# Analysis of the Effectiveness of LLMs in Handwritten Essay Recognition and Assessment

Daisy Cristine Albuquerque da Silva, Carlos Luiz Ferreira, Sérgio dos Santos Cardoso Silva and Juliano Bruno de Almeida Cardoso

*Military Engineering Institute (IME), Praça General Tibúrcio, 80, Urca, Rio de Janeiro, RJ, Brazil*

Keywords: Large Language Models, Automated Essay Scoring, Learning Analytics, Education, Handwritten Texts.

Abstract: This study investigates the application of Large Language Models (LLMs) for handwritten essay recognition and evaluation within the Military Institute of Engineering (IME) selection process. Utilizing a two-stage methodology, 100 handwritten essays were transcribed using LLMs and subsequently evaluated against predefined linguistic and content criteria by both open-source and closed-source LLMs, including GPT-3.5, GPT-4, o1, LLaMA, and Mixtral. The evaluations were compared to those conducted by IME professors to assess reliability, alignment, and limitations. Results indicate that closed-source models like o1 demonstrated strong reliability and alignment with human evaluations, particularly in language-related criteria, though they exhibited a tendency to assign higher scores overall. In contrast, open-source models displayed weaker correlations and lower variance, limiting their effectiveness for nuanced assessment tasks. The study highlights the potential of LLMs as complementary tools for automated essay evaluation while identifying challenges such as variability in human and model evaluations, the need for advanced prompt engineering, and the necessity of incorporating diverse essay formats for improved generalizability. These findings provide insights into optimizing LLM performance in educational contexts.

## 1 INTRODUCTION

The evaluation of handwritten texts is an indispensable yet challenging step in selection processes where essays play a central role. At the Military Institute of Engineering (IME), essays constitute the second phase of the selection exam, following an initial stage comprising objective and open-ended questions in mathematics, physics, and chemistry. Given the large number of candidates, the correction of essays becomes a significant logistical and operational challenge, often requiring between 4 to 6 months to complete.

Despite the considerable time and effort invested by evaluators, the feedback provided to candidates is often delayed and highly variable. Like other institutions, IME faces difficulties related to the scarcity of qualified human resources to efficiently perform this task, highlighting the need for technological solutions to ease workload.

Recent advances in Artificial Intelligence (AI) and Large Language Models (LLMs), such as GPT-4 (OpenAI, 2024a) and OpenAI's o1 model (OpenAI, 2024b), Meta's LLaMA 3 (AI@Meta, 2024), and Mistral AI (Jiang et al., 2024), open up new possibilities in the educational context (Kasneci et al., 2024). These models have a high proficiency in processing, analyzing and generating natural language, offering significant potential to improve teaching and assessment processes. For example, LLMs can assist in lesson preparation through automated question generation (Bhat et al., 2022), facilitate teacher collaboration via conversational AI tools (Ji et al., 2023), and support the correction and feedback generation for student texts (Bewersdorff et al., 2023).

In addition, LLM applications span various educational domains, including language learning (Muñoz et al., 2023), mathematics (Nguyen et al., 2023), and life sciences (Bewersdorff et al., 2024). Integrating these tools can optimize teachers' time management, allowing them to focus on other critical activities while simultaneously enhancing the quality and consistency of assessments, fostering more personalized and interactive learning experiences.

Given these opportunities, the question arises: to what extent can LLMs serve as viable alternatives or, at least, complementary tools in the evaluation of handwritten essays? Although AI-generated feed-



Albuquerque da Silva, D. C., Ferreira, C. L., Silva, S. S. C. and Cardoso, J. B. A.
Analysis of the Effectiveness of LLMs in Handwritten Essay Recognition and Assessment.
DOI: 10.5220/0013353700003890

In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART 2025) - Volume 2*, pages 776-785
ISBN: 978-989-758-737-5; ISSN: 2184-433X

back is increasingly being incorporated into educational applications, significant gaps remain in understanding its quality and effectiveness. Some initial approaches use AI to provide feedback on student texts (Seßler et al., 2023), while others focus on Automated Essay Scoring (AES) systems (Ramesh and Sanampudi, 2022). However, detailed assessments based on specific criteria remain scarce.

Previous research, often focusing on holistic scores (Sawatzki et al., 2021) or a limited set of general criteria (Mizumoto and Eguchi, 2023; Naismith et al., 2023), often fails to fully capture the complexity and nuances of student texts, particularly in the IME selection process, where essays are assessed with technical and narrative/discursive rigor. Additionally, the lack of adequate data resources and the absence of clear and consistent standards—often based on subjective evaluations by teachers—pose additional challenges.

To explore the potential of LLMs in the IME selection process, this study aims to address these gaps by evaluating the effectiveness of open-source and closed-source models in the transcription and analysis of candidates' essays based on predefined content and linguistic criteria. The study compares human evaluations with those generated by models such as GPT-3.5, GPT-4 (OpenAI, 2024a), o1 (OpenAI, 2024b), LLaMA 3-70B (AI@Meta, 2024), and Mixtral 8x7B (Jiang et al., 2024), thoroughly investigating their performance and limitations in different categories of evaluation.

The central objective of this study is to understand how well LLMs aligns with teacher evaluations and to identify areas where these models excel or require improvement. The main research questions include:

- **RQ1.** How reliably do open and closed LLM models perform in essay evaluation?
- **RQ2.** How do LLM evaluations correlate with those conducted by IME teachers?
- **RQ3.** What are the limitations of using LLMs to assess qualitative aspects of essays beyond providing a basic holistic score?

