Automatic Scoring of Elementary School Essays in Brazilian Portuguese with LLMs: Comparing Gemini, GPT-4o, Claude, and Mistral

Jamilla Lobo[1], Lenon Anthony[2], Andreza Falcão[2], Cleon Xavier[3], Newarney Torrezão[3], Seiji Isotani[4], Ig Ibert[5], Luiz Rodrigues[6], Rafael Mello[1,2]

[1] Centro de Estudos e Sistemas Avançados do Recife (CESAR School)

[2]Universidade Federal Rural de Pernambuco

[3]Instituto Federal Goiano

[4]Universidade de São Paulo

[5]Universidade Federal de Alagoas

[6]Universidade Tecnológica Federal do Paraná

{jsl}@cesar.school

Abstract. *Writing is an essential skill for the development of students' critical thinking, communication, and language competencies. However, evaluating written productions efficiently and within an appropriate timeframe remains a challenge, especially in contexts of high teaching demands. This study investigates the use of Large Language Models (LLMs) for the automated analysis of essays with a focus on achieving accurate and consistent assessments. Four advanced models such as Gemini, GPT-4o, Claude 3.7 and Mistral were examined and applied to the evaluation of narrative texts written in Brazilian Portuguese by elementary school students. The results indicate that the Gemini 2.0 Pro model demonstrated greater accuracy in score assignment, while Claude 3.7 stood out for consistency in alignment with human evaluation. The findings highlight the potential of LLMs to support pedagogical practices by providing consistent assessments and contributing to the development of students' writing skills. The study proposes viable alternatives for the use of artificial intelligence in educational contexts with limited resources.*
Keywords: Automated Essay Scoring · Large Language Models · Artificial Intelligence in Education.

1. Introduction

Writing plays a crucial role in shaping students' understanding of the world, as well as in enhancing their ability to share information, communicate effectively, and improve both language and reading skills [Graham and Harris 2019]. Written essays are widely used in education to assess students' progress, particularly in elementary education, where it is often the first time students engage with complex lexical structures, interpret contextual meanings, and plan their writing [Marrs et al. 2016]. However, the essay evaluation process often fails to provide meaningful and continuous feedback to support students in their learning due to the workload, subjectivity involved in written essays, and the need

DOI: 10.5753/sbie.2025.12161

to adhere to specific criteria that result in a slow and exhausting process for educators [Er et al. 2021].

This complexity of the evaluation process deepens at the intersection between the critical need for formative feedback and the operational challenges of the Brazilian educational system. Pedagogical literature consistently demonstrates that clear and targeted feedback is one of the most influential factors in improving student performance [Aucejo and Wong 2024]. However, due to large class sizes and the increasing intensification of workload demands, providing individualized and continuous support becomes difficult [Gasparini et al. 2022]. This gap between pedagogical demand and practical capacity creates a scenario in which intelligent automation tools, such as LLMs, represent not only a gain in efficiency but also an opportunity to promote educational equity, ensuring that more students receive high-quality guidance to enhance their writing skills.

Automated Essay Scoring (AES) [Page 1966] emerged as a potential solution to the challenges of assessing essays using computers to help grade written texts. Almost 60 years later, with the emergence of artificial intelligence combined with AES systems, models and techniques have evolved to provide high-quality feedback within seconds, as well as scores that are consistent with students' writing [Wang and Wang 2021]. This approach not only optimizes the assessment process for educators but also encourages continuous revision by students, reinforcing the knowledge acquired [Wang and Wang 2021].

AES systems have seen increased use, initially relying on models trained with textual features extracted from student essays. These earlier approaches, while incorporating traditional methods such as the writing approach, focusing on ongoing revision and Computer-Assisted Language Learning (CALL) software [Wang and Wang 2021], and benefiting from advances in machine learning for improved accuracy [Salim et al. 2019], often faced limitations in deep understanding of natural languages and generating contextualized human-like feedback. Consequently, collaborative human-AI models emerged, with AI acting as a decision-support assistant. This landscape has now shifted significantly with the rise of Large Language Models (LLMs), new models that can overcome the performance limitations of prior AI models.

LLMs, which are based on training with massive amounts of text data, have shown strong capabilities for writing support by scoring and generating feedback. In [Liew and Tan 2024], the potential of Open-AI's GPT-3.5 is demonstrated in International English Language Testing System (IELTS) essays, highlighting the model's ability to produce human-like text and assist in understanding the nuances of the instructions. Particularly in AES, [Seßler et al. 2025] explores the performance of various LLMs in scoring essays written by German elementary school students, emphasizing the higher reliability of closed-source models, while also noting that open-source models, when coupled with effective prompt engineering, have the potential to produce better results.

