Neuro-symbolic Approaches for Rubric-Based Automatic Evaluation of ENEM Essays

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

Abstract

Trait-specific automated scoring of essays written for the standardized Brazilian National Entrance Exam (ENEM) has received significant attention in recent years. The task is both important in a classroom setting, to provide timely and personalized learning feedback, and in the official exam, to make the scoring process more scalable and consistent. The state-of-the-art systems approach the task as a purely statistical predictive task, ignoring the knowledge provided to human graders and test takers in the form of rubrics and guidelines. Aiming to produce more interpretable and informative formative feedback in this work, we leverage the official ENEM Grader's handbook and develop two neuro-symbolic approaches to trait-specific essay scoring. The first approach uses a Large Language Model (GPT4o) to write an evaluative explanation of the essay score according to the subcriteria described in the guidelines; the explanation is then fed into a statistical model to effectively predict the score; the good performance of the scoring validates the quality of the explanations. The second approach formalizes the Guideline grading rubrics as logical rules that derive the essay score as a function of subcriteria, mimicking the recommended human grader's scoring approach. In order to provide weak supervision in training and to evaluate the quality of the model, we build a dataset of 63 essays annotated with their subcriteria by two expert human graders. Our empirical results suggest that both approaches perform on par with purely statistical methods while providing more helpful and fine-grained feedback.

1 Introduction

Automated Essay Evaluation (AEE) seeks to ease the burden on teachers and scale up personalized feedback for students (Page, 1966; Shermis and Burstein, 2013). Most existing AEE systems are developed for and evaluated purely by their ability to produce essay scores that are aligned with human-assigned scores.

In Brazil, the *Exame Nacional do Ensino Médio* (ENEM) is a standardized exam used by the majority of higher education institutions as part of the admission process. The exam is divided into two parts: the first evaluates subject-matter knowledge through a multiple-choice questionnaire, and the second evaluates critical thinking and writing skills by means of an argumentative essay. Thus, mastery of essay writing is crucial for students who desire to enter higher education. Consequently, being able to simulate its evaluation is of major importance in High School. Being a high-stakes standardized exam implies that the ENEM essay exam has well-defined grading criteria and important societal and economic importance.

It is not surprising that the majority of the literature on AEE for Brazilian Portuguese consists of training machine learning models on essays submitted for mock ENEM exams hosted on online web portals (Amorim and Veloso, 2017; Marinho et al., 2022a; Silveira et al., 2024). Those works cast essay scoring as a simple prediction task. More recently, Barbosa et al. (2025) have investigated the use of instruction-guided LLM-based approaches that utilize additional materials such as official exam rubrics. Still, their end task consisted in providing a trait-specific score for each essay.

In this work, we leverage the released official ENEM Grader Guidelines[1] to provide more fine-grained feedback in the form of explanations and rubric-based scoring. More precisely, the rubrics are stated in the form of high-level logical rules over sub-score criteria for each trait. We develop two neuro-symbolic approaches to score the essays that make use of the official rubrics.

The first approach builds upon the work by Bar-

[1]Available at: https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem/outros-documentos.

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 790–799

bosa et al. (2025) and instructs a generative LLM model to justify (explain) a given essay score according to the subcriteria. Those explanations are then used to train a BERTimbau-base model in a supervised fashion to assign a trait-specific score based on the explanation instead of the essay text. We show that the resulting model performs comparably to models learned directly from essay texts, providing evidence that LLM's explanations are indeed accurate.

The second approach formalizes the official rubrics as logical inferences of a knowledge base whose propositions' truth-values are inferred by statistical models. The models are trained through gradient-based learning using the Semantic Loss (Xu et al., 2018), which evaluates the quality of probabilistic classifications under logical consistency constraints. In principle, that loss allows us to train the entire models end-to-end without any annotation of the subcriteria. To escape poor local optima and initialization issues, we develop a small dataset of subcriteria-labeled essays. The dataset contains 63 essays, submitted to mock ENEM exams, which were evaluated by two professional annotators with respect to a total of 71 subcriteria.

Our experiments show that either approach is capable of scoring the essays with a quality similar to that of the state-of-the-art predictive models, while providing more helpful feedback — either a text-based explanation in terms of subcriteria, or predictions for the subcriteria values.