To address these research questions, the study was divided into two distinct stages.

The first stage involved the transcription of 100 handwritten essays collected from candidates who participated in the IME admission process. For this purpose, the essays were digitized in high resolution and manually transcribed using LLMs. The effectiveness of the process was evaluated by comparing the transcribed texts with the original handwritten versions, using one of the predefined evaluation categories, the Presentation category. This stage aimed to assess the model's ability to handle variations in legibility present in the handwritten essays.

In the second stage, the transcribed texts were evaluated by LLMs based on a predefined rubric that included criteria like theme, types of text, structure of the text, argumentative strength and coehrence, cohesion and grammatical structure. The results of these automated evaluations were then compared to the analyses performed by IME faculty members, using correlation metrics such as Spearman's r to measure the alignment between the methods. Additionally, qualitative analyses of discrepancies between teacher and automated evaluations were conducted to identify limitations and potential biases in the models, contributing to a better understanding of their performance in text assessment tasks.

This study evaluated the performance of open-source and closed-source LLMs in analyzing essays from candidates in the IME selection process using predefined criteria, comparing their results to teacher evaluations. The findings highlight that the closed-source o1 model demonstrated strong alignment with human assessments, particularly in language-related criteria, but consistently assigned higher overall scores. In contrast, open-source models like LLaMA and Mixtral showed limited effectiveness due to weak correlations with human evaluations.

The variability in both LLM and human evaluations underscores the need for robust aggregation mechanisms and the integration of subjective factors into LLM training frameworks. Despite these challenges, advances in closed-source models, particularly OpenAI, demonstrate increasing reliability and potential for use in educational contexts.

## 2 RELATED WORK

### 2.1 LLM for Handwritten Text Recognition

**Handwritten Text Recognition in Portuguese.** The ICDAR 2024 Competition on Handwritten Text Recognition in Brazilian Essays – BRESSAY aimed to advance handwritten text recognition (HTR) in Brazilian academic essays, challenging participants to handle diverse handwriting styles and irregularities such as smudges and erasures (Neto et al., 2024a). The competition, featuring 14 participants from various countries, utilized the BRESSAY dataset, which comprises 1,000 handwritten pages in Brazilian Portuguese. The challenges were structured across three

levels: line, paragraph, and page recognition, evaluated using the metrics Character Error Rate (CER) and Word Error Rate (WER). The best-performing submissions achieved CERs of 2.88% for line-level recognition, demonstrating the effectiveness of deep learning models and preprocessing techniques, as highlighted by (Gatos et al., 2014) and (Neto et al., 2024b). This study underscores the importance of the BRESSAY dataset as a benchmark for future HTR research, particularly in addressing real-world challenges of handwritten texts in educational contexts.

Recent research has made significant strides in the field of Handwritten Text Recognition (HTR) by leveraging advanced machine learning models and hybrid techniques. Early approaches, such as combining Hidden Markov Models (HMM) with Artificial Neural Networks (ANN) (Graves et al., 2009), demonstrated the potential for hybrid architectures. Building on these foundations, novel frameworks like Gated-CNN-BGRU have been introduced to enhance Handwritten Digit String Recognition (HDSR), particularly in noisy environments with limited training data (LeCun et al., 1998; Gatos et al., ). Efforts have also extended to specific applications, such as the automatic detection and summarization of handwritten content on whiteboards (Breuel, 2005), and robust CNN-based approaches for recognizing text in handwritten notes and whiteboard images (Wang and Li, 2020). Moreover, researchers have explored handwritten character recognition using neural networks (Bluche et al., 2014), aiming to transform handwritten or printed documents, such as doctors' notes, into digital formats for better analysis and accessibility (Bishop, 2006).

Further advancements include segmentation of cursive handwritten words using methods like the Kaiser window to address challenges in preprocessing and word segmentation (Doermann and Tombre, 2014). Specialized applications, such as Smart RE frameworks for capturing workshop notes (Rice, 1999) and Bank Cheque Handwritten Text Recognition (BCHWTR) systems for Indian cheques (Graves et al., 2006), highlight the versatility of HTR. Other studies have delved into the use of Multi-Layer Perceptrons (MLP) and Deep Convolutional Networks (CNN) for handwritten digit recognition (Graves et al., 2008), and convolutional architectures combined with Long Short Term Memory (LSTM) for improved text-to-digital conversions (Koutník et al., 2014). Cutting-edge innovations, including deformable convolutions for accommodating diverse writing styles and 2D Self-Organized Neural Networks (ONNs) for enhancing accuracy, have demonstrated significant reductions in Character Error Rate (CER) and Word Error Rate (WER) across datasets such as IAM English (Bowman et al., 2016; Jaderberg et al., 2014). These advancements collectively underline the growing potential of modern HTR systems to tackle real-world challenges with precision and adaptability.

## 2.2 LLM for Text Assessment

**Applying Open-Source LLMs to Essay Data.** Recent studies have explored the potential of open-source LLMs for generating feedback on essays. (Stahl et al., 2024a) evaluated various prompting strategies, such as zero- and few-shot learning, to determine the effectiveness of Mistral 7B (Jiang et al., 2023) in generating feedback for essays. Although this approach appeared promising, the overall impact of automated essay scoring (AES) on feedback quality was minimal. The study highlighted that combining AES with feedback generation could enhance scoring performance but emphasized the risks of relying on LLMs to evaluate feedback from another LLM. Such practices could perpetuate model biases and lack the nuanced understanding that human experts, like teachers, provide. Additionally, the study omitted critical information about the qualifications of the 12 human raters, raising concerns about the reliability of their feedback assessments. This underscores the need to compare LLM-generated feedback with feedback from qualified human experts.

