From Zero-shot to Self-generated References: Leveraging LLMs for Scoring ENEM Essays

Matheus Yasuo Ribeiro Utino[1], **Paulo Mann**[2]

[1] Institute of Mathematics and Computer Science, University of São Paulo

[2]Institute of Computing, Federal University of Rio de Janeiro

matheusutino@usp.br, paulomannjr@gmail.com

Abstract. *This study investigates the application of Large Language Models (LLMs) to Automated Essay Scoring (AES) in the context of Brazil's Exame Nacional do Ensino Médio (ENEM). We evaluate five state-of-the-art LLMs across three prompting scenarios: zero-shot, one-shot (with high-score references), and a novel self-generated reference approach, where the model generates its own ideal reference before evaluation. Using the Essay-BR corpus, we assess performance using both classification and regression metrics. Results show that one-shot prompting consistently yields the best metrics, while the self-generated reference method presents a viable alternative when no real references are available. Our findings highlight the promise of LLMs for educational scoring.*

1. Introduction

The *Exame Nacional do Ensino Médio* (ENEM) is Brazil's primary educational assessment, playing a central role in public university admissions and access to government programs such as Sisu, Prouni, and FIES [Inep 2020]. With millions of participants annually, ENEM evaluates a broad range of competencies, including written production, which is graded by human evaluators based on objective criteria. The essay component, in particular, represents a critical stage, as it synthesizes students' linguistic, argumentative, and socio-cognitive skills.

More than a mere evaluation tool, ENEM functions as a mechanism for social mobility and the democratization of higher education access [Pires 2023]. Consequently, accurately predicting student performance, especially in the essay section, can inform the design of public education policies, the personalization of pedagogical interventions, and the early identification of risks related to dropout or poor academic achievement.

With recent advances in Large Language Models (LLMs), it has become possible to explore novel approaches to complex tasks requiring sophisticated textual comprehension and production, such as Automated Essay Scoring (AES) [Atkinson and Palma 2025]. These models have demonstrated capabilities not only in classifying texts with high accuracy but also in providing structured textual justifications, making them particularly promising for educational applications where interpretability is a fundamental requirement. These outcomes are closely tied to the emergent abilities of LLMs, which enable them to perform complex reasoning and explanation tasks beyond their training objectives [Berti et al. 2025].

Within this context, the present study investigates the use of LLMs to predict ENEM essay scores, evaluating different usage configurations (zero-shot, one-shot, and

a novel approach herein termed self-generated reference). In addition to analyzing the predictive performance of the models, we explore their capacity to simulate human evaluative behavior, generate patterns of textual excellence, and provide formative feedback. Through this work, we aim to contribute to the advancement of AES in Portuguese, emphasizing scalable, transparent, and pedagogically valuable solutions.

The main contributions of this work are:

- The proposal of a novel self-generated reference strategy, in which the model creates its own ideal essay based solely on the theme and uses it as a benchmark for evaluation.
- A comparative evaluation of five open-source LLMs across three usage configurations (zero-shot, one-shot with real reference, and self-generated reference) in the context of ENEM essay scoring.
- The creation and public release of a new synthetic dataset containing LLM-generated essays, predicted scores, and detailed justifications for each ENEM competency.
- Evidence that smaller LLMs can be competitive with larger models in specific scenarios, highlighting important trade-offs between performance and computational efficiency.

2. Related Work

[Mello et al. 2024] present solutions developed for the AES competition focused on narrative texts written by middle school students in Brazil. The study compares several approaches for predicting students' proficiency levels across four evaluation criteria: Formal Register, Narrative Rhetorical Structure, Thematic Coherence, and Textual Cohesion, using a corpus of 1,235 essays. The most effective systems relied on pre-trained language models, such as BERTimbau and Albertina, applied to both classification and regression tasks. Notably, the PiLN model achieved the best performance in Thematic Coherence and Formal Register. The study highlights the potential of pre-trained language models for AES in Portuguese, despite the inherent complexity and subjectivity of the task.

[Marinho et al. 2022a] publicly released a corpus of argumentative essays written by Brazilian high school students, evaluated according to the ENEM criteria. The study also presents experiments for predicting the overall essay score using two feature-based approaches: one employing linear regression with 19 handcrafted features proposed by [Amorim et al. 2018], and another using 681 features with a gradient boosting regressor inspired by [Fonseca et al. 2018]. We rely on these baselines, which employ linear regression and gradient boosting using handcrafted features, for performance comparison purposes.

[Marinho et al. 2022b] developed independent AES models for each of the five ENEM writing competencies, employing three different approaches: handcrafted feature engineering, Doc2Vec embeddings, and Long Short-Term Memory (LSTM) neural networks. Their findings indicate that feature-based methods were more effective for Competencies 1 and 2 (which focus on formal language aspects), while LSTM networks performed better on Competencies 3, 4, and 5 (which involve more subjective dimensions). Their models showed improved results over previous works in Portuguese, achieving moderate agreement using the QWK metric.

[Silveira et al. 2024] evaluated different approaches for predicting competency-based essay scores, including classifiers, traditional regressors, and ordinal regressors based on the pre-trained BERTimbau model, using essays aligned with the ENEM format. The authors also explored supervised and unsupervised pretraining strategies with data from two distinct sources. Their results show that ordinal regression with BERTimbau-base performs competitively, outperforming traditional feature-based models. The study underscores the potential of combining pre-trained language models with additional resources to further advance AES in Portuguese.

3. Methodology

This section presents the dataset, the models, and the experimental setup.

3.1. Dataset: The Essay-BR Corpus

To evaluate our proposed approaches, we used the Essay-BR Corpus, a publicly available dataset designed for the task of AES in Brazilian Portuguese [Marinho et al. 2022a]. This corpus fills a significant gap in the field, as most AES resources are focused on English, and there is a shortage of manually annotated corpora in Portuguese that follow the criteria of the Brazilian ENEM. The essays were scraped from two public platforms, *Vestibular UOL* and *Educação UOL*, and underwent a normalization process to align their grading with ENEM's official rubric. The dataset includes both holistic scores and detailed sub-scores per competence, allowing for more granular analyses.

The Essay-BR Corpus consists of 4570 argumentative essays written by Brazilian high school students, distributed across 86 topics that cover topics such as politics, public health, human rights, and contemporary social issues. Each essay was manually scored by experts based on the five official ENEM competences, which assess grammar, textual structure, argument development, coherence, and the proposal of a solution to the discussed problem. Each competence is scored from 0 to 200 in increments of 40, leading to a total score ranging from 0 to 1000. To structure the corpus, we adopt the original data split proposed by the dataset authors: 70% for training (3,198 essays), 15% for development (686 essays), and 15% for testing (686 essays).

Each of the five ENEM competences captures a distinct dimension of writing quality. Competence 1 (C1) evaluates the command of the formal norms of the Portuguese language, focusing on grammar and orthography. Competence 2 (C2) assesses the understanding of the proposed topic and the ability to apply knowledge from various areas to construct a coherent argument. Competence 3 (C3) measures the selection, organization, and interpretation of information, facts, and opinions to defend a point of view. Competence 4 (C4) examines the proper use of linguistic mechanisms to structure arguments and ensure textual cohesion. Finally, Competence 5 (C5) evaluates the ability to elaborate a proposal for solving the problem discussed in the essay, respecting human rights and demonstrating feasibility and coherence [INEP 2024].

3.2. Large Language Models

LLMs are deep neural architectures with billion of parameters and trained on extensive text corpora, capable of understanding, generating, and reasoning over natural language [Minaee et al. 2024]. Based primarily on the Transformer architecture, these models achieve state-of-the-art performance across diverse tasks including text classification,

question answering, summarization, and dialogue generation [Qin et al. 2024]. Recent advances have extended their applicability beyond traditional NLP to domains such as education, finance, healthcare, and law [Xu et al. 2024, Chen et al. 2024]. However, their generalization capabilities in tasks requiring structured understanding or quantitative reasoning, such as essay scoring in Portuguese, remain underexplored.

We evaluate five open-source state-of-the-art LLMs with varying parameter sizes and architectural designs. Our objective is to compare their performance across different usage paradigms (zero-shot, one-shot, and self-generated reference) and assess the relationship between model scale, reasoning ability, and predictive quality.

- **Qwen3-8B**: An 8-billion-parameter model from the Qwen3 series, optimized for instruction-following and efficiency. It uniquely supports seamless switching between complex reasoning and general dialogue modes, with strong multilingual capabilities and superior alignment to human preferences.
- **DeepSeek-R1-0528-Qwen3-8B**: A fine-tuned variant of Qwen3-8B using the DeepSeek R1 framework, enhanced for improved task generalization, reasoning, and response consistency, particularly in multilingual and zero-shot settings.
- **Mistral-Small-3.1-24B-Instruct-2503**: A 24-billion-parameter instruction-tuned model designed for both reasoning-intensive and high-throughput tasks. It supports long context windows and excels in text and vision understanding, while enabling efficient local deployment.
- **LLaMA-3.1-nemotron-ultra-253b-v1**: A 253-billion-parameter reasoning-focused model derived from Meta's LLaMA 3.1. Leveraging Neural Architecture Search for an optimal accuracy-efficiency tradeoff, it is fine-tuned for advanced reasoning, chat, tool use, and supports very long contexts.
- **DeepSeek-R1T-Chimera**: A 685-billion-parameter mixture-of-experts (MoE) model merging DeepSeek-R1 and DeepSeek-V3. By dynamically activating parameter subsets during inference, it optimizes computational resources while integrating reasoning and token efficiency from prior DeepSeek versions.

These models represent a diverse spectrum of sizes and reasoning capabilities, enabling a comprehensive analysis of how scale and architectural innovations impact performance in automated evaluation tasks. By encompassing architectures ranging from compact, efficient models to extremely large-scale mixture-of-experts systems, the evaluation sheds light on the trade-offs between computational resource demands and predictive accuracy. Moreover, the inclusion of models with specialized fine-tuning for reasoning and instruction-following allows us to investigate the extent to which targeted training enhances performance on tasks that require structured understanding and complex inference. This comprehensive comparison is crucial for identifying the most suitable model configurations for real-world applications, particularly in educational contexts where nuanced evaluation and scalability are essential. Ultimately, these insights contribute to guiding future model development toward balancing efficiency, interpretability, and advanced reasoning capabilities.

3.3. Experimental Setup

In this study, we investigate the use of LLMs for predicting the score category of ENEM essays, considering three different evaluation scenarios:

1. **Zero-shot**: the model receives the theme description, the essay title, and the full student essay to be evaluated. No reference examples are provided. The model must predict the score category directly based on this input alone.
2. **One-shot**: in addition to the theme description, the essay title, and the essay to be evaluated, the model is given a single reference essay — the highest-scoring essay available for that same theme. This reference serves as a comparative basis to guide the model's prediction.
3. **Self-generated reference**: a novel approach proposed in this study, where the model is first prompted to generate a hypothetical essay that would achieve a perfect score (1000) based solely on the theme description and the essay title, without access to any real student submissions. This generated essay is then used as a reference in a one-shot evaluation setting, where the model assesses the actual student essay. This approach aims to simulate human-like evaluation behavior by allowing the model to define its own ideal standard of excellence.

The evaluation prompt provides the model with a detailed and authoritative context, simulating the role of an official ENEM essay evaluator. It establishes clear expectations regarding the scoring process by grounding the model in the five official competencies defined by INEP. These competencies encompass language mastery, thematic comprehension, argument structure, textual cohesion, and proposal of intervention. Furthermore, the prompt instructs the model to adopt an impartial, rigorous, and criterion-based approach, explicitly referencing the normative guidelines of the ENEM scoring paradigm. It also includes information about disqualifying criteria that result in an automatic zero, ensuring the model adheres to institutional correction standards.