Despite the increasing interest in LLMs for AES, research on their potential in essay from Brazilian elementary school students remains limited. On the one hand, past research has investigated LLMs' performance in AES for Brazilian secondary school students. On the other hand, studies have explored AES for Brazilian elementary school students, but limited to traditional AI models like BERT [Mello et al. 2025]. As a re-

sult, there is an empirical gap in terms of how LLMs perform in scoring essays written in Brazilian Portuguese by elementary school students.

Therefore, this study aims to investigate the performance of state-of-the-art LLMs in evaluating essays written in Brazilian Portuguese by elementary school students. The study compares the scores generated by the LLMs with reference scores assigned by human evaluators. This comparison assesses the accuracy and reliability of LLM-based essay scoring in this educational context, informing practitioners and researchers on the potential of relying on these models to assess Brazilian elementary school students' essays.

2. Related Work

The increasing popularity of LLMs has motivated research on their potential for various tasks, including AES. Given initial insights that these models are more likely to generalize than traditional models like BERT, at least in tasks like Automatic Short Answer Grading [Mello et al. 2025]. Accordingly, this section reviews related work on AES with a focus on studies that either investigate LLMs or concern elementary school essays, providing an overview of previous contributions to date as well as how this paper differs from the state of the art.

To provide a broad context on the application of LLMs in AES and the importance of techniques like prompt engineering, we first consider studies examining standardized tests before narrowing our focus to elementary education. The study by [Liew and Tan 2024] explores GPT-3.5's applicability to IELTS essays scoring, addressing key challenges in AES. Recognizing the utility of AES as a time-saving solution and an efficient grading approach, their study was important for our investigation in gaining a deeper understanding of how this method works. In this regard, the Quadratic Weighted Kappa (QWK) result of 0.68, achieved by them in AES with GPT-3.5, was relevant for our study by illustrating the effectiveness with which AES can operate in that context. Furthermore, [Liew and Tan 2024] emphasis on prompt engineering to optimize performance in AES reinforced the relevance of our own methodological approach for elementary school competencies.

Building on the understanding of LLM capabilities in broader AES tasks, the challenges, and specificities of applying these models to younger learners become pertinent. A direct examination of LLMs in an elementary school context is offered by [Seßler et al. 2025]. AES using LLMs has gained prominence, particularly to support educators in grading school texts. In this study, a comparison was made between closed-source models such as GPT-3.5, GPT-4, and o1-preview, and open-source models like LLaMA 3 and Mixtral, in evaluating essays written by German elementary school students. The authors found that closed-source models demonstrated greater consistency and alignment with human evaluations, particularly in linguistic aspects such as literal speech and spelling, with GPT-o1 achieving overall result of Spearman's r = 0.74, while open-source models exhibited lower variability and accuracy. The study highlights limitations of LLMs in analyzing complex content-related aspects, such as coherence and textual logic. This analysis is crucial for the current study in selecting models for comparison and identifying areas where LLM performance can be enhanced in the context of Brazilian Portuguese and elementary education.

Given the identified limitations and the desire to improve LLM performance without extensive retraining, research has explored methods like incorporating linguistic features into prompts, as demonstrated by [Hou et al. 2025] presented a hybrid model to enhance the AES with LLMs, incorporating linguistic features directly into the prompts, such as the number of unique words, use of connectives, and complex vocabulary. The study showed that the inclusion of these features improved the performance of the models, particularly with Mistral-7B, especially in out-of-distribution data. Although the supervised BERT model showed the best accuracy with average result of QWK equal to 0.545, the proposed approach offers a practical and scalable solution, without the need for extensive training or adjustments. This methodology of enriching prompts with linguistic features inspires our research, which explores the use of LLMs like Gemini, GPT-4o, and Claude to evaluate essays written by Brazilian elementary school students, aiming to improve the accuracy and adaptation of the models to Brazilian Portuguese. Like in the study by Hou et al. [Hou et al. 2025], our work aims to offer an efficient and easily adaptable solution for automated essay grading, without the need for complex fine-tuning, differing by the focus on elementary school essays written in Brazilian Portuguese.