**Traditional Automated Essay Scoring (AES).** Automated Essay Scoring systems have been evolving since 1966 (Ramesh and Sanampudi, 2022). Earlier approaches relied on statistical features to analyze text (Ke and Ng, 2019). With the advent of deep learning, methods like LSTMs and Transformer-based models (e.g., BERT) enabled more advanced syntactic and semantic analysis (Devlin et al., 2019). For instance, BERT has been used to extract features for regression models, output class labels (Doewes and Pechenizkiy, 2021; Sung et al., 2019; Xue et al., 2021), or combine with Bi-LSTM for essay scoring (Beseiso et al., 2021). Studies show that incorporating handcrafted features alongside these models improves performance (Uto et al., 2020). However, traditional AES methods up to 2022 have focused more on linguistic elements than content and often neglected coherence and cohesion in essays (Ramesh and Sanampudi, 2022). Additionally, these methods primarily analyzed English texts and relied on holistic scoring, overlooking the multidimensional aspects of essay quality.

**Applying GPT Models to English Essay Data.** The emergence of closed-source LLMs like GPT-3 and

GPT-4 in 2022 introduced new possibilities for automated essay grading. For instance, (Chiang and yi Lee, 2023) compared GPT-3's ratings on 400 English text fragments to those of three teachers across categories such as grammaticality, cohesion, and relevance. While GPT-3 aligned well with human ratings on relevance, its performance in other categories showed weak correlations. Additionally, (Mizumoto and Eguchi, 2023) used GPT-3 to evaluate over 12,000 essays from the TOEFL11 dataset across four dimensions, discovering that combining GPT-3's insights with linguistic features yielded the best results. Further studies with GPT-3.5 and GPT-4 demonstrated moderate accuracy, with alignment rates ranging from 70% within a small deviation to 56% for exact matches in discourse coherence (Naismith et al., 2023). However, many of these studies relied on holistic scores and did not explore the multidimensional nuances of human evaluations.

**Applying GPT Models to Portuguese Argumentative Writing Data.** Among the studies applying LLMs to the evaluation of texts in the Portuguese language, I highlight the paper that motivated this study, which investigates the effectiveness of LLMs, particularly GPT-4, in assessing argumentative essays and generating feedback for military school students as part of the Mario Travasso Project, aimed at encouraging writing in Brazilian military schools.(da Silva et al., 2024) The research seeks to enhance students' critical writing skills by integrating automated feedback with human evaluation. It is structured into two phases: the first involves quantitative and qualitative comparisons between evaluations conducted by instructors and feedback generated by GPT-4 in categories such as topic choice, development, and references. The results revealed consistency in task-level feedback but highlighted GPT-4's limitations in addressing self-regulatory aspects and more complex contextual elements. The second phase evaluates the students' ability to improve their work based on the feedback received. Grounded in Hattie and Timperley's feedback model (2007), the study emphasizes the importance of combining AI capabilities with human oversight to optimize the educational impact of feedback. References include foundational works on feedback mechanisms (Hattie and Timperley, 2007) and recent advancements in educational applications of LLMs (Biswas, 2023; Firat, 2023).

# 3 METHODOLOGY

This study aims to analyze the performance of LLMs in evaluating essays from candidates in the IME selection process, based on a predefined rubric composed of seven categories. The following subsections detail the study's aspects, including the essays and evaluation criteria used, the participants, the application of the LLMs, and the metrics employed for analysis, as illustrated in Figure 1.

## 3.1 Candidate Essay Dataset

The entrance exam for the IME is one of the most competitive in Brazil, attracting thousands of candidates each year seeking admission to one of the country's most traditional institutions. In 2024, over 4,500 candidates registered to compete for around 80 available spots, with previous editions, such as 2017/2018, recording up to 6,290 registrations. The spots are divided into two categories: the Active option, aimed at candidates who wish to pursue a military career, and the Reserve option, for those who intend to work as civil engineers. The competition is extremely fierce, with a ratio of 68.08 candidates per spot in the Active option and 52.17 in the Reserve option in the 2017/2018 edition. This high level of competition reflects the academic rigor and stringent selection process, qualities that ensure the IME's tradition and excellence in training both military and civil engineers in Brazil.

The IME entrance exam consists of two phases. The first phase is an objective test with 40 questions, divided among Mathematics (15), Physics (15), and Chemistry (10). To be approved, candidates must answer at least 40% of the questions in each subject and achieve a minimum of 40 correct answers overall. In the second phase, candidates face essay-type exams: on the first day, questions in Mathematics, Physics, and Chemistry; on the following days, essay exams in Physics and Chemistry; and finally, objective and essay exams in Portuguese (including an essay) and English. Approval in the second phase depends on satisfactory performance in each area and achieving a good final score.

For this study, 100 essays from candidates in the 2024 entrance exam for the IME were randomly selected. The choice of essays was made to represent a diverse sample of the candidate pool, considering the variety of approaches and writing styles present in the exam essays.

## 3.2 Essay Assessment Category

In this study, the evaluation of essays was conducted across multiple categories, each with specific criteria and scoring systems. Table 1 presents the seven evaluation criteria.

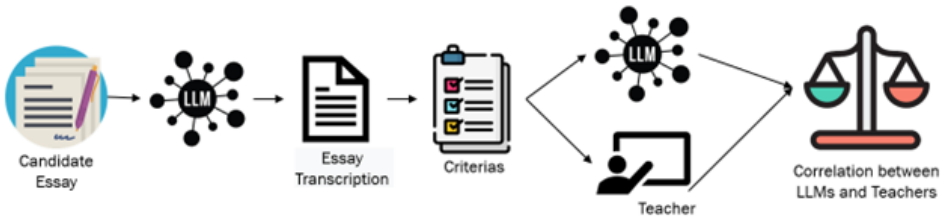


Figure 1: Design and Workflow the study.

Table 1: The 7 evaluation criteria, where each criterion is evaluated from 0 to 4 scores.

| No. | Title | Description |
|---|---|---|
| 1 | Theme | Assesses adherence to the proposed theme. |
| 2 | Types of Text | Focuses on conformity to the argumentative-essay genre. |
| 3 | Presentation | Evaluates text legibility and visual organization. |
| 4 | Structure of the Text | Examines the organization into introduction, body, and conclusion, as well as paragraph division. |
| 5 | Argumentative Strength / Coherence | Measures the ability to present consistent and cohesive arguments. |
| 6 | Cohesion | Evaluates the use of connectors and the hierarchy of ideas. |
| 7 | Grammatical Structure | Examines adherence to orthographic conventions, morphosyntactic, syntactic, and semantic rules. |

**Theme.** Assesses adherence to the proposed theme. Essays receive a score of 0 if there is a complete deviation from the theme, making it impossible to evaluate other criteria. Partial adherence to the theme, referred to as "tangential approach," results in a score of 1. The maximum score (2) is awarded to essays that fully address the theme.

**Text Type.** Focuses on conformity to the argumentative-essay genre. Essays that do not fit this genre receive a score of 0, while those that partially or fully meet this criterion are awarded scores of 1 and 2, respectively.

**Presentation.** Evaluates text legibility and visual organization. Essays with illegible sentences, excessive erasures, skipped lines, or lack of paragraph indentation receive a score of 0. Essays with clear handwriting and semantically autonomous paragraphs are awarded a score of 1.

**Text Structure.** Examines the organization into introduction, body, and conclusion, as well as paragraph division. Essays lacking this basic structure receive a score of 0. The maximum score (4) is given to essays that include all structural elements and well-divided paragraphs of more than three lines each.

**Argumentative Strength and Coherence.** Measures the ability to present consistent and cohesive arguments. Essays with no internal or external coherence or those employing extreme idealizations receive a score of 0. Scores progress (from 1 to 4) based on criteria such as logical flow between ideas and the use of facts and concepts to support arguments.

**Cohesion.** Evaluates the use of connectors and the hierarchy of ideas. Essays that fail to achieve thematic progression or hierarchy receive a score of 0. Mastery of connectors and other cohesive resources increases the score, reaching 4 when these elements are used with excellence.

**Grammatical Structure.** Examines adherence to orthographic conventions, morphosyntactic, syntactic, and semantic rules. Essays with seven or more errors receive a score of 0, while a score of 4 is awarded to essays with up to three errors and full compliance with grammatical norms.

In summary, these categories form a rigorous and well-rounded evaluation system, ensuring a thorough analysis of essays based on objective and standardized criteria.

### 3.3 LLM Essay Scoring

The automated evaluation of candidates' essays was conducted based on seven predefined criteria, and different LLMs were selected to compare the performance of various foundational models. Among the closed-source models, GPT-3.5 (gpt-3.5-turbo-0125), GPT-4 (gpt-4o-2024-05-13) (OpenAI, 2024a), and o1 (o1-preview) (OpenAI, 2024b) were utilized and integrated into the evaluation process via the OpenAI API. For open-source models, LLaMA 3-70B (AI@Meta, 2024) and Mixtral 8x7B (Jiang et al., 2024) were chosen. Preliminary tests included smaller variants of these models; however, due to their

```
You are a teacher. Analyze the essay written by the candidate according to the criteria indicated. Return a scalar number from 0 to 4 for each criterion. 0 means the criterion was not met, 4 means is completely accomplished. Return just one JSON. ## Criteria = {criteria}; ## Essay = {text}; ## Assessment =
```

Figure 2: Zero-shot prompt employed at all LLMs to ensure a fair comparison of your essay assessment performance.

inferior performance compared to their larger counterparts, these smaller variants were excluded from subsequent analyses.

Before starting the automated evaluation, each handwritten essay was transcribed by the selected LLMs using specific techniques to convert the text into digital format. The presentation criterion, included in the evaluation rubric, was used to assess the accuracy and quality of the transcription performed by the models. This step allowed for evaluating the LLMs' ability to handle different levels of legibility present in the handwritten essays, ensuring that the transcribed texts were faithful representations of the originals.

The configuration of the LLMs followed a zero-shot approach, where each model was instructed to evaluate one essay at a time. To ensure the independence of evaluations, a new session was initiated for each essay. Prompts were meticulously designed to reflect the predefined evaluation criteria, ensuring consistency across all assessments conducted by the models. This strategy enabled a systematic and impartial comparison of LLM performance.

The prompt design, illustrated in Figure 2, assigned the model the role of a teacher, contextualized the essay as part of the IME selection process, specified the analysis task based on the established criteria, and defined the output format in JSON. This prompt format was uniformly applied throughout all experiments to ensure standardization of results. Although variations in prompts could influence outcomes (Stahl et al., 2024b), prompt engineering was not the primary focus of this study.