To summarize, our contributions are: we present two novel neuro-symbolic approaches for AEE that provide sub-score feedback, which is arguably more helpful for student learning; we make available a dataset of essays annotated with their subcriteria by expert human graders.[2]

The rest of the work is organized as follows: related work (Section 2), description of the subcriteria annotated essay dataset (Section 3), the development of our neurosymbolic approaches (Section 4), presentation and discussion of empirical analyzes (Section 5), additional investigation on the second approach (Section 6), and conclusion (Section 7).

2 Related Work

There are many AEE datasets for Portuguese, each with different peculiarities and applications. For European Portuguese, Ribeiro et al. (2024a) created a dataset to identify text readability — that is, how proficient the reader must be to understand the essay — according to the Common European Framework of Reference for Languages. In Brazilian Portuguese, there are generally three different groups that can be used for essay scoring: narrative essays, diplomate essays, and ENEM essays.

The narrative dataset (Mello et al., 2024; Oliveira et al., 2025) is composed of essays written by middle school students in Brazil. The essays were written for two different prompts and evaluated by two professionals according to four distinct traits — only the average of the graders is presented in the data. Differing from the previous dataset, Diplomatrix (Cavalcanti et al., 2025) is composed of essays written for the diplomatic entrance exam, a graduate level exam. They were scored by professionals in a holistic fashion — that is, not divided by traits.

Finally, datasets of (mock) ENEM essays started with the work of Amorim and Veloso (2017). The authors scraped essays along with their trait-based scores from websites that simulate the exam. Notably, the scores at that time were not on the same scale as the original exam. Afterward, the Essay-br dataset (Marinho et al., 2021, 2022a) followed the same strategy and scraped two websites that simulate the exam. The essays were grouped together, and in cases where the scores were not on the proper scale, they were adjusted according to a heuristic. Next, AES-ENEM (Silveira et al., 2024) scraped the same websites as Essay-br but showed that the websites had different score distributions and thus divided them into Source A and Source B. Source A was then scored by two independent professionals, and it was shown that they had better agreement between themselves than with the scores available on the website. In this work, we hired experts to annotate essays from Source B not only with the scores but also according to each subcriteria evaluated in each trait.

Along with the datasets, several models have already been employed for scoring. These models can be divided into interpretable feature-based models, black-box Transformer-based models, and hybrid models.

The first group has been widely explored in the context of Brazilian Portuguese. Bazelato and Amorim (2013) developed a model based on the probability of a word occurring in a given score. Amorim and Veloso (2017) used common, more complex features employed in English systems. Marinho et al. (2022b) defined a different group

[2]Available at: https://huggingface.co/collections/igorcs/enem-subcriteria-dataset

of features that are expected to be useful for each ENEM trait. In the narrative essay competition (Mello et al., 2024), one of the baselines was a model based on the TF-IDF representation of the essays. Finally, Silveira et al. (2025a) leveraged the NILC-Metrix (Leal et al., 2024) to calculate several features in different levels of discourse. The key strength of these models is that it is possible to explain which elements influence the scoring: the presence of certain words, the value of a feature being too low, to name a few possible explanations.

Transformer-based models, on the other hand, usually have better performance than previous models. However, their improvement comes at the cost of losing explainability. For the narrative shared task (Mello et al., 2024), models competed to achieve the best performance in each trait. Each trait was evaluated according to two metrics. The winner in at least one metric was always a Transformer-based model: a baseline BERTimbau model (Souza et al., 2020), a model that predicted all the traits together (de Sousa et al., 2024), and one that decided which (Transformer-based) model to use depending on the trait (Ribeiro et al., 2024b). For the ENEM, Barbosa et al. (2025) compared feature-based models against Transformer-based models, and the latter had better performance.

Lastly, hybrid models attempt to combine the explainability of one type of model with the high performance of another. Liu et al. (2019) employed a two-stage approach to scoring: a BERT model learns to predict the essay score, and this score is used along with other features as input to a traditional, interpretable method. This strategy was also used by Silveira et al. (2025b) applied to narrative essays dataset. Barbosa et al. (2025) used the ENEM Graders Guideline as input for conversational models, such as ChatGPT, Maritaca and DeepSeek, requesting the models to relate the essay to the characteristics described in the guideline and then, based on these relationships, score the essay. Through this, there is a weak relationship between the score and the reasons for that output. We validate their approach by comparing the performance of a system that uses the essay as input against one that uses the identified relationships. Moreover, our approach based on logical restrictions has a strong relationship between the assigned grade and its rationale.

Little has been studied regarding automated feedback for essays in Portuguese. As far as we know, no work has been done on applications other than the ENEM. Anchiêta et al. (2025) investigated querying conversational models to generate feedback that suggests what to change in the essay in order to improve the score in each trait. Our feedback differs from theirs, as ours is an explanation for the grade rather than a suggestion on how to write better.

	AES-ENEM	SUB
C1	.57	.41
C2	.54	.30
C3	.59	.29
C4	.45	.18
C5	.64	.60
# essays	380	63

Table 1: QWK comparison between graders in AES-ENEM and in our subcriteria dataset (at trait level).

3 Subcriteria AES Dataset

One prominent characteristic of the writing part of the ENEM is that it has two sets of guidelines: one set consists of a document for the student, stating what is expected in the essay, broadly presenting each of the five traits and the meaning of each score in each trait. This document is about fifty pages long. The second set of guidelines available is the Graders Guideline. Here, each trait is explained in a different document — the smallest is 42 pages long, and the largest is 78 pages long. In these documents, each trait is decomposed into many disjoint characteristics — or subcriteria — and each score is defined in terms of which subcriteria must be satisfied to be assigned that grade.

In most cases, the subcriteria can be organized into dimensions. For instance, for trait C1, there are 9 subcriteria: (non-existing syntax, deficitary syntax, many mistakes, regular syntax, some mistakes, good syntax, few mistakes, excellent syntax, and maximum 2 mistakes). These subcriteria can be organized into two dimensions: syntax (5 subcriteria) and mistakes (4 subcriteria), where one and only one of the subcriteria must be true within each dimension. The final score for C1 is a logical consequence of which subcriteria are true. The same principles can be applied to all the other traits. Importantly, given a set of true subcriteria, only one score is deduced; however, one score can be assigned due to different assignments of subcriteria. As an example, consider the case for this same trait,

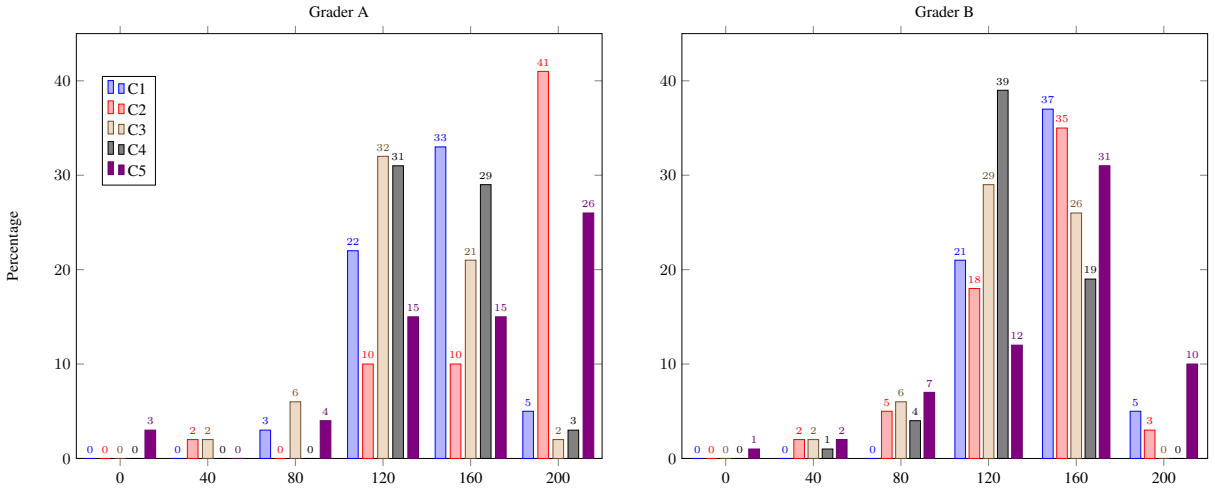

Figure 1: Per-trait score distributions for both graders.

where, for the sake of simplicity, each subcriterion in each dimension corresponds to one score, and the final score is the minimum score of each dimension. Then, it is easy to see that the same final score can be reached by different assignments of subcriteria, but given an assignment, only one final score is possible.