For each competency, the model provides a predicted score and a detailed explanation justifying that score. This design ensures both granularity and transparency in the evaluation process. In addition to competency-level feedback, the model also generates general comments that synthesize the essay's overall strengths and weaknesses, along with specific suggestions for improvement. These supplementary fields enhance the pedagogical value of the feedback by offering actionable guidance, thereby supporting students in refining their writing skills. By combining structured scoring, interpretability, and formative commentary, this output format emulates human-like evaluation practices and facilitates meaningful educational interventions. Details about prompts and implementation are available on GitHub[1].

Furthermore, for comparison purposes, we adopt as baselines the results reported by [Marinho et al. 2022a], which include a linear regression model with 19 handcrafted features originally proposed by [Amorim et al. 2018], here referred to as LinearHandcrafted-19, and a gradient boosting regressor trained on 681 handcrafted features following the approach of [Fonseca et al. 2018], referred to as BoostedHandcrafted-681.

4. Metrics

For the classification tasks, we use the Quadratic Weighted Kappa (QWK), a metric that measures the level of agreement between two raters, adjusted for chance, and penalizes disagreements based on the squared distance between categories. This is especially useful

[1]https://github.com/Matheusutino/LLM-essay-scoring

for ordinal classification problems, such as the categorization of essay scores, where the order of the classes matters. Formally, the QWK is defined as:

$$\text{QWK} = 1 - \frac{\sum_{i,j} W_{i,j} O_{i,j}}{\sum_{i,j} W_{i,j} E_{i,j}} \tag{1}$$

where O is the observed confusion matrix, E is the expected matrix assuming independence between raters, and W is a weight matrix computed as:

$$W_{i,j} = \frac{(i-j)^2}{(k-1)^2} \tag{2}$$

with k being the total number of possible categories. QWK values range from -1 (complete disagreement) to 1 (perfect agreement), with 0 indicating agreement at chance level.

For regression tasks, model performance is assessed using the Root Mean Squared Error (RMSE). This metric calculates the square root of the average squared differences between predicted and actual values, assigning greater weight to larger errors and thus being particularly sensitive to outliers.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3}$$

5. Results and Discussions

As shown in Table 1, in the classification task, LLM-based models outperformed the baseline across all competencies except for Competency C4. This finding suggests that these models still face challenges in accurately assessing textual cohesion in essays. Among the evaluated models, LLaMA-3.1-nemotron-ultra-253b-v1 stood out by achieving the highest scores in both Competency C1 and the overall evaluation, underscoring its potential as the leading model for this classification task. DeepSeek-R1T-Chimera also performed strongly, particularly excelling in Competencies C2 and C5. Moreover, it demonstrated consistent results across all settings, making it a standout model. These findings suggest that larger models may be better equipped to capture subtle nuances during essay evaluation.

Figure 1 shows that the one-shot setting achieved, on average, the best performance across all competencies. This result suggests that providing a high-quality essay as an example on the same topic offers an effective reference point for LLMs, significantly contributing to the accuracy of their evaluations. Additionally, the self-generated reference setting led to improvements in competencies C1, C4, and in the overall score in relation of zero-shot. This indicates that, although it does not reach the same performance levels as the one-shot setting, the LLM is still capable of generating texts with sufficient quality to serve as a reference. Therefore, this approach appears promising, especially in scenarios where no real example base is available for applying the one-shot strategy.

Table 2 reveals a more heterogeneous pattern of results, in which DeepSeek-R1T-Chimera once again delivered consistent performance across different settings. Nevertheless, it is noteworthy that the simplest model, Qwen3-8B, achieved the lowest over-

Table 1. QWK results on the test set for each competency (C1 to C5) and the overall total. The metric was computed for different models under settings. Underlined values indicate the best result within each setting, while bolded values highlight the best overall result per competency across all settings.