To evaluate the reliability of predictions and account for the stochastic nature of the models, each essay was evaluated ten times by each LLM using a temperature setting of 0.7. The average of these ten evaluations was used as the final score assigned by the LLMs, analogous to the average scores provided by three human evaluators for the same essay. This approach ensured a robust analysis aligned with traditional evaluation standards.

### 3.4 Analysis

To address the proposed research questions, we conducted a systematic analysis of the evaluation results obtained from both human raters and LLMs.

**RQ1** focuses on the reliability and quality of the evaluations. To address this, we analyzed multiple runs of each model on the same text, calculating the intraclass correlation and treating each run as an individual rater. Throughout the subsequent analyses, the average of the ten runs was considered the final score.

**RQ2** investigates the relationship between the evaluations performed by LLMs and those conducted by teachers. We used Spearman's correlation coefficient to identify similarities between the scores assigned by the models and the human evaluators.

Finally, **RQ3** examines the overall holistic scores and the multidimensional aspects defined in the evaluation categories, encompassing language and content-related criteria. By comparing the distribution of scores between LLMs and human raters and applying the Mann-Whitney U test, we identified differences in how specific criteria are assessed by each group.

## 4 RESULTS AND DISCUSSIONS

The results highlighted the discrepancies between the evaluations conducted by teachers and those generated by open- and closed-source LLMs for the candidates' essays, aiming to address the three main research questions.

### 4.1 RQ1: Reliability of Model Predictions

To assess the reliability of each model's evaluations in the context of RQ1, multiple runs were performed using the same prompt. Each data point was evaluated ten times, following an approach similar to that used by human raters. The Intraclass correlation coefficient (ICC) for all the foundational models analyzed are presented in Table 2.

The results indicate that closed-source models exhibit considerable consistency in their evaluations, with ICC scores ranging from moderate to good (0.73 - 0.84) (Koo and Li, 2016). In contrast, open-source models such as LLaMA and Mixtral showed high variability, reflected in low ICC values and poor agreement across different runs. This disparity highlights the importance of accounting for these inconsistencies in subsequent analyses.

Table 2: ICC values comparing.

| | GPT-3.5 | GPT-4 | GPT-o1 | LLaMA | Mixtral |
|---|---|---|---|---|---|
| **ICC** | 0.84 | 0.73 | 0.80 | -0.04 | 0.01 |

### 4.2 RQ2: Correlation Analysis Between LLM Evaluations and Teacher Assessments

In the analysis of RQ2, the evaluations conducted by LLMs and teachers were compared using Spearman correlation coefficients (*r)* for all evaluation criteria, as shown in Table 3. The o1 model stood out as the most aligned with human assessments, achieving significance in six out of the seven analyzed categories. Notably, it was the only model to demonstrate a high and significant correlation of 0.742 with teacher evaluations in the presentation category.

GPT-4 showed significant correlations in five out of the seven categories, indicating moderate agreement with human evaluators in most cases. GPT-3.5, on the other hand, achieved significant correlations in only two categories, highlighting the improvements in the more recent versions of the models. Conversely, Mixtral exhibited weak or non-significant correlations across all criteria, while LLaMA demonstrated nearly zero correlations and, in some cases, even negative correlations with teacher evaluations, underscoring its inconsistency in replicating human assessments.

### 4.3 RQ3: Rating Comparison Between LLM and Teacher

The analysis of the discrepancies between teacher and LLM evaluations compared the average scores assigned to each criterion, aiming to deepen the insights presented in this study. Overall, the GPT models displayed higher averages across all individual criteria, while teachers adopted a stricter approach in their assessments. The LLaMA model showed averages similar to those of the teachers. In contrast, the Mixtral model consistently exhibited substantially higher averages across all categories. These differences, reflected in the overall scores, are also evident in the evaluations of each criterion, indicating that the final ratings align with the more detailed analyses conducted by both teachers and LLMs.

Additionally, the variations observed in Table 4 are noteworthy. While the LLaMA model demonstrated average scores comparable to those of the teachers, its variance was significantly smaller, with scores clustered around the midpoint. The Mixtral model followed a similar pattern, also showing reduced variance. On the other hand, GPT-3.5, o1, and the teachers displayed higher variances, suggesting a greater differentiation in their assessments. The low ICC values recorded for the LLaMA and Mixtral models corroborate these findings, highlighting limited agreement among their evaluations, resulting in median scores and reduced variability.

Table 4 also presents the p-values from the Mann-Whitney U test, comparing the score distributions assigned by teachers and LLMs for each criterion. No significant differences were observed between teacher evaluations and those of GPT-4 for criteria such as textual structure and grammatical structure, all of which are language-related. Similarly, the o1 model showed no significant differences in grammatical structure and textual structure, also language-related criteria. Since LLMs are trained on large volumes of textual data encompassing various writing styles and formalities, they are optimized to identify patterns and stylistic elements. Consequently, superficial aspects such as grammar and sentence organization can be analyzed in a manner comparable to human evaluators. Notably, LLMs demonstrated the ability to evaluate candidate texts accurately, despite limited exposure to this type of content during training, highlighting their efficient generalization capabilities.