We extracted the subcriteria from all traits and hired two professionals who are experienced with official ENEM grading. We asked them to identify which subcriteria are true according to each dimension. In the process, they could see which score would be assigned by that classification. In total, they had to classify each essay according to the following dimensions (between brackets, the number of subcriteria within that dimension):

C1: syntax (5) and mistakes (4);

C2: theme (3), typology (5), conclusion (2), repertoire (4), pertinence (2), and usage (2);

C3: direction (2), project (4), development (4), and contradiction (2);

C4: cohesiveness (6), repetitions (5), inadequacies (5), and monoblock (2);

C5: human rights (2), adherence (4), elements (6), and conditional (2);

Compared to the traditional scoring task, this fine-grained annotation is even more laborious and time-consuming. As a consequence, only a few essays were annotated by each grader — the same essay was graded independently by both. We randomly selected 9 essays from 7 different prompts from Source B of AES-ENEM, which are also potentially present in Essay-br, summing up to 63 essays. By doing this, we can safely use this dataset without data leakage when evaluating models in Source A.

Our dataset allows for comparing agreement at the dimension and trait levels. In Table 1, we compare the trait level agreement in our dataset with AES-ENEM. The agreement was calculated using the standard Quadratic Weighted Kappa (QWK), which is widely used for scoring agreement (Doewes et al., 2023).

Although the agreement is always lower in our dataset, it might be due to several reasons: (i) the number of instances, as AES-ENEM is about six times larger; (ii) as the granularity of a task increases, the agreement tends to decrease (Ji et al., 2025); (iii) the graders might not actually use these subcriteria when grading an essay, thus yielding lower performance; (iv) the essays from Source B tend to have less grade diversity, hindering QWK.

In Figure 1, we present the grade distribution for each grader. We can see that for grades 0 and 40, their distributions are similar. For C1, they are very similar, while for C2, there is a significant difference between them. Overall, Grader A seems to be more lenient, assigning higher grades, while Grader B has grades that are more distributed over the grading scale. From this figure, we infer that, except for C2, the low QWK is not due to different distributions, but rather that having only one essay difference in each score impacts significantly, as one essay is very significant in a dataset of 63 essays. This impacts the expected and observed agreement used to calculate QWK, so we conclude

	Grade		Sub				
	Train	Test	D	U	Z	Train	Test
C1	.51	.29	2	0	0	.45	.38
C2	.27	.31	6	1	3	.24	.28
C3	.27	.29	4	2	0	.28	.58
C4	.16	.21	4	1	0	.08	.12
C5	.57	.64	4	1	2	.73	.63
#essays	36	27	–	–	–	36	27

Table 2: On the left, QWK comparison between graders in the trait level. On the right, D denotes the number of dimensions, U the number of undefined dimensions, Z the number of dimensions with random agreement, and then and average between all the remaining dimensions in each split.

that the metric would be higher with more data.

Finally, we split our dataset according to a different pattern than AES-ENEM and Essay-br. The latter randomly split the data into training, validation, and test sets. The former divided the three sets according to prompts, such that essays written for the same prompt are in the same set. In this work, we follow the same strategy of having no overlap between prompts in different sets. Our approach differs in that, instead of having one of the three sets, we have only two — train and test — and that there is a set for each trait instead of a global set. In other words, an essay might be in the training set for one trait but in the test set for another trait. We chose this approach because we have few essays, and by doing this, we can try to keep the agreement balanced between the two sets. The agreement between graders in each set is presented in the left part of Table 2.

One additional piece of information we calculate is the average agreement at the subcriteria level, which is presented in the right part of Table 2. Here, we face two problems due to the lack of diversity in our data: if all instances have the same label, then QWK is undefined — the number of dimensions with undefined QWK is presented in the U column. Sometimes, the correlation was zero due to the high prevalence of a class and one instance of a different class assigned by one of the graders — the number of such dimensions is presented in the Z column. The D column presents the total number of dimensions. Finally, the QWK agreement shown in the Train and Test columns is the average of the dimensions that are not in U and Z.

From the previous table, we highlight three points. First, four traits had at least one dimension where all instances were judged by both graders as having the same label, thus making it impossible to be learned. Second, in C2 and C5, the graders had zero agreement in half of the dimensions. Finally, only C1, the simplest of them, did not have any of the previous problems, and the average of the subcriteria agreements can be higher or lower than the agreement at the trait level.

4 Methodology

In this work, we are interested in leveraging the subcriteria that define the grading guidelines for the ENEM. By leveraging them, we can provide both the score and the explanation for the score. Thus, we pose the research question: how well do these models perform compared to the state-of-the-art scoring models?