In addition to overall scoring, LLMs are also being explored for more detailed analyses of textual components in student writing. [Ferreira Mello et al. 2025], for example, investigated the use of these models to detect narrative elements, a task relevant to many writing activities in elementary and high school education. As part of that investigation, they developed a system for the automatic detection of narrative component elements in essays written by middle school students, focusing on identifying structural categories such as narrator, character, and inciting event, using a combination of Random Forest with TF-IDF and BERT embeddings, along with the LLM model GPT-4 Turbo. The study showed that the BERT and Random Forest combination proved most effective across most categories, exemplified by BERT's Mean Absolute Error (MAE) of 0.64, which was the lowest MAE for generalizability. Meanwhile, GPT-4 Turbo, even with a simple prompt, also yielded strong results, although it encountered challenges in more complex categories. The research contributes to the development of automated methods for essay grading, offering a practical solution for identifying narrative structures, with an emphasis on the use of LLMs in this process. The work of Mello et al. inspires our research, as it also uses LLMs for essay evaluation, but with a focus on the detection of textual elements, whereas we focus on scoring competencies of Brazilian elementary education from the essay.

The adaptation of LLMs and prompt engineering techniques to specific national educational contexts is a critical step, particularly for languages other than English. [da Silva and de Araujo 2025]. address this by applying LLMs to the Brazilian high-stakes testing scenario. [da Silva and de Araujo 2025] proposed a system based on GPT-4o-mini for automated evaluation of ENEM (Exame Nacional do Ensino Médio) essays. They employed a chain prompt engineering technique that enables the separate scoring of competencies outlined in official criteria while providing interpretable, pedagogical feedback. The method's robustness, requiring no extensive fine-tuning, demonstrated the feasibility of LLMs for the Brazilian context, potentially enhancing the scalability and accessibility of evaluations. In our case, given that elementary education competencies differ from those assessed in ENEM, we adopted the approach of constructing more comprehensive and specific prompts tailored to the unique characteristics of elementary

education. This adaptation aims to ensure precision and efficiency in the automated assessment of student essays at this educational stage.

The study conducted by [Filho et al. 2023] has made an important contribution to the development and understanding of our research. The main reason for this connection is that both studies use the exact same dataset, which consists of narrative essays written in Brazilian Portuguese by elementary school students. Their work was crucial for understanding the dataset itself and for identifying existing approaches for certain competencies, such as Formal Register. This foundation allowed us to explore how these insights could be replicated and extended to other elementary school essay competencies. Thus, it becomes easier to highlight the differences in our approach, which uses LLMs, differing from the traditional techniques employed by Filho and his team. Their work serves as a reliable benchmark for this dataset, against which we can demonstrate the advantages and details of our LLM-based method.

Finally, the PROPOR'24 Competition on Automatic Essay Scoring of Portuguese Narrative Essays [Mello et al. 2024b] stands as a key reference for this paper. This competition utilized the exact same dataset of narrative essays in Brazilian Portuguese as our study to compare pre-trained language models such as Albertina and BERTimbau. Their results demonstrated fair to moderate Kappa agreement with human evaluations. Specific examples include INESC-ID's Kappa of 0.548 for Thematic Coherence and PiLN's 0.414 for Formal Register. This highlights the growing notability of this research area and positions our work as providing valuable comparative results to serve as a basis for future studies.

2.1. Research Question

Drawing upon existing literature and recent advancements in LLMs, this paper investigates the efficacy of employing LLMs to assess narrative texts written by elementary school students in Portuguese. Specifically, the research examines how effectively these advanced computational tools can evaluate and score student narratives, contributing to more reliable, consistent, and scalable educational assessment methods in primary education contexts.To achieve this, we will investigate the following research questions:

> **RESEARCH QUESTION 1 (RQ1):**
> *How accurate are LLMs in scoring Brazilian elementary school essays?*
>
> **RESEARCH QUESTION 2 (RQ2):**
> *What is the agreement between the actual essay scores and those assigned by the LLMs?*

The resulting analysis can guide writing evaluators in optimizing their time and workload, making the integration of artificial intelligence into manual essay evaluation more tangible.

3. Method

As shown in Figure 1, the pipeline for this study consists of two main stages. First, selected dataset transcriptions will be tested using chosen LLMs with a custom prompt adapted for Brazilian elementary school essay scoring. The second stage involves comparing and interpreting the results with the dataset score references, then extracting relevant information for future work.

