Automated ENEM Essay Scoring and Feedbacks: A Prompt-Driven LLM Approach

Wesley Alves da Silva[1], **Cristiano Coelho de Araujo**[1]

[1]Centro de Informatica (CIn) – Universidade Federal de Pernambuco (UFPE)
Mailbox 7851 – 50.732-970 – Recife – PE – Brasil

{was4,cca2}@cin.ufpe.br

Abstract. *This paper presents a novel LLM-based solution for automated evaluation of essays from the National High School Exam (ENEM), the largest educational test in Brazil, addressing the significant gap in Portuguese-language automated essay scoring (AES). Leveraging a chain of meticulously crafted prompts and advanced LLM architectures, our system achieves 100% adjacent agreement with human raters across all five ENEM competencies and the final score, significantly outperforming existing methods. Beyond accurate scoring, the system generates detailed, competency-specific feedback, transforming it into a valuable learning tool for students. Our approach, based on prompt engineering and adhering strictly to official ENEM scoring rubrics, offers a robust and scalable solution for large-scale educational assessment in Portuguese, enhancing both efficiency and educational value.*

1. Introduction

Over the past few years, essay evaluation has emerged as a significant challenge in the educational domain, particularly in large-scale examinations such as the National High School Exam (Exame Nacional do Ensino Médio, ENEM) in Brazil. The ENEM essay assesses students' ability to express ideas clearly and coherently and aims to evaluate their argumentative and critical thinking skills on contemporary societal issues.

The ENEM is Brazil's most extensive educational assessment test and the most taken test in the Portuguese language worldwide. Its importance transcends national boundaries, serving as a gateway to higher education opportunities not just within Brazil but also in Portugal and select institutions in the United States. For Brazilian students, a good ENEM score can open doors to public universities through the Unified Selection System (SiSU), scholarships via the Programa Universidade para Todos (ProUni), and funding through the Fundo de Financiamento Estudantil (FIES). Additionally, it serves as a pathway for Portuguese universities, owing to bilateral agreements between Brazil and Portugal, which recognize the ENEM scores for admissions. Furthermore, some American universities accept the ENEM as part of their admissions criteria, making it a pivotal examination for students seeking diverse academic opportunities.

With approximately four million participants in 2023 alone [Brasil], the ENEM stands out as one of the country's most significant and critical educational assessments. Students prepare for this challenge and engage in continuous text production practices, necessitating detailed and practical feedback to enhance their writing skills.

However, the meticulous grading of essays demands considerable time and resources from educators, often resulting in the limited quantity and quality of feedback provided to students. Adopting technological solutions, such as Large Language Models (LLMs) [Mansour et al. 2024], presents a promising opportunity to optimize and enhance the essay evaluation process.

This work aims to streamline and enrich the essay evaluation process by providing students with detailed and personalized feedback beyond mere grading. The proposed solution seeks to help students understand their strengths and areas for improvement, thus effectively enhancing their writing skills. By integrating the evaluative principles of the ENEM, as outlined in the Participant Handbook [INEP 2023] and the manuals for essay examiners [INEP 2020a, INEP 2020b, INEP 2020c, INEP 2020d, INEP 2020e], the proposed solution aims to ensure quality and objectivity in the assessment of students' written texts.

This paper is organized as follows: Section 2 provides an overview of the ENEM essay and its five competency-based scoring rubric. Section 3 reviews existing research on automated essay scoring (AES), large language models (LLMs), and prompt engineering techniques. Section 4 describes our methodology, encompassing data acquisition, model selection, prompt design, and the evaluation framework. Section 5 presents the architecture of our proposed LLM-based AES system, with a detailed examination of prompt design. Section 6 presents and analyzes the results, including a comparison with a related study and an illustrative example of the generated feedback. Section 7 concludes the paper by summarizing our findings and outlining potential avenues for future research.

2. The ENEM Essay Style and Assessment

In Brazilian higher education, the National High School Exam (ENEM) essay is a pivotal assessment tool, challenging students to engage in sophisticated argumentative discourse. This written task, conducted in Portuguese, demands a nuanced approach to contemporary issues of public interest, requiring candidates to demonstrate their linguistic prowess and their capacity for critical analysis and problem-solving. At its core, the ENEM essay can be conceptualized as a unique discursive genre [Silva and de Lima Cavalcante 2023].