Setting	Model	C1	C2	C3	C4	C5	Total
Zero-shot	Qwen3-8B	0.25	0.21	0.30	0.20	0.19	0.29
	DeepSeek-R1-0528-Qwen3-8B	0.17	0.24	0.34	0.25	0.22	0.32
	Mistral-Small-3.1-24B-Instruct-2503	0.30	0.42	0.43	0.22	0.32	0.48
	LLaMA-3.1-nemotron-ultra-253b-v1	0.40	0.38	0.37	0.24	0.28	0.45
	DeepSeek-R1T-Chimera	0.37	0.43	0.46	0.32	0.36	0.50
One-shot	Qwen3-8B	0.40	0.41	0.40	0.38	0.33	0.51
	DeepSeek-R1-0528-Qwen3-8B	0.29	0.37	0.38	0.37	0.32	0.46
	Mistral-Small-3.1-24B-Instruct-2503	0.35	0.48	**0.53**	0.31	0.34	0.53
	LLaMA-3.1-nemotron-ultra-253b-v1	**0.43**	0.50	0.50	0.35	0.35	**0.57**
	DeepSeek-R1T-Chimera	0.38	**0.54**	0.52	0.38	**0.40**	0.56
Self-generated reference	Qwen3-8B	0.33	0.30	0.33	0.29	0.27	0.39
	DeepSeek-R1-0528-Qwen3-8B	0.26	0.24	0.28	0.24	0.24	0.34
	Mistral-Small-3.1-24B-Instruct-2503	0.35	0.37	0.40	0.24	0.32	0.50
	LLaMA-3.1-nemotron-ultra-253b-v1	0.40	0.35	0.42	0.31	0.35	0.51
	DeepSeek-R1T-Chimera	0.41	0.43	0.48	0.36	0.36	0.53
LinearHandcrafted-19	Linear Regression	0.35	0.44	0.39	0.37	0.34	0.47
BoostedHandcrafted-681	Gradient Boosting	0.42	0.46	0.40	**0.45**	0.36	0.51

Table 2. RMSE results on the test set for each competency (C1 to C5) and the overall total. The metric was computed for different models under settings. Underlined values indicate the best result within each setting, while bolded values highlight the best overall result per competency across all settings.

Setting	Model	C1	C2	C3	C4	C5	Total
Zero-shot	Qwen3-8B	41.06	41.88	36.97	48.63	57.39	186.54
	DeepSeek-R1-0528-Qwen3-8B	50.35	47.71	40.98	54.06	67.32	208.65
	Mistral-Small-3.1-24B-Instruct-2503	45.28	37.35	35.13	52.19	61.34	165.35
	LLaMA-3.1-nemotron-ultra-253b-v1	33.81	38.54	35.62	57.04	60.53	165.81
	DeepSeek-R1T-Chimera	35.26	36.27	34.18	46.32	55.21	159.98
One-shot	Qwen3-8B	35.42	39.68	36.17	**40.32**	49.68	**151.35**
	DeepSeek-R1-0528-Qwen3-8B	41.85	43.11	40.69	47.61	60.65	177.92
	Mistral-Small-3.1-24B-Instruct-2503	42.10	36.40	33.25	51.88	61.64	167.08
	LLaMA-3.1-nemotron-ultra-253b-v1	34.56	38.61	35.46	54.19	58.83	160.25
	DeepSeek-R1T-Chimera	36.53	**34.01**	**33.21**	46.60	54.45	157.41
Self-generated reference	Qwen3-8B	39.71	41.88	37.81	44.81	53.99	171.63
	DeepSeek-R1-0528-Qwen3-8B	42.13	43.36	42.52	52.99	63.61	188.52
	Mistral-Small-3.1-24B-Instruct-2503	39.50	38.58	35.62	51.84	60.03	157.68
	LLaMA-3.1-nemotron-ultra-253b-v1	35.42	46.57	38.42	51.59	58.41	165.40
	DeepSeek-R1T-Chimera	**33.77**	37.63	33.56	44.21	54.06	153.24
LinearHandcrafted-19	Linear Regression	34.82	36.14	40.27	42.23	49.02	163.92
BoostedHandcrafted-681	Gradient Boosting	34.16	35.76	40.02	40.48	**48.28**	159.44