To answer this question, we first devise two models. The first builds upon the work of Barbosa et al. (2025). We use a decoder model to relate the essay (and prompt, when necessary) to the subcriteria and then use its output — in natural language — as input to an encoder model that learns to predict the score. We use BERTimbau-base (Souza et al., 2020) as the encoder model, as it has been shown to be a good compromise between size and performance, and GPT4o as the decoder, for it has shown good performance overall compared to Sabiá 3 and DeepSeek R1.

The output of the decoder can be seen as the explanation that the model would provide for its scoring. As the model is not deterministic, it was queried several times, which raises the question of how many explanations to use. We tried three strategies: *Single Explanation*, where we use only the first explanation given by the model; *Multiple Explanations*, where the model receives all the explanations, one at a time, and the final answer is decided through majority voting; *Concatenated Explanations*, where all the explanations are concatenated and passed as input to the model — not requiring the majority voting stage to decide the final answer.

The second model — which we will call NeSy — uses the subcriteria in its architecture. Instead of using only one BERT model, we use one for each dimension (which we call subnetworks) and treat the problem as probabilistic learning over boolean formulae. Many tools can be used for this, such as DeeProbLog (Manhaeve et al., 2018,

model	C1	C2	C3	C4	C5
BERTimbau-base (Barbosa et al., 2025)	.60	.36	.35	.55	**.63**
BERTimbau-large (Barbosa et al., 2025)	**.68**	.32	.24	**.60**	.46
GPT4o-Grader-full-context (Barbosa et al., 2025)	.53	**.47**	**.57**	.50	.32
GPT4o-Grader-essay-only (Barbosa et al., 2025)	.29	.18	.11	.45	.52
Single Explanation	.59	.29	.50	.36	.44
Multiple Explanations	.56	.33	.54	.40	.46
Concatenated Explanations	.64	.39	.51	.44	.47
Dist. Learning NeSy using Single Explanation	.66	.26	.46	.47	.47
Dist. Learning NeSy using Multiple Explanations	.60	.31	.49	.37	.42
Dist. Learning NeSy using Concatenating Explanations	.58	.32	.50	.23	.38
Dist. Learning NeSy using essay	.62	.24	.29	.56	.52
Pretrained (grader A) + Dist. Learning NeSy using essay	**.68**	.35	.29	.49	.61
Pretrained (grader B) + Dist. Learning NeSy using essay	.66	.41	.35	.51	.52

Table 3: Comparison of performances (QWK) of different models. In bold, the best performance in each competence.

2021), NeurASP (Yang et al., 2020), DPASP (Geh et al., 2023, 2024), and Scallop (Huang et al., 2021; Li et al., 2023). We employ the latter, as it allows for easy use of the GPU.[3] Here, we can use any input in natural language — we experiment with two different inputs: the explanation given by the Decoder model, as explained previously, and the essays. In these two cases, the model is learning in a *distant learning* fashion.

This second model allows us to use our subcriteria dataset to *pretrain* and test the subnetworks. In this case, we train the models for each dimension in the subcriteria dataset. As we have two graders, we can pretrain the subnetworks to simulate either of them. Then, we use AES-ENEM to further train in a distant supervision fashion. When performing finetuning in AES-ENEM, we kept the same learning rate (10^-5), minibatch size (1), and patience (20) for all traits. This decision was made to simplify the experiments, as it worked for all of them. The finetuning process stops when the QWK on the validation set is below the historical best for 20 consecutive epochs. Then, the best performing model is selected.

In order to fit the BERT models onto the GPU, we have to make some simplifications to the problem, as our equipment can only support up to four BERTs. These simplifications were made based on whether we had data to test that dimension. The final configuration was: C1 remained as it was; C2 had *theme* removed, and *pertinence* and *usage* became one dimension with 3 subcriteria; C3 had *direction* and *contradiction* removed; C4 remained the same; C5 kept only *elements*.

[3]The codes are available at: https://github.com/SilveiraIgor/NeuroSymbolic.

5 Results

Our first experiment is training the BERT models using the explanation output from GPT4o. The results are compared against two baseline models — BERTimbau-base and GPT4o-Grader-full-context — whose performance was reported by Barbosa et al. (2025). We tested the three possible usages of the explanation: Single, Multiple and Concatenated. The results are shown in Table 3. In the first section, we report the baseline performances, and in the second section, we present the performance of this experiment.

First, comparing the different usages of the explanations, the results tend to be close: the largest difference is in C2, where the amplitude is 0.1 QWK. Then, in four out of five cases, the best performance came from concatenating all the explanations — in the case where it lost, it was only by a small margin. This is reasonable, as it has access to more information than using a single one, and it can learn to combine the differing explanations, which is not done by Multiple Explanations.

Then, comparing Concatenated against the BERTimbau-base, we see that for the first three traits, our approach had better results, and for the last two, using the essay achieved better outcomes. If we take into consideration the confidence in-

terval reported by the authors[4], our results are either inside the interval or sufficiently close that a bootstrap would yield overlapping intervals. This suggests that explaining the score according to the guidelines results in a text that is as informative as the original essays. This is of particular importance for C2 and C3, where models should be able to relate the essay to additional material that does not fit into the context of traditional Encoder models.

It is also possible to compare the confidence intervals of GPT4o-full-context and we would reach the same conclusion. This raises the question: is this process of using the explanations as good as using only GPT4o? To answer this question, we check that only in trait five BERTimbau-base and GPT4o do not have overlapping intervals: the lower bound of BERTimbau is 0.51, while the upper bound of GPT4o is 0.43, our approach of concatenating the explanations is in the middle of the two. This suggests that this approach is doing something different from the two models independently.

Following, in the third section of Table 3, we investigate whether we can combine NeSy with the explanations — in this case, we do not pretrain the subnetworks. Here we see a slightly different pattern from the second section: when using NeSy as architecture, relying only on the first justification is the best approach in three out of five cases. Thus, we are using this method as representative of the third section. Comparing NeSy using First Explanation with Concatenated Justification, we see that the logical program has inferior performance in two traits (C2 and C3). If we take into account that the NeSy represents many BERT models together, we also need to consider the influence of having more parameters in the model. One way to do this is to compare the reported performances of BERTimbau-base and BERTimbau-large. We see that it increased in C1 and C4, and decreased in the others. Comparing the second and third sections row-wise, we see that all strategies lost performance in C2 and C3, but from the three that should increase, one always decreases — overall, it follows the pattern. From this analysis, we validate that the approach based on NeSy is competitive with the traditional one-network approach. Furthermore, we highlight that NeSy seems to be influenced by the number of parameters in a manner similar to increasing the model size.

[4]Available at: https://huggingface.co/datasets/kamel-usp/jbcs2025_experiments_report

Next, we turn to NeSy using the essays. The result is presented in the fourth section of Table 3. Comparing it with the baseline BERTimbau-base and BERTimbau-large, we can see that the NeSy approach is also competitive when using the essays. This shows that this approach was not dependent on the structured output of GPT4o.

In the last section of the same table, each network is first pretrained in our dataset to identify each dimension — according to each one of the graders. Then, they are finetuned in a distant supervision fashion — using the AES-ENEM dataset. In these cases, the performance remained the same or increased in four traits (C1, C2, C3 and C5), when compared against the case without pretraining. This presents evidence that the subnetworks may be doing what they are supposed to do, as the pretraining provided a better initialization of the weights. As with the first approach, the increases and decreases in performance are not enough to make the methods statistically different.

6 Investigating NeSy's Learning

In this section, we investigate some aspects that were not covered previously: could we achieve better results if we had invested in more training time? Are the subnetworks of NeSy with pretraining and distant learning learning what they are supposed to?

In order to answer these questions, we devised two additional experiments: the first involves comparing the average agreement of the subnetworks with the annotators, and the second consists of running it for a longer time — in other words, increasing the patience from 20 to 100. For these experiments, we are using C1, as it is the competence that does not require any adaptation for its usage. The performances evaluated in QWK are presented in Figure 2.

We notice two interesting points from the previous image. First, although there is a spike in performance on the validation set, the respective performance on the test set does not reflect the same spike — moreover, there are many similar performances on the test set that come from relatively smaller performances on the validation set. Nonetheless, after 20 epochs, the performances on the test set rest between .65 and .70. The second point is related to this performance range: the network still did not overfit — the training loss was at .009 in the last epoch —, so additional training

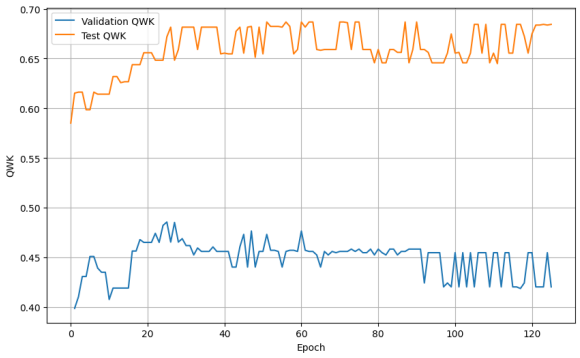

Figure 2: QWK performances in validation and test for C1.

Epoch	validation	test	syntax	mistake
0	–	.58	.40	.39
25	.48	.68	.39	.47
125	.42	.68	.39	.46

Table 4: Model performance at different epochs. Except for Epoch, all columns are reporting QWK. Syntax and Mistake are the average w.r.t the two graders.

might lead to better results. With this, we can see that we might not have reached the limit of this approach. Additionally, the fact that we used the same hyper-parameters for all traits might suggest that specialized training routines might significantly increase performance.

Then, we test whether the subnetworks are learning what they are supposed to learn. In order to do so, we first pretrain them on our dataset and then test their performance on AES-ENEM and their agreement with the annotators at the subcriteria level on our subcriteria dataset. After that, we train them through distant learning using only AES-ENEM. The results are presented in Table 4.

The first row of the table indicates the zero-shot performance in the AES-ENEM dataset — in other words, the subnets were only trained on the subcriteria dataset. Here we already have an interesting result, as it shows that they were capable of properly transferring the knowledge to AES-ENEM (.58 QWK) and that it reached the agreement of the graders (.395 here against .38 in Table 2). In the following rows, the subnets were trained in a distant supervision fashion using the AES-ENEM dataset. The second row presents the model selected by validation, the last row shows the model when the last epoch (patience 0) is reached. Naturally, the test performance increased compared to the first row. Nonetheless, the syntax performance decreased slightly, while the mistake performance increased. Summing up these facts, the quality of the network's scoring and its justification increased. This shows the potential of this approach.

7 Conclusion

In this work, we presented two major contributions. The first consists of two novel neuro-symbolic approaches designed to leverage the subcriteria present in the Graders Guideline of the Brazilian National Entrance Exam (ENEM). The second is a small dataset of essays annotated according to these subcriteria.

Our first neuro-symbolic approach uses the large context window of GPT4o in order to relate essays to the subcriteria. Then, we treat this output as an explanation and use it as input instead of the essays — this can be seen as compacting the relevant information. We showed that this novel strategy is competitive with the current state-of-the-art. Although we used GPT4o for our experiment to generate the explanation for the grade and then used BERT to grade the essay, the same process could be done using any Decoder for the first stage and any fine-tunable model for the second. The strong point of this first approach is validating the explanation.

In order to test whether the subcriteria can be properly used for grading, we constructed a dataset in which essays were annotated independently by two experienced graders according to the subcriteria. This task is much more laborious than the original, as each essay has to be evaluated across 20 dimensions. As a consequence, we were only able to annotate 63 essays. Then, we used our dataset to pretrain the networks of the second approach and showed that it could lead to better results. Our dataset suffers from the same setback as all other AES datasets: the data available is not representative enough to learn all patterns — especially for low scoring essays. We propose further investigation on how to mitigate this by employing a strategy similar to Rocha et al. (2023) combined with the subcriteria. In this scenario, it is possible to ask conversational models to alter an essay in order to have a defined characteristic that will be used only by one network.

Our second neuro-symbolic approach combines the subcriteria through boolean formulae in order to produce a score. We showed that this approach

is also competitive with previous methods when using the essay or the explanations. With this formulation, it is possible to take one step further: instead of just scoring, it is also possible to provide a reason for the score. Finally, we demonstrated, using trait C1, that the subcriteria can be learned from our small dataset and subsequently used on another dataset with a good level of performance. Additionally, further (distant) training on the second dataset increased performance in the second dataset and also increased agreement with annotators on the first dataset. This further highlights the potential of this approach. In this work, we did not explore the full potential of this approach, as it allows for changing the subnetworks, using different models for different dimensions, and varying the learning rate for each network, to name a few. Here, our intent was to demonstrate its feasibility and advantages: competitive in performance, more explainable, and it can be improved through distant learning.

Limitations

As stated previously, the limitations of our work can be divided into three parts. First, our dataset is not representative enough to cover all grading patterns. Second, we employed only GPT4o as the decoder model — further research could employ other proprietary or free models. Finally, our NeSy model was not fully optimized, as our goal was not to surpass the state-of-the-art but rather show that it is competitive with it.

Acknowledgments

This work was partially supported by the São Paulo Research Agency (FAPESP) Grant no. 2022/02937-9, CNPq Grant no. 305136/2022-4 and CAPES 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.

Rafael Anchiêta, Anthony Luz, Shara Lopes, and Raimundo Moura. 2025. A zero-shot prompting approach for automated feedback generation on enem essays. In *Proceedings of the 31st Brazilian Symposium on Multimedia and the Web*, pages 511–515, Porto Alegre, RS, Brasil. SBC.

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.

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

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.

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.

Renato Lui Geh, Jonas Gonçalves, Igor C. Silveira, Denis D. Mauá, and Fabio G. Cozman. 2024. dPASP: A Probabilistic Logic Programming Environment For Neurosymbolic Learning and Reasoning. In *Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning*, pages 731–742.

Renato Lui Geh, Jonas Gonçalves, Igor Cataneo Silveira, Denis Deratani Mauá, and Fabio Gagliardi Cozman. 2023. dpasp: A comprehensive differentiable probabilistic answer set programming environment for neurosymbolic learning and reasoning. *Preprint*, arXiv:2308.02944.

Jiani Huang, Ziyang Li, Binghong Chen, Karan Samel, Mayur Naik, Le Song, and Xujie Si. 2021. Scallop: From probabilistic deductive databases to scalable differentiable reasoning. In *Advances in Neural Information Processing Systems*, volume 34, page 25134–25145. Curran Associates, Inc.

Shiyu Ji, Farnoosh Hashemi, Joice Chen, Juanwen Pan, Weicheng Ma, Hefan Zhang, Sophia Pan, Ming Cheng, Shubham Mohole, Saeed Hassanpour, Soroush Vosoughi, and Michael Macy. 2025. A generalizable rhetorical strategy annotation model using

LLM-based debate simulation and labelling. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20482–20503, Suzhou, China. Association for Computational Linguistics.

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.

Ziyang Li, Jiani Huang, and Mayur Naik. 2023. Scallop: A language for neurosymbolic programming. *Proc. ACM Program. Lang.*, 7(PLDI).

Jiawei Liu, Yang Xu, and Yaguang Zhu. 2019. Automated essay scoring based on two-stage learning. *Preprint*, arXiv:1901.07744.

Robin Manhaeve, Sebastijan Dumancic, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. 2018. Deepproblog: Neural probabilistic logic programming. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Robin Manhaeve, Sebastijan Dumančić, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. 2021. Neural probabilistic logic programming in deepproblog. *Artificial Intelligence*, 298:103504.

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.

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. 2024a. Automatic text readability assessment in European Portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 97–107, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Eugénio Ribeiro, Nuno Mamede, and Jorge Baptista. 2024b. 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.

Victor Hugo Nascimento Rocha, Igor Cataneo Silveira, Paulo Pirozelli, Denis Deratani Mauá, and Fabio Gagliardi Cozman. 2023. Assessing good, bad and ugly arguments generated by chatgpt: a new dataset, its methodology and associated tasks. In *Progress in Artificial Intelligence: 22nd EPIA Conference on Artificial Intelligence*, page 428–440.

Mark D Shermis and Jill Burstein. 2013. Handbook of automated essay evaluation. *NY: Routledge*.

Igor Cataneo Silveira, André Barbosa, Daniel Silva Lopes da Costa, and Denis Deratani Mauá. 2025a. 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.

Igor Cataneo Silveira, Eugénio Ribeiro, Nuno Mamede, and Jorge Baptista. 2025b. Aprendizado por transferência para correçao automática de redaçao. *Linguamática*, 17(2):99–116.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 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)*.

Jingyi Xu, Zilu Zhang, Tal Friedman, Yitao Liang, and Guy Van den Broeck. 2018. A semantic loss function for deep learning with symbolic knowledge. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5502–5511. PMLR.

Zhun Yang, Adam Ishay, and Joohyung Lee. 2020. Neurasp: Embracing neural networks into answer set programming. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 1755–1762. International Joint Conferences on Artificial Intelligence Organization. Main track.