However, discrepancies between teacher evaluations and those of the GPT-4 and o1 models were significant in criteria such as theme, text type, argumentative strength and coherence, and cohesion, all of which are content-related categories. These differences may be attributed to several factors. Although LLMs are capable of producing coherent texts, they still lack the deep semantic understanding that humans possess. This limitation manifests in logical reasoning errors, particularly regarding contextual details and maintaining logical consistency. Additionally, during training, the models may acquire and perpetuate biases, resulting in more lenient evaluations. It is important to note the variations among model versions, as evidenced earlier in Table 4. This is surprising, given that newer versions typically outperform their predecessors in complex tasks, aligning more closely with human evaluative standards. These findings underscore the ongoing challenges in aligning LLM assessments with human judgments, especially in criteria with a strong emphasis on content.

## 5 FUTURE WORKS

As future work, it is proposed to enhance prompt engineering with techniques such as Chain-of-Thought (CoT) and Few-Shot Learning, aiming to improve the

Table 3: Spearman correlation coefficients ($r$) and the corresponding p-values comparing teacher ratings with those of the five LLM-models. Significant correlations are highlighted in bold.

| Category | GPT-3.5 | | GPT-4 | | GPT-o1 | | LLaMA | | Mistral | |
|---|---|---|---|---|---|---|---|---|---|---|
| | r | p | r | p | r | p | r | p | r | p |
| **Theme** | -0.005 | 0.984 | 0.159 | 0.504 | **0.699** | **0.001** | 0.131 | 0.581 | 0.252 | 0.284 |
| **Text Type** | 0.313 | 0.179 | 0.325 | 0.162 | 0.127 | 0.594 | 0.014 | 0.953 | 0.348 | 0.133 |
| **Presentation** | 0.418 | 0.067 | **0.575** | **0.008** | **0.742** | **0.000** | 0.091 | 0.703 | 0.311 | 0.182 |
| **Text Structure** | **0.520** | **0.019** | **0.626** | **0.003** | **0.675** | **0.001** | 0.177 | 0.455 | 0.211 | 0.372 |
| **Argumentative Strength and Coherence** | 0.279 | 0.234 | 0.386 | 0.092 | **0.466** | **0.038** | -0.094 | 0.694 | 0.376 | 0.103 |
| **Cohesion** | 0.425 | 0.062 | **0.585** | **0.007** | **0.608** | **0.004** | -0.032 | 0.893 | 0.442 | 0.051 |
| **Grammatical Structure** | **0.728** | **0.000** | **0.846** | **0.000** | **0.814** | **0.000** | 0.406 | 0.076 | 0.005 | 0.984 |

Table 4: Criteria-based evaluation of all essays, including p-values from the Mann-Whitney U test comparing the distribution of the LLM scores to teacher scores for each individual category. Significant differences are indicated in bold.

| Category | Teacher | GPT-3.5 | p | GPT-4 | p | GPT-o1 | p | LLaMA | p | Mixtral | p |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Theme** | 3.61±1.16 | 4.10±0.82 | 0.10 | **4.40±0.60** | **0.00** | **4.96±1.12** | **0.00** | 3.89±0.45 | 0.75 | 4.19±0.59 | 0.09 |
| **Text Type** | 3.85±0.79 | 4.34±0.91 | 0.07 | **4.45±0.47** | **0.01** | **4.87±0.55** | **0.00** | 3.77±0.43 | 0.39 | **4.53±0.53** | **0.00** |
| **Presentation** | 3.65±0.86 | 3.85±0.81 | 0.30 | **4.34±0.66** | **0.02** | **4.41±0.68** | **0.01** | 3.55±0.41 | 0.97 | **4.60±0.41** | **0.00** |
| **Text Structure** | 3.74±0.97 | 3.72±0.98 | 0.92 | 4.01±0.68 | 0.34 | 4.11±0.86 | 0.24 | 3.69±0.45 | 0.96 | **4.45±0.50** | **0.01** |
| **Argumentative Strength and Coherence** | 3.77±0.94 | 4.24±1.14 | 0.17 | **4.38±0.57** | **0.01** | **4.78±0.53** | **0.00** | 3.82±0.51 | 0.53 | **4.33±0.51** | **0.02** |
| **Cohesion** | 3.93±0.88 | 4.13±1.17 | 0.66 | **4.81±0.59** | **0.00** | **5.19±0.68** | **0.00** | 3.85±0.51 | 0.83 | **4.57±0.47** | **0.01** |
| **Grammatical Structure** | 3.43±1.16 | 3.80±1.05 | 0.36 | 3.35±0.91 | 0.83 | 2.91±0.86 | 0.14 | 3.62±0.72 | 0.66 | **4.40±0.47** | **0.00** |

models' interpretation of subjective and contextual criteria. Additionally, plans include expanding training with diverse data, incorporating essays from different academic levels, writing styles, and texts with intentional errors, which will increase the models' ability to generalize and provide accurate corrections.

Other approaches involve integrating closed and open models into a hybrid solution, balancing costs and performance, and applying statistical normalization techniques to reduce variability in evaluations. Priority will also be given to adapting the models to the cultural and linguistic context of Brazilian Portuguese, ensuring greater precision in analyses. Finally, complementary metrics combining qualitative and quantitative analyses will be developed, enabling a more comprehensive evaluation of criteria such as coherence, style, and originality.

# 6 CONCLUSION

This study assessed the performance of both open-source and closed-source LLMs in evaluating essays from candidates participating in the IME selection process, comparing their outcomes with teacher evaluations. The analysis aimed to identify the strengths and limitations of these models in essay assessment. Among the models evaluated, the o1 model demonstrated notable reliability and strong alignment with teacher evaluations, particularly in language-related criteria. However, it exhibited a consistent tendency to assign higher overall scores than human raters. In contrast, open-source models like LLaMA and Mixtral displayed low variance and weak correlations with teacher assessments, which limited their effectiveness in accurately evaluating essays.

The study also underscores several limitations and potential directions for future research in the use of LLMs for automated essay evaluation. One key limitation was the absence of prompt engineering, which could improve the performance of open-source models by incorporating advanced techniques such as Chain-of-Thought (CoT) Engineering, already integrated into the o1 series. Additionally, the research focused on a narrow set of models (GPT-3.5, GPT-4, o1, LLaMA, and Mixtral), suggesting that future studies should include alternative models like Claude or Gemini to enable a more comprehensive evaluation of LLM performance across different architectures. The study's scope was further limited to a single essay format, highlighting the need for more diverse datasets and essay types, such as argumentative essays, to enhance the generalizability and robustness of findings.

Another significant challenge identified was the inherent variability in LLM evaluations, especially among open-source models. This highlights the necessity of employing mechanisms to aggregate multiple scores to reduce the impact of outliers. Furthermore, the variability observed among human raters points to the absence of a definitive gold standard, emphasizing the importance of incorporating this subjective nature into LLM training and evaluation frame-

works.

Despite these challenges, the study revealed a consistent trend of improvement in OpenAI's closed-source models, with newer iterations demonstrating increased reliability and closer alignment with human assessments. These findings, combined with ongoing advancements in the field, indicate that LLMs are becoming increasingly viable tools for automated essay evaluation, offering valuable support in educational contexts.

## REFERENCES


AI@Meta (2024). Llama 3 model card. Technical documentation.

Beseiso, M., Alzubi, O. A., and Rashaideh, H. (2021). A novel automated essay scoring approach for reliable higher educational assessments. *Journal of Computing in Higher Education*, 33:727–746.

Bewersdorff, A., Hartmann, C., Hornberger, M., Seßler, K., Bannert, M., Kasneci, E., Kasneci, G., Zhai, X., and Nerdel, C. (2024). Taking the next step with generative artificial intelligence: The transformative role of multimodal large language models in science education. *arXiv preprint*, arXiv:2401.00832.

Bewersdorff, A., Seßler, K., Baur, A., Kasneci, E., and Nerdel, C. (2023). Assessing student errors in experimentation using artificial intelligence and large language models: A comparative study with human raters. *Computers and Education: Artificial Intelligence*, 5:100177.

Bhat, S., Nguyen, H. A., Moore, S., Stamper, J. C., Sakr, M., and Nyberg, E. (2022). Towards automated generation and evaluation of questions in educational domains. In Mitrovic, A. and Bosch, N., editors, *Proceedings of the 15th International Conference on Educational Data Mining (EDM)*, pages 701–704, Durham, United Kingdom. International Educational Data Mining Society.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York, NY.

Biswas, S. (2023). Role of chatgpt in computer programming.: Chatgpt in computer programming. *Mesopotamian Journal of Computer Science*, 2023:8–16.

Bluche, T., Kermorvant, C., and Louradour, J. (2014). Joint learning of convolutional neural networks and label trees for grapheme-based handwriting recognition. pages 527–543.

Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A. M., Jozefowicz, R., and Bengio, S. (2016). Generative adversarial networks for text generation.

Breuel, T. M. (2005). The ocropus open source ocr system. volume 6076, page 60760K. SPIE.

Chiang, C.-H. and yi Lee, H. (2023). Can large language models be an alternative to human evaluations? In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

da Silva, D. C. A., de Mello, C. E., and Garcia, A. C. B. (2024). Analysis of the effectiveness of large language models in assessing argumentative writing and generating feedback. In *ICAART (2)*, pages 573–582.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Doermann, D. and Tombre, K. (2014). *Handbook of Document Image Processing and Recognition*. Springer, London, UK.

Doewes, A. and Pechenizkiy, M. (2021). On the limitations of human-computer agreement in automated essay scoring. In *Proceedings of the 14th International Conference on Educational Data Mining (EDM21)*, pages 475–480, Paris, France. International Educational Data Mining Society.

Firat, M. (2023). What chatgpt means for universities: Perceptions of scholars and students. *Journal of Applied Learning and Teaching*, 6(1).

Gatos, B., Louloudis, G., Causer, T., Grint, K., Romero, V., Sánchez, J. A., Toselli, A. H., and Vidal, E. (2014). Ground-truth production in the transcriptorium project. In *2014 11th IAPR International Workshop on Document Analysis Systems*, pages 237–241.

Gatos, B., Pratikakis, I., and Perantonis, S. Adaptive degraded document image binarization. *Pattern Recognition*.

Graves, A., Fernández, S., and Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. pages 369–376. ACM.

Graves, A., Liwicki, M., Bunke, H., Schmidhuber, J., and Fernández, S. (2008). A novel approach to on-line handwriting recognition based on bidirectional long short-term memory networks. *Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR)*, 2:367–371.

Graves, A., Liwicki, M., Fernández, S., Bertolami, R., Bunke, H., and Schmidhuber, J. (2009). A novel connectionist system for unconstrained handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):855–868.

Hattie, J. and Timperley, H. (2007). The power of feedback. *Review of educational research*, 77(1):81–112.

Jaderberg, M., Simonyan, K., Vedaldi, A., and Zisserman, A. (2014). End-to-end text recognition with convolutional neural networks. pages 1–11. BMVA Press.