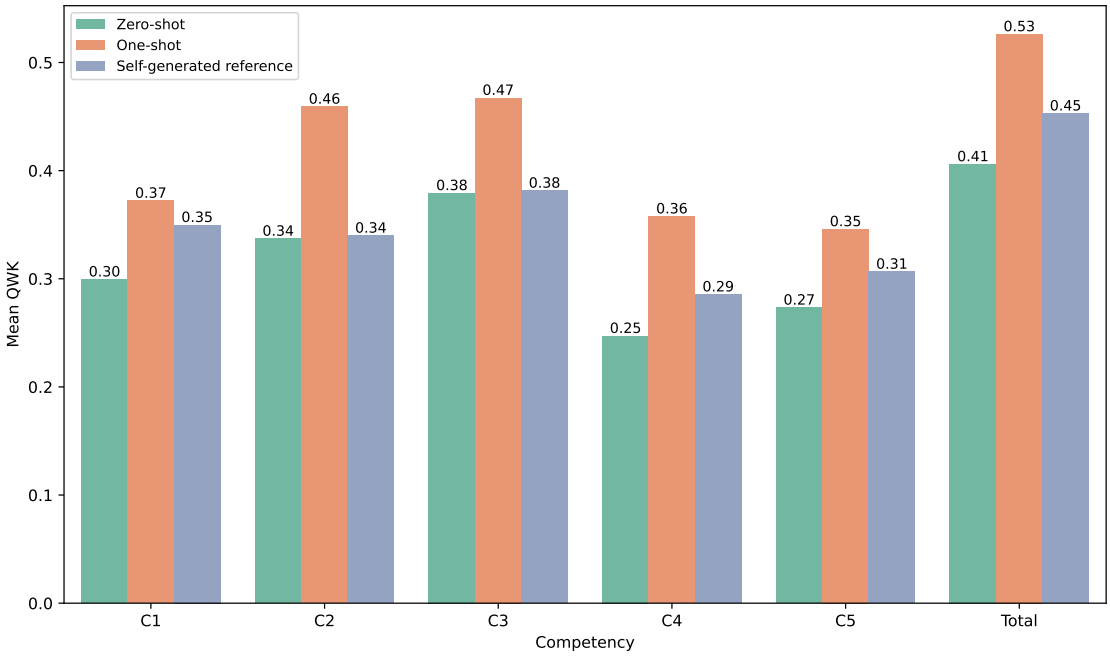

Figure 1. Mean QWK for each competency (C1 to C5) and the overall average, grouped by setting.

all RMSE, suggesting that smaller models may be sufficient for regression tasks in this context. Despite this, the results also indicate a general difficulty across all models in accurately estimating C5, highlighting a persistent challenge in modeling this particular dimension.

Figure 2 shows that, similarly to the classification task, the one-shot approach outperformed the other configurations across all competencies. However, as a counterpoint, for competencies C2 and C3, the self-generated reference approach resulted in a slight performance drop compared to the zero-shot version, while still maintaining an advantage in the remaining competencies.

According to Table 3, the one-shot setting demonstrated greater consistency in the QWK metric across all competencies, showing the lowest dispersion values. However, it outperformed in terms of RMSE in only two competencies. This suggests that results in this scenario were more stable overall. In contrast, the zero-shot setting exhibited the highest variability, particularly in the total RMSE (20.27), indicating greater instability in predictions. The self-generated reference produced intermediate standard deviations, with more stable performance than the zero-shot setting, reinforcing its potential as a viable alternative when real reference examples are not available.

Table 3. Standard deviation of QWK (left) and RMSE (right) on the test set for each competency (C1 to C5) and the overall total. Values are reported for different evaluation settings.

Setting	C1	C2	C3	C4	C5	Total
Zero-shot	0.095 | 6.90	0.105 | 4.62	0.063 | 2.66	0.045 | 4.26	0.069 | 4.60	0.097 | 20.27
One-shot	0.051 | 3.62	0.069 | 3.43	0.070 | 3.06	0.028 | 5.35	0.033 | 4.96	0.043 | 10.17
Self-generated reference	0.063 | 3.42	0.069 | 3.63	0.080 | 3.36	0.050 | 4.22	0.051 | 4.10	0.083 | 13.81