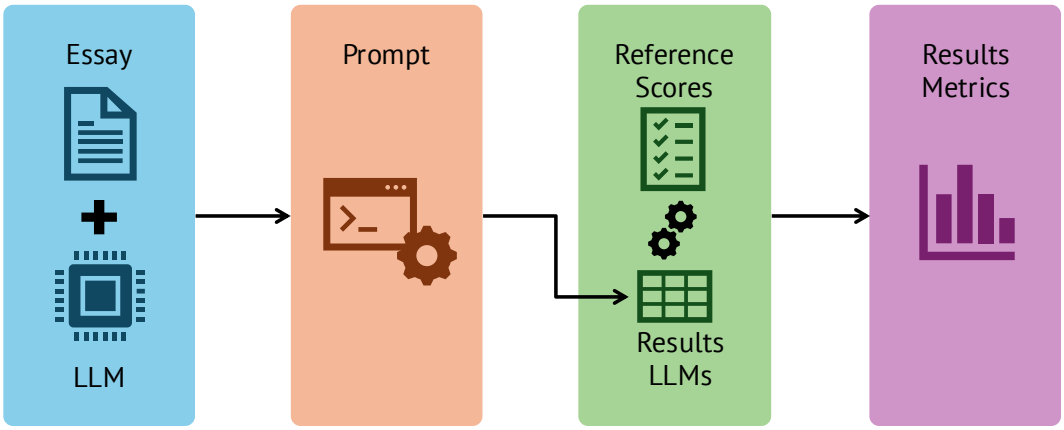

Figure 1. Pipeline diagram used in the proposed system.

3.1. Dataset

The dataset used in this study is the Brazilian Portuguese Narrative Essays, which consists of 1,235 essays written in Portuguese by fifth-grade students from elementary schools in Brazil. Introduced by [Mello et al. 2024a], this reference dataset was the subject of related work [Ferreira Mello et al. 2025, da Silva Filho et al. 2023], as well as a competition in AES for Brazilian Portuguese, highlighting its adequacy for our research purposes.

Following Brazilian guidelines, the dataset's narrative-style essays were evaluated into four competencies: (C1) Mastery of formal written Portuguese; (C2) Knowledge of linguistic mechanisms to build coherent arguments; (C3) Understanding of the structure of the essay and the integration of interdisciplinary knowledge to develop a narrative in prose; and (C4) Selection, organization, and interpretation of information, facts, opinions, and arguments to support a point of view [Mello et al. 2024b]. Each competency was scored on a scale from 1 (lowest) to 5 (highest), allowing a maximum total score of 20 points per essay.

The construction process involved three steps: data collection, annotation, and discrepancy resolution. Teachers photographed student essays, which were then transcribed. Two linguistically trained teachers independently scored essays across competencies over twelve weeks. A third, more experienced evaluator was included to resolve discrepancies between the initial two annotators. Although the dataset was originally split into training, validation, and test sets, we opted to use all available essays to obtain more comprehensive results regarding the overall score distribution shown in Table 1.

Additionally, the dataset includes special formatting tokens that reflect structural and handwriting features observed in the original handwritten essays. These tokens provide metadata about paragraph breaks, erasures, symbols, and other formatting peculiarities. Table 2 summarizes the meaning of each token category.

These tokens were preserved in the digital version of the essays to maintain fidelity to the original handwritten documents, enabling analyses that consider both linguistic content and formatting patterns. This feature is particularly relevant for studies investigating writing mechanics, revision behaviors, or stylistic coherence in early literacy development. Next, we discuss how we handled these essays in our analysis.

Table 1. Essays distribution by proficiency level / score for each evaluated competency.

Score	C1 - Formal Register	C2 - Thematic Coherence	C3 - Textual Typology	C4 - Cohesion
1	42 (3.40%)	356 (28.83%)	31 (2.51%)	40 (3.24%)
2	188 (15.22%)	291 (23.56%)	29 (2.35%)	187 (15.14%)
3	803 (65.02%)	513 (41.54%)	198 (16.03%)	824 (66.72%)
4	185 (14.98%)	63 (5.10%)	751 (60.81%)	157 (12.71%)
5	17 (1.38%)	12 (0.97%)	226 (18.30%)	27 (2.19%)

Table 2. Flags/Tokens and their meanings in handwritten essay analysis

Flag/Token	Meaning
[P], [P], {P}, [p], {p}	Presence of a new paragraph in that spot
[S], [s]	This token was a symbol in the handwritten essay
[T], [t], {t}	The following tokens represent the title of the essay
[R], [X], [X˜], [r], [x], {x}	This token was an erasure in the handwritten essay
[?], {?}, [?}, {?]	Unknown token in the handwritten essay
[LC], [LT], [lt]	The following tokens didn't follow a straight line in the essay

3.2. LLM Models and Prompt

This study employed four LLMs to evaluate narrative essays authored by elementary school students (grades 5 through 9). To ensure a representative sample of the current LLM landscape, our selection features proprietary models from prominent AI companies. This approach aligns with literature recommendations, which stress the importance of comparing different model families in the evaluation of educational solutions [Singh et al. 2024]. Our evaluation framework specifically incorporates four distinct LLM implementations:

- **OpenAI GPT-4o**[1]: the selection was based on previous research on its application in AES, allowing a direct comparison of our results with the literature. The 4o model variant was chosen because it represented, at the time of the experiments, the most cost-optimized and high-performing multimodal model available.
- **Google Gemini 2.0 Flash**[2]: included as a light counterpoint (flash), it allows large test scaling. It delivers next-gen features, with a wide range of superior speed inferences, multimodal resources, and a 1M token context window.
- **Anthropic Claude 3.7 Sonnet**[3]: its capabilities include an extended reasoning model, the ability to adhere to instructions, and robust graduate-level reasoning, complemented by significant extended thinking.
- **Mistral Large**[4]: we opted for Mistral due to its single-node inference system and its multilingual training, which is beneficial for understanding Portuguese words. Its proficiency in handling precise instructions was another key factor.

[1]https://platform.openai.com/docs/models
[2]https://cloud.google.com/vertex-ai/generative-ai/docs/models
[3]https://www.anthropic.com/news/claude-3-7-sonnet
[4]https://mistral.ai/news/mistral-large

They have facilitating support from the LangChain framework [Chase 2023], which offers consistent interfaces and abstractions between various model providers. In line with best practices for assessment tasks, as outlined by [Liu et al. 2023], all models were configured with minimal temperature settings. This configuration prioritizes consistency and reliability in the evaluation process over the generation of creative or varied responses.

Prompt design is crucial for eliciting consistent and accurate essay evaluations [Ferreira Mello et al. 2025, da Silva and de Araujo 2025]. Accordingly, our prompt was structured to ensure models would:

1. Assume the role of an experienced teacher evaluating elementary school narrative essays
2. Evaluate essays based on four specific competencies (C1-C4)
3. Provide numerical scores (0-5) for each competency
4. Consider both the essay text and its corresponding prompt text

As a result, the prompt includes detailed rubrics for each competency:

- **C1**: Formal Register (Orthography, Grammar, and Punctuation)
- **C2**: Thematic Coherence (Theme Maintenance and Textual Progression)
- **C3**: Textual Typology (Narrative Structure)
- **C4**: Cohesion (Idea Sequencing)

Each competency's rubric was designed with five performance levels, providing clear criteria for scoring. This approach follows the recommendations of [Singh et al. 2024], who suggest that explicit scoring criteria enhance the reliability of LLM evaluations in educational contexts.

The prompt also includes specific instructions for handling technical tokens that might be present in digitized essays, ensuring models would focus only on the actual content written by students. We directed models to output only the scores in a structured format (e.g., "c1: 4") without explanations, to facilitate consistent parsing and analysis of results.

Our implementation includes handling for different model types, addressing the unique requirements of each LLM architecture. We leverage LangChain's structured output capabilities for all models, ensuring a standardized evaluation format across our entire analysis. Based on that context, our final prompt was the one shown in Table 3:

Table 3. Prompt Elements for the Assessment Task Translated to English

Element	English text
Instruction	Evaluate the essay based on the prompt text, assigning scores from 0 to 5 for each competence, according to the Competency-Based Correction Matrix.
Criteria	C1 - Formal Register (Spelling, Grammar, and Punctuation); C2 - Thematic Coherence (Maintaining the Theme and Textual Progression); C3 - Text Typology (Narrative Structure); C4 - Cohesion (Chaining of Ideas). *Important to notice that the full description of each competency used in this prompt is in the Mello et al. [Mello et al. 2024b] study*
Additional Instruction	Completely ignore the technical tokens present in the essay. They are not part of the actual content and were added by whoever digitized the text. Tokens to ignore include: *tokens shown in Table 2*. Evaluate only the actual content of the essay. Ignore spacing, line breaks, or incorrect formatting. Compare the content of the essay with the prompt text. Do not generate explanations, only the scores.
Role	You are a teacher correcting essays written by elementary school students (5th to 9th grade).
Output	Respond only with the scores, following the format: c1: [score 0 to 5], c2: [score 0 to 5], c3: [score 0 to 5], c4: [score 0 to 5]. Essay to be evaluated: {essay_text} Prompt text: {prompt_text}