This perspective illuminates the essay's distinctive characteristics, including its structural composition, thematic focus, and verbal style, all contributing to its role as a specialized form of academic discourse. The stylistic elements of successful ENEM essays are multifaceted and encompass a range of discursive strategies. These include the judicious use of modalizers and argumentative techniques, which position the author's voice within the text and manage the interplay of various perspectives [Prado and Morato 2017]. Moreover, the essay's effectiveness hinges on its cohesion, coherence, and thematic progression, achieved through careful lexical selection and adherence to formal Portuguese writing conventions. Thematically, the ENEM essay grapples with contemporary issues of societal importance, spanning political, economic, social, and cultural domains. The ability to accurately interpret the essay prompt and maintain a laser-like focus on the central theme is paramount.

Candidates must navigate this thematic landscape precisely, ensuring their argumentation remains relevant and syntactically sound. Another distinguishing feature of the ENEM essay is its emphasis on practical problem-solving. Beyond presenting a well-reasoned argument, candidates are tasked with proposing viable interventions to address the issue. These proposals must be relevant to the theme and demonstrate respect for human rights, underscoring the essay's role in fostering socially responsible thinking. In essence, the ENEM essay represents a complex intersection of linguistic skills, critical thinking, and social awareness. By synthesizing these elements within the framework of a specific discursive genre, it serves as a powerful tool for assessing a student's readiness for higher education and their potential to contribute meaningfully to societal discourse. As such, understanding the nuances of this assessment is crucial for both students and educators in the Brazilian educational landscape.

The ENEM essay evaluation process employs a rigorous dual-rater system to ensure impartial assessment. Each essay undergoes independent evaluation by two experts, who assign scores across five competencies on a $0-200$ point scale. To guarantee reliability, essays with a score discrepancy exceeding 100 points between raters are subjected to a third evaluation. The final score is determined by calculating the mean of the initial two scores (for discrepancies within acceptable limits) or the closest two scores among the three (for significant discrepancies). It evaluates five skills, which are described in the following sections.

2.1. Competency 1 (C1)

Competency 1 of the ENEM Essay, titled *"Demonstrating Mastery of the Formal Written Portuguese Language,"* serves as the cornerstone for evaluating the quality of the dissertative-argumentative essay produced by participants. In this initial analysis stage, the focus falls on the candidates' linguistic proficiency, attesting to their ability to express themselves clearly, concisely, and correctly per the Portuguese language's grammatical and orthographic norms [INEP 2020a].

The assessment of Competency 1 extends beyond the mere absence of grammatical errors. It encompasses a broader range of aspects that demonstrate mastery of formal written Portuguese, such as:

- **Appropriate Lexical Choices:** The candidate should select rich and precise vocabulary, avoiding colloquial terms, slang, or informal expressions that compromise the formality of the writing.
- **Textual Structure:** The essay should follow the dissertation-argumentative structure, with a clearly defined introduction, development, and conclusion, each fulfilling its specific function in constructing the text.
- **Punctuation:** Correct use of punctuation is crucial to ensure the clarity and proper interpretation of the text, demonstrating knowledge of the orthographic and punctuation norms of the Portuguese language.

2.2. Competency 2 (C2)

In the journey of the ENEM Essay, Competency 2, titled *"Understanding the Essay Prompt and Applying Concepts from Various Areas of Knowledge to Develop the Theme Within the Structural Limits of the Dissertative-Argumentative Text"*, emerges as the second crucial challenge, requiring candidates to demonstrate more than mere linguistic skills. At this stage, the focus shifts to the ability to argue consistently and constructively, showcasing intellectual maturity and mastery of various areas of knowledge [INEP 2020b].

The assessment of Competency 2 seeks to identify the candidate's ability to:

- **Understand the Essay Prompt:** The essay should demonstrate that the candidate comprehends the proposed theme, its implications, and the exam instructions.

- **Articulate Ideas and Arguments:** Ideas must be logically and cohesively connected, constructing a clear and convincing line of reasoning.
- **Use Sociocultural Repertoire:** The candidate should present concrