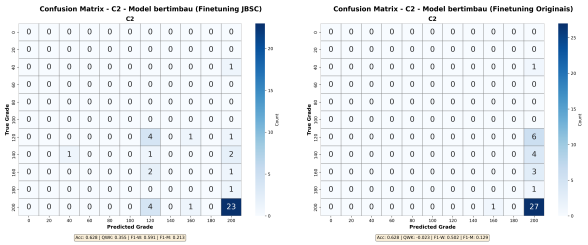

Figure 4: BERTimbau — Confusion matrices for C2 under (left) Pretrained Fine-tuning and (right) Exclusive Fine-tuning.

tuned models still outperform the exclusive fine-tuning ones, although by a smaller margin. This indicates that the pretrained models not only learned the dominant class but also captured more subtle distinctions in the dataset.

In addition to the encoder-based models, the experiment with GPT-4o on the test set revealed that, even though it is much larger, with an estimated hundreds of billions of parameters compared to roughly 110 million in the BERT models, it did not outperform the encoders. In several competencies, its performance was similar to or below that of the zero-shot models, and it was consistently behind the pretrained and fine-tuned versions. Since GPT-4o is more expensive to run and requires more computational resources, and BERT models are free and can be executed locally, using GPT-4o for automated scoring is simply not cost-effective in this context. Therefore, for scoring ENEM essays, smaller and well-trained encoder models offer a better cost-benefit trade-off than a much larger model like GPT-4o.

The textual analysis of the datasets also supports these findings. Using NILC-Metrix, it was observed that the main linguistic and structural characteristics of the non-official essays are very similar to those of the official ENEM essays in this corpus. Together with the encoder model results, this strengthens the argument that non-official essays are a suitable and useful resource for training automatic scoring models.

Overall, these results show that pre-training on non-official ENEM-like essays is valuable. Since official ENEM essays are difficult to obtain due to lack of public access, simulated essays are much easier to collect in large quantities. Our experiments demonstrate that models pretrained on such data can be successfully adapted to real ENEM scoring, and that this combination leads to better performance than training exclusively on the small official dataset.

7 Conclusion

In this work, a new dataset of 157 official ENEM essays is introduced and used to analyze how well current AES models trained on mock exam data perform on real ENEM essays. The first step consists of comparing the linguistic characteristics of this dataset with those of previous mock datasets using NILC-Metrix. The results show that the most important textual features are very similar in both cases, suggesting that simulated essays can approximate the structure and style of official ENEM writing.

Three encoder-based models (mBERT, BERTuguês, and BERTimbau) are then evaluated under three scenarios: zero-shot, pretrained fine-tuning, and exclusive fine-tuning on official ENEM essays. The zero-shot results indicate that models trained only on mock essays already generalize reasonably well to official essays. When the models are fine-tuned using the available official data, performance improves in most competencies and across all models.

Finally, the comparison between pretrained fine-tuning and exclusive fine-tuning showed the strongest difference. Models that were first trained on mock essays and only then fine-tuned on the official dataset performed much better than models trained only on the official essays. On average, the gain was about 0.27 in QWK, which is a large improvement given the small size of the dataset. This result, combined with the textual analysis, indicates that mock test essays are a useful resource for training ENEM scoring models.

In summary, these findings suggest a practical approach for building AES systems for the ENEM: use large collections of mock essays for pre-training and then adapt the models using a smaller set of official essays.

Limitations

The limitations of this work can be summarized in two main points. First, the dataset is relatively small, since collecting real ENEM essays is difficult. In addition, it is biased toward higher grades, as most of the participants who submitted essays are currently enrolled in higher education. Second, the experiments were restricted to encoder-based models. Future work could extend this analysis to decoder models, which are typically larger and may require more data to train effectively.

Acknowledgments

This work received generous financial support from the São Paulo Research Agency (FAPESP) grant no. 2022/02937-9, CNPq grant no. 305136/2022-4 and Finance Code 001.

References

Evelin Amorim and Adriano Veloso. 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.

André Barbosa, Igor Cataneo Silveira, and Denis Deratani Mauá. 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):858–871.

Rodrigo Cavalcanti, Gabriela Casini, Gabriel Assis, Livy Real, Daniela Vianna, Paulo Mann, and Aline Paes. 2025. Diplomatrix-br: Um corpus paralelo de redaçoes de autoria humana e de llms no concurso de diplomacia brasileira. In *Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, pages 192–205. SBC.

Rogério F de Sousa, Jeziel C Marinho, Francisco AR Neto, Rafael Anchiêta, and Raimundo S Moura. 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Afrizal Doewes, Nughthoh Arfawi Kurdhi, and Akrati Saxena. 2023. Evaluating quadratic weighted kappa as the standard performance metric for automated essay scoring. In *Proceedings of the 16th International Conference on Educational Data Mining*, pages 103–113, Bengaluru, India. International Educational Data Mining Society.

Erick Rocha Fonseca, Ivo Medeiros, Dayse Kamikawachi, and Alessandro Bokan. 2018. Automatically grading brazilian student essays. In *Proceedings of International Conference on Computational Processing of the Portuguese Language*, pages 170–179.

Sidney Evaldo Leal, Magali Sanches Duran, Carolina Evaristo Scarton, Nathan Siegle Hartmann, and Sandra Maria Aluísio. 2024. NILC-Metrix: Assessing the Complexity of Written and Spoken Language in Brazilian Portuguese. *Language Resources and Evaluation*, 58(1):73–110.

Jeziel Marinho, Rafael Anchiêta, and Raimundo Moura. 2021. Essay-br: a brazilian corpus of essays. In *Anais do III Dataset Showcase Workshop*, pages 53–64. Sociedade Brasileira de Computação.

Jeziel Marinho, Rafael Anchiêta, and Raimundo Moura. 2022a. Essay-br: a brazilian corpus to automatic essay scoring task. *Journal of Information and Data Management*, 13(1):65–76.

Jeziel Marinho, Fábio Cordeiro, Rafael Anchiêta, and Raimundo Moura. 2022b. Automated essay scoring: An approach based on enem competencies. In *Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional*, pages 49–60.

Ricardo Mazza Zago and Luciane Agnoletti dos Santos Pedotti. 2024. Bertugues: A novel bert transformer model pre-trained for brazilian portuguese. *Semina: Ciências Exatas e Tecnológicas*, 45:e50630.

Rafael Ferreira Mello, Hilário Oliveira, Moésio Wenceslau, Hyan Batista, Thiago Cordeiro, Ig Ibert Bittencourt, and Seiji Isotanif. 2024. PROPOR'24 Competition on Automatic Essay Scoring of Portuguese Narrative Essays. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 2*, pages 1–5.

Hilário Oliveira, Rafael Ferreira Mello, Péricles Miranda, Hyan Batista, Moésio Wenceslau Silva da Filho, Thiago Cordeiro, Ig Ibert Bittencourt, and Seiji Isotani. 2025. A benchmark dataset of narrative student essays with multi-competency grades for automatic essay scoring in brazilian portuguese. *Data in Brief*, 60:111526.

Ellis B. Page. 1966. The imminence of... grading essays by computer. *The Phi Delta Kappan*, pages 238–243.

Eugénio Ribeiro, Nuno Mamede, and Jorge Baptista. 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.

Igor Cataneo Silveira, André Barbosa, Daniel Silva Lopes da Costa, and Denis Deratani Mauá. 2025. Investigating Universal Adversarial Attacks Against Transformers-Based Automatic Essay Scoring Systems. In *Intelligent Systems*, pages 169–183, Cham. Springer Nature Switzerland.

Igor Cataneo Silveira, André Barbosa, and Denis Deratani Mauá. 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.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: pretrained bert models for brazilian portuguese. In *Brazilian conference on intelligent systems*, pages 403–417. Springer.

A Appendix: Influence of Rounding Strategies on QWK

In this appendix, we investigate the influence that different rounding strategies have on the values of Quadratic Weighted Kappa (QWK), as discussed in Section 4.

We devised four strategies to manage real scores on a 20-point scale and model predictions on a 40-point scale. The first strategy is called *no changes*, where we keep the real scores on the 20-point scale. Consequently, the performance of the models will be underestimated, as they cannot output on this fine-grained scale and will be penalized for predicting the wrong class. This is the strategy used to report all results in Section 5.

The remaining strategies change the 20-point scale to match the other. The second strategy is called *duplicate bounds*, where every real score is turned into two: one rounded up to the nearest multiple of 40 and other rounded down to the nearest multiple of 40. Note that for scores that are already multiples of 40, two new scores are created, both of which are equal to the original. This is done to mitigate the change in the distribution of labels. Notably, in this strategy, when the two different scores are created, if the model gets one of them right, it necessarily gets the other wrong by one class.

The other two strategies simply round the 20-point score to either its *floor* or *ceiling* on the 40-point scale. Note that these two strategies agglomerate scores, changing the distribution and making it more concentrated, which may be penalized by QWK, as QWK incentivizes identifying less frequent classes.

We present in Table 6 the performance of the Pretrained Fine-tuned models according to each strategy, along with their amplitude — the difference between the largest and smallest values of the strategies. We can see that the largest value of each column comes either from *no changes* or *floor*. The first strategy yielding high numbers may come from the fact that the grades are more distributed and that the penalty of QWK takes into account the number of classes. The reason why *floor* sometimes yields high results is not obvious, it might be due to the final distribution of real scores and predictions.

Additionally, we can see that the strategies impact the models differently — for the same competence, the amplitude is different for each model. Except for mBERT C5, the lowest amplitude was 0.3 and the largest was 0.131, which is a significant difference and directly impacts the decision to accept or reject a model. Finally, we can see that if we were interested in comparing which model is better, the number ordering of the models could be impacted by the strategy — for example, BERTuguês is better than mBERT in C1 in two strategies and worse in the other two.

Summary Across experiments, *no changes* and *floor* tend to yield the highest QWK. The relative ordering of models can change according to the strategy used, but for our research questions, the Pretrained Finetuned models performed better than their other variants, independent of the strategy used.

	mBERT					BERTuguês					BERTimbau				
	C1	C2	C3	C4	C5	C1	C2	C3	C4	C5	C1	C2	C3	C4	C5
Mock Exams	.520	.220	.350	.500	.000	.620	.330	.290	.540	.360	.600	.360	.350	.550	.630
no changes	.456	.464	.519	.727	.000	.467	.449	.286	.685	.489	.393	.355	.453	.658	.385
duplicate	.386	.456	.475	.687	.000	.384	.442	.263	.642	.471	.331	.350	.411	.618	.371
floor	.448	.473	.458	.686	.000	.501	.456	.284	.625	.491	.410	.369	.402	.639	.411
ceiling	.343	.438	.488	.688	.000	.308	.426	.246	.660	.450	.279	.330	.420	.596	.329
amplitude	.113	.035	.061	.041	.000	.193	.030	.040	.060	.041	.131	.039	.051	.062	.082

Table 6: QWK by competency for the **pretrained finetuned** models under different evaluation protocols. Mock Exams stands for the performance of the pretrained model in their original paper.