4. Results

Table 4 shows the results of exact match accuracy in the essay evaluations (RQ1). We observe that the Gemini 2.0 flash model achieved the best results in competencies C1 (41.62%), C3 (31.26%) e C4 (43.16%). Additionally, Claude Sonnet 3.7 demonstrated higher precision in competency C2 (47.37%). On the other hand, the Mistral model presented the lowest exact match rates across all evaluated dimensions.

Table 4. LLMs' exact match accuracy in assessing Brazilian Portuguese essays by competence (C1 to C4).

Models	C1	C2	C3	C4
Gemini 2.0 flash	**41.62%**	32.79%	**31.26%**	**43.16%**
Mistral Large	28.26%	29.15%	18.62%	33.44%
Claude Sonnet 3.7	33.52%	**47.37%**	30.61%	26.40%
GPT-4o	28.42%	46.07%	19.35%	37.09%

In terms of F1-score and QWK in Table 5 (RQ2), the Claude Sonnet 3.7 model stood out, showing the highest values in several competencies, notably in C2 (F1-score

= 0.5117; QWK = 0.6228) and C3 (F1-score = 0.3432; QWK = 0.3766). Interestingly, although Gemini 2.0 flash showed high exact match precision in some competencies, its QWK values were low or negative, indicating that when Gemini misses a score, it tends to be a substantial error compared to other LLMs.

Table 5. F1-score and QWK by LLM and competency

Model	F1-score				QWK			
	C1	C2	C3	C4	C1	C2	C3	C4
Gemini	**0.4347**	0.2776	0.3330	**0.4491**	-0.0083	0.0225	-0.0172	0.0065
Mistral	0.3169	0.2983	0.2128	0.3763	-0.0166	0.0504	-0.0023	0.0176
Claude	0.3722	**0.5117**	**0.3432**	0.3098	**0.3811**	**0.6228**	**0.3766**	0.3043
GPT-4o	0.3100	0.4693	0.2190	0.4112	0.3057	0.3972	0.2610	**0.3534**

The analysis of evaluation errors, through RMSE and MAE, is presented in Table 6. It reveals that the Claude Sonnet 3.7 model tended to present the lowest errors across all competencies, indicating closer proximity between its scores and the reference ones (e.g., for C2, an RMSE of 0.9012 and a MAE of 0.6146). In contrast, the Mistral model generally exhibited the highest errors, ranging from 1.1975 to 1.6777 and 0.8996 to 1.3733 for RMSE and MAE, respectively.

Table 6. RMSE and MAE by LLM and competency

Model	RMSE				MAE			
	C1	C2	C3	C4	C1	C2	C3	C4
Gemini	1.0643	1.3768	1.3501	1.0772	0.7522	1.0405	1.0097	0.7457
Mistral	1.3049	1.3949	1.6777	1.1975	1.0146	1.0794	1.3733	0.8996
Claude	0.9939	0.9012	1.1770	1.1481	0.7692	0.6146	0.9126	0.9231
GPT-4o	1.0906	0.9828	1.3513	0.9902	0.8688	0.6680	1.1296	0.7425

5. Discussion

The findings of this study reveal the capacity of different LLMs to perform AES in narrative essays written by elementary school students in Brazilian Portuguese. The variability in performance among the models and the evaluated competencies highlights the complexity of replicating human judgment in writing assessment. This insight leads to important implications for research and practice in AES, which are discussed next.

Gemini 2.0 flash exhibited high exact match accuracy in some competencies, but its low weighted agreement raises questions about the consistency of this precision (RQ1). Claude Sonnet 3.7 demonstrated the highest overall accuracy, suggesting a better ability to capture nuances in writing quality. GPT-4o showed reasonable accuracy, with lower average errors compared to reference scores. In contrast, Mistral Large showed the lowest overall accuracy among the models evaluated for this specific task. The comparative analysis indicates that the accuracy of the LLMs varies significantly depending on the model and the metric considered. This highlights that model selection and context are crucial, as what performs well in one setting may not directly translate to another, emphasizing the need for adjusted approaches to meet the right linguistic and pedagogical demands of specific educational levels.

The agreement between the actual essay scores and those assigned by the LLMs (RQ2) also varied among the models. Claude Sonnet 3.7 presented the highest agreement with human evaluations, as indicated by its high QWK values. Gemini 2.0 flash, despite its high exact match in some areas, demonstrated lower weighted agreement (QWK), suggesting that when discrepancies occur, the scores tend to deviate more from human evaluations. This low weighted agreement, despite high exact match, emphasizes the importance of consistency of precision, indicating a critical area for future research to develop more robust models or refine evaluation metrics that better capture subtle deviations from human scores. GPT-4o showed competitive agreement, complemented by its low average errors. When comparing our results with [Liew and Tan 2024], we find that the GPT-4 from the related study, showed QWK values close to those of our GPT-4o results. In some cases, as previously mentioned, our Gemini models even performed better agreement. On the other hand, Mistral Large exhibited the lowest overall agreement with human evaluations. These findings underscore that the alignment between LLM assessments and human ratings heavily depends on the specific model, with models like Claude and Gemini showing considerable potential for this assessments.

This study significantly contributes to the literature by providing a comparative analysis of state-of-the-art LLMs in evaluating narrative essays written by Brazilian Portuguese elementary school students. This is a specific context that presents distinct linguistic and pedagogical challenges. By employing a diverse set of metrics, including those focused on agreement (exact and weighted) and scalar error, we offer a richer, more detailed perspective than studies relying on a single metric. However, the use of a single and potentially unbalanced dataset is a limitation that may have influenced our results, thereby limiting the generalizability of the models' true performance and their ability to predict outcomes on new, unseen data. Our findings advance the understanding of each model's capabilities and limitations, underscoring that the choice of LLM and evaluation metrics directly influences how their performance is interpreted.

It is important to acknowledge other limitation of this study. The decision to adopt complexity of replicating human judgment a zero-shot approach, without specific prompt or hyperparameter optimization for each LLM, while allowing for a comparison under more direct conditions, may not have explored the full potential of each model. Therefore, future research should explore the impact of different prompting and fine-tuning strategies on LLM performance in AES, tailoring prompts to specific student developmental stages. Furthermore, the single evaluation of each essay per LLM does not capture the potential variability in their output. Based on these limitations, future research could benefit from using multiple datasets to validate the generalization of the results, exploring the impact of different prompting and fine-tuning strategies, and analyzing the stability of LLM evaluations through multiple inferences per essay.

6. Conclusion

This study evaluated the capability of state-of-the-art LLMs to automatically score elementary school narrative essays in Brazilian Portuguese. Our findings show that while these models show promise for reducing teacher workload and providing timely student feedback, their performance varies significantly across different metrics and specific competencies. In particular, Claude Sonnet 3.7 exhibited the strongest agreement with human evaluators, suggesting its potential for reliable automated grading, while Gemini 2.0 flash

showed high accuracy in specific essay competencies.

The primary implication of this research is that the choice of both the LLM and the evaluation metrics is critical for determining a model's suitability for automated essay scoring tasks. This work advances the literature by providing a detailed empirical analysis of state-of-the-art models in a specific educational context, highlighting the importance of a multifaceted approach to assessing their capabilities and limitations. Concretely, it contributes by establishing comparative baselines for four LLMs applied to essays in Brazilian Portuguese, aligning the evaluation with a rubric that reflects classroom practice and reporting performance at the competence level through multiple metrics. Furthermore, it provides practical guidance on model selection according to different pedagogical priorities, such as the need for greater consistency or higher accuracy in specific competencies. These results support practitioners in choosing configurations that balance reliability and coverage in real school contexts.

6.1. Future Work

Future research should explore the application of these models to different text genres and educational levels, expand the diversity of datasets to ensure greater representativeness, investigate fine-tuning strategies based on national educational rubrics, and analyze the stability of evaluations through multiple runs. It is also relevant to study aspects of equity, cost, and the integration of LLMs into pedagogical workflows that combine automated scoring with formative feedback.

Acknowledgments

We acknowledge the use of Generative AI to help in drafting and revising this paper, which was revised accordingly, and take full responsibility for its content.

Ethical Concerns

This study relies on a previous collected / published dataset. Accordingly, it is not eligible for assessment by an Institutional Review Board.

Artifacts Availability

This study's artifacts will be made available from the corresponding authors.

References

Aucejo, E. M. and Wong, K. (2024). The effect of feedback on student performance. *Journal of Public Economics*, 224:105274.

Chase, H. (2023). Langchain. `https://github.com/langchain-ai/langchain`.

da Silva, W. A. and de Araujo, C. C. (2025). Automated enem essay scoring and feedbacks: A prompt-driven llm approach. In *Proceedings of the ... (complete as appropriate)*, Recife, Brazil.

da Silva Filho, M. W., Nascimento, A. C., Miranda, P., Rodrigues, L., Cordeiro, T., Isotani, S., Bittencourt, I. I., and Mello, R. F. (2023). Automated formal register

scoring of student narrative essays written in portuguese. In *Workshop de Aplicações Práticas de Learning Analytics em Instituições de Ensino no Brasil (WAPLA)*, pages 1–11. SBC.

Er, E., Dimitriadis, Y., and Gašević, D. (2021). Collaborative peer feedback and learning analytics: Theory-oriented design for supporting class-wide interventions. *Assessment & Evaluation in Higher Education*, 46(2):169–190.

Ferreira Mello, R., Pereira Junior, C., Rodrigues, L., Pereira, F. D., Cabral, L., Costa, N., Ramalho, G., and Gasevic, D. (2025). Automatic short answer grading in the llm era: Does gpt-4 with prompt engineering beat traditional models? page 93–103.

Filho, M. W. d. S., Nascimento, A. C. A., Miranda, P., Rodrigues, L., Cordeiro, T., Isotani, S., Bittencourt, I. I., and Mello, R. F. (2023). Automated formal register scoring of student narrative essays written in portuguese. In *Anais do Congresso Brasileiro de Informática na Educação (CBIE)*. The paper mentions CBIE 29[cite: 1]. Specific page numbers would be found in the full proceedings.

Gasparini, S. M., Barreto, S. M., and Assunção, A. A. (2022). O professor, as condições de trabalho e os efeitos sobre sua saúde. *Educação & Pesquisa*, 48(2):e242423.

Graham, S. and Harris, K. R. (2019). Evidence-based practices in writing. In *Best Practices in Writing Instruction*.

Hou, Z. J., Ciuba, A., and Li, X. L. (2025). Improve llm-based automatic essay scoring with linguistic features. *arXiv preprint arXiv:2404.19064*. Available at `https://github.com/JoeyHou/essay_eval`.

Liew, P. Y. and Tan, I. K. T. (2024). On automated essay grading using large language models. In *Proceedings of the 2024 8th International Conference on Computer Science and Artificial Intelligence (CSAI)*, page 8, Beijing, China. ACM.

Liu, H.-C., Wang, C., Keefer, M. W., Kim, S., Glaser, K., van der Wegen, R., and Rus, V. (2023). Evaluating LLMs for grading undergraduate student essays. *arXiv preprint arXiv:2502.09497*.

Marrs, S. et al. (2016). Exploring elementary student perceptions of writing feedback. *Journal on Educational Psychology*, 10(1):16–28.

Mello, R. F., de Oliveira, H. T. A., Wenceslau, M., Batista, H., Cordeiro, T., Bittencourt, I. I., and Isotani, S. (2024a). Brazilian portuguese narrative essays dataset. Accessed: May 15, 2025.

Mello, R. F., Oliveira, H., Wenceslau, M., Batista, H., Cordeiro, T., Bittencourt, I. I., and Isotanif, S. (2024b). 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.

Mello, R. F., Rodrigues, L., Sousa, E., Batista, H., Lins, M., Nascimento, A., and Gasevic, D. (2025). Automatic detection of narrative rhetorical categories and elements on middle school written essays.

Page, E. B. (1966). The imminence of grading essays by computer. *The Phi Delta Kappan*, 47(5):238–243.

Salim, Y., Stevanus, V., Barlian, E., Sari, A. C., and Suhartono, D. (2019). Automated english digital essay grader using machine learning. In *2019 IEEE International Conference on Engineering, Technology and Education (TALE)*, pages 1–6. IEEE.

Seßler, K., Fürstenberg, M., Bühler, B., and Kasneci, E. (2025). Can ai grade your essays? a comparative analysis of large language models and teacher ratings in multidimensional essay scoring. page 462–472.

Singh, A., Tan, D., Hepworth, C., and Seeland, M. (2024). Comparing LLM responses for education across model families. *arXiv preprint arXiv:2502.08450*.

Wang, D. and Wang, J. (2021). The impact mechanism of aes on improving english writing achievement. *Journal Unknown*.