Ji, H., Han, I., and Ko, Y. (2023). A systematic review of conversational ai in language education: Focusing on the collaboration with human teachers. *Journal of Research on Technology in Education*, 55(1):48–63.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F.,

Lengyel, G., Lample, G., Saulnier, L., et al. (2023). Mistral 7b. Technical report, Mistral AI.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., de las Casas, D., Hanna, E. B., Bressand, F., et al. (2024). Mixtral of experts. Technical report, Mistral AI.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., and Hüllermeier, E. (2024). Can ai grade your essays? *Educational Assessment and Artificial Intelligence Review*. Forthcoming.

Ke, Z. and Ng, V. (2019). Automated essay scoring: A survey of the state of the art. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 6300–6308. International Joint Conferences on Artificial Intelligence Organization.

Koo, T. K. and Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155–163.

Koutník, J., Greff, K., Gomez, F., and Schmidhuber, J. (2014). Handwriting recognition with large multidimensional long short-term memory recurrent neural networks. pages 2193–2201. Curran Associates, Inc.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Mizumoto, A. and Eguchi, M. (2023). Exploring the potential of using an ai language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2):100050.

Muñoz, S. A. S., Gayoso, G. G., Huambo, A. C., Tapia, R. D. C., Incaluque, J. L., Aguila, O. E. P., Cajamarca, J. C. R., Acevedo, J. E. R., Rivera, H. V. H., and Arias-Gonzáles, J. L. (2023). Examining the impacts of chatgpt on student motivation and engagement. *Social Space*, 23(1):1–27.

Naismith, B., Mulcaire, P., and Burstein, J. (2023). Automated evaluation of written discourse coherence using gpt-4. In Kochmar, E., Burstein, J., Horbach, A., Laarmann-Quante, R., Madnani, N., Tack, A., Yaneva, V., Yuan, Z., and Zesch, T., editors, *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 394–403, Toronto, Canada. Association for Computational Linguistics.

Neto, A. F. S., Bezerra, B. L. D., Araújo, S. S., Souza, W. M. A. S., Alves, K. F., Oliveira, M. F., Lins, S. V. S., Hazin, H. J. F., Rocha, P. H. V., and Toselli, A. H. (2024a). Bressay: A brazilian portuguese dataset for offline handwritten text recognition. In *Document Analysis and Recognition - ICDAR 2024: 18th International Conference, Athens, Greece, August 30–September 4, 2024, Proceedings, Part II*, page 315–333, Berlin, Heidelberg. Springer-Verlag.

Neto, A. F. S., Bezerra, B. L. D., Araujo, S. S., Souza, W. M. A. S., Alves, K. F., Oliveira, M. F., Lins, S. V. S., Hazin, H. J. F., Rocha, P. H. V., and Toselli, A. H. (2024b). Bressay: A brazilian portuguese dataset for offline handwritten text recognition. In *18th International Conference on Document Analysis and Recognition (ICDAR)*, Athens, Greece. Springer.

Nguyen, H. A., Stec, H., Hou, X., Di, S., and McLaren, B. M. (2023). Evaluating chatgpt's decimal skills and feedback generation in a digital learning game. In Viberg, O., Jivet, I., Muñoz-Merino, P. J., Perifanou, M., and Papathoma, T., editors, *European Conference on Technology Enhanced Learning*, pages 278–293, Cham. Springer Nature Switzerland.

OpenAI (2024a). Gpt-4 technical report.

OpenAI (2024b). Openai o1 system card. Technical report, OpenAI.

Ramesh, D. and Sanampudi, S. K. (2022). An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review*, 55(3):2495–2527.

Rice, S. V. (1999). *Optical Character Recognition: An Illustrated Guide to the Frontier*. Springer, Boston, MA.

Sawatzki, J., Schlippe, T., and Benner-Wickner, M. (2021). Deep learning techniques for automatic short answer grading: Predicting scores for english and german answers. In Cheng, E. C. K., Koul, R. B., Wang, T., and Yu, X., editors, *International Conference on Artificial Intelligence in Education Technology*, pages 65–75, Singapore. Springer Nature Singapore.

Seßler, K., Xiang, T., Bogenrieder, L., and Kasneci, E. (2023). Peer: Empowering writing with large language models. In Viberg, O., Jivet, I., Muñoz-Merino, P. J., Perifanou, M., and Papathoma, T., editors, *Responsive and Sustainable Educational Futures*, pages 755–761, Cham. Springer Nature Switzerland.

Stahl, M., Biermann, L., Nehring, A., and Wachsmuth, H. (2024a). Exploring llm prompting strategies for joint essay scoring and feedback generation. *arXiv preprint*, arXiv:2404.15845. [cs.CL].

Stahl, M., Biermann, L., Nehring, A., and Wachsmuth, H. (2024b). Exploring llm prompting strategies for joint essay scoring and feedback generation. *arXiv*, 2404.15845.

Sung, C., Dhamecha, T. I., and Mukhi, N. (2019). Improving short answer grading using transformer-based pretraining. In Isotani, S., Millán, E., Ogan, A., Hastings, P., McLaren, B., and Luckin, R., editors, *Artificial Intelligence in Education*, pages 469–481, Cham. Springer International Publishing.

Uto, M., Xie, Y., and Ueno, M. (2020). Neural automated essay scoring incorporating handcrafted features. In Scott, D., Bel, N., and Zong, C., editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6077–6088, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Wang, Z. and Li, Y. (2020). Handwritten text recognition: Benchmarking of current state-of-the-art. arXiv:2003.12294.

Xue, J., Tang, X., and Zheng, L. (2021). A hierarchical bert-based transfer learning approach for multidimensional essay scoring. *IEEE Access*, 9:125403–125415.