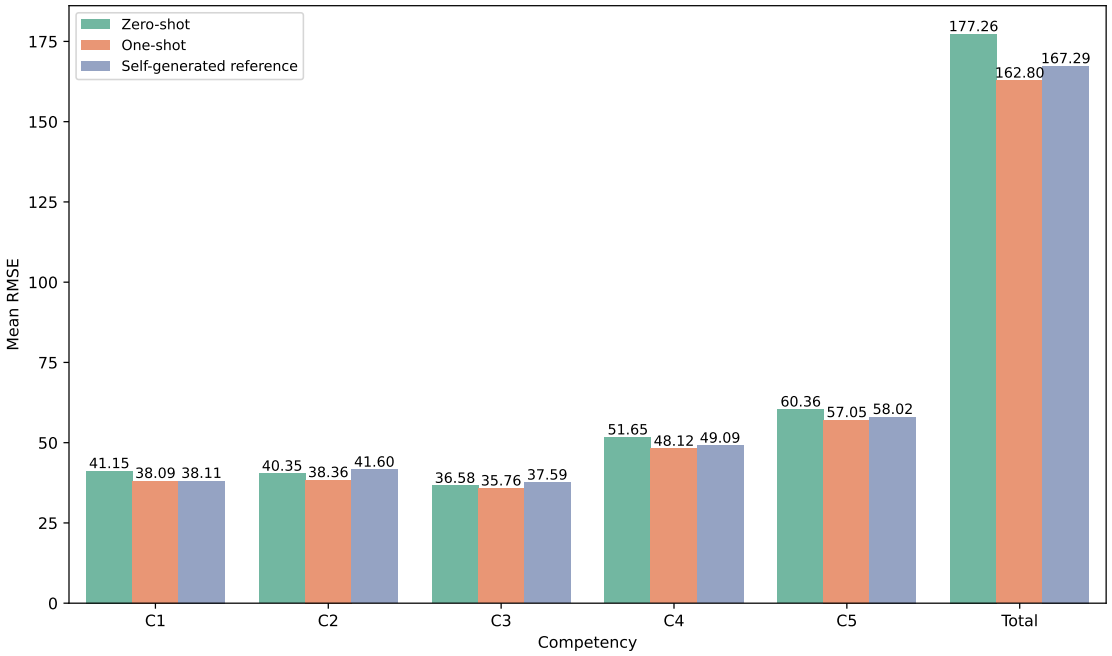

Figure 2. Mean RMSE for each competency (C1 to C5) and the overall average, grouped by setting.

Based on these findings, it becomes clear that larger models may offer some advantages but do not necessarily guarantee superior results across all tasks or competencies. While large-scale models such as LLaMA-3.1-Nemotron-Ultra-253B and DeepSeek-R1T-Chimera stood out in certain metrics, smaller models like Qwen3-8B demonstrated competitive performance, particularly in regression tasks. This highlights an important trade-off between computational cost and performance, suggesting that smaller and more efficient models can be viable alternatives, especially in resource-constrained scenarios.

Beyond quantitative improvements, one of the key advantages of LLMs over traditional automated essay scoring methods lies in their ability to provide detailed justifications for each evaluated competency. Unlike classical approaches based solely on regression or classification, the analyzed LLMs generate explanations that accompany the assigned scores, as well as overall comments on the strengths and weaknesses of each essay. This feature adds transparency to the evaluation process and brings the model's behavior closer to that of human raters. To support further research, we release the new synthetic dataset of essays generated by the LLMs, along with their predictions and detailed explanations. This dataset, derived from Essay-BR, is available on our GitHub repository, providing a valuable resource for researchers to study and build upon interpretable automated essay scoring.

In this context, this study investigates the use of LLMs for predicting ENEM essay scores, evaluating different usage configurations (zero-shot, one-shot, and an original approach called self-generated reference). In addition to analyzing the predictive performance of the models, we explore their ability to simulate human evaluative behaviors, generate patterns of textual excellence, and provide formative feedback. Through this, we aim to contribute to the advancement of automated text evaluation in Portuguese, with an

emphasis on scalable, transparent, and pedagogically useful solutions. This approach is particularly important to provide support methods that can help improve education in a country marked by profound social inequalities and a significant educational deficit, such as Brazil [Magalhães 2023, Garcia 2024].

6. Conclusion

Our study has demonstrated that LLMs, particularly when provided with reference essays, can approximate human-level performance in the automated scoring of ENEM-style essays. The proposed self-generated reference approach emerged as a promising alternative in contexts where real references are unavailable, underscoring the capacity of LLMs to autonomously generate idealized textual benchmarks. The findings highlight the potential of LLMs to enhance educational assessment systems by offering not only accurate predictions but also structured justifications and formative feedback with pedagogical value.

As future work, fine-tuning LLMs on locally curated corpora annotated according to ENEM's official criteria emerges as a promising direction to better align model behavior with the linguistic and cultural characteristics of the Brazilian context. Additionally, enhancing pedagogical feedback through the incorporation of personalized student data, such as performance history, previous essays, and prior interactions with the model, could support longitudinal tracking of student progress and enable the delivery of context-aware, adaptive guidance. This approach holds potential for fostering more effective and learner-centered educational interventions.

References

Amorim, E., Cançado, M., and Veloso, A. (2018). Automated essay scoring in the presence of biased ratings. In Walker, M., Ji, H., and Stent, A., editors, *NAACL 2018, Volume 1*, pages 229–237.

Atkinson, J. and Palma, D. (2025). An llm-based hybrid approach for enhanced automated essay scoring. *Scientific Reports*, 15(1):14551.

Berti, L., Giorgi, F., and Kasneci, G. (2025). Emergent abilities in large language models: A survey. *arXiv preprint arXiv:2503.05788*.

Chen, Z. Z., Ma, J., Zhang, X., Hao, N., Yan, A., Nourbakhsh, A., Yang, X., McAuley, J., Petzold, L., and Wang, W. Y. (2024). A survey on large language models for critical societal domains: Finance, healthcare, and law. *arXiv preprint arXiv:2405.01769*.

Fonseca, E., Medeiros, I., Kamikawachi, D., and Bokan, A. (2018). Automatically grading brazilian student essays. In *PROPOR 2018*, page 170–179, Berlin, Heidelberg. Springer-Verlag.

Garcia, G. (2024). Desigualdade: 63% da riqueza do brasil está nas mãos de 1% da população, diz relatório da oxfam. Accessed on: Jun 6, 2025.

Inep (2020). Histórico enem. `https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem/historico`. Accessed at: 7 jun. 2025.

INEP (2024). *The ENEM Essay: Participant's Handbook 2024*. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). Accessed: 2025-06-07.

Magalhães, T. (2023). Brasil tem baixo desempenho e estagna em ranking mundial da educação básica. Accessed on: Jun 6, 2025.

Marinho, J. C., Anchiêta, R. T., and Moura, R. S. (2022a). Essay-br: a brazilian corpus to automatic essay scoring task. *Journal of Information and Data Management*, 13(1).

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

Mello, R. F., Oliveira, H., Wenceslau, M., Batista, H., Cordeiro, T., Bittencourt, I. I., and Isotanif, S. (2024). Propor'24 competition on automatic essay scoring of portuguese narrative essays. In *Proc. PROPOR 2024, Vol. 2*, pages 1–5.

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M. A., Socher, R., Amatriain, X., and Gao, J. (2024). Large language models: A survey. *ArXiv*, abs/2402.06196.

Pires, G. (2023). Enem 2023: entenda como o exame é capaz de mudar a vida dos estudantes. Accessed at: 7 jun. 2025.

Qin, L., Chen, Q., Feng, X., Wu, Y., Zhang, Y., Li, Y., Li, M., Che, W., and Yu, P. S. (2024). Large language models meet nlp: A survey. *arXiv preprint arXiv:2405.12819*.

Silveira, I. C., Barbosa, A., and Mauá, D. D. (2024). A new benchmark for automatic essay scoring in portuguese. In *Proc. PROPOR 2024, Vol. 1*, pages 228–237.

Xu, H., Gan, W., Qi, Z., Wu, J., and Yu, P. S. (2024). Large language models for education: A survey. *arXiv preprint arXiv:2405.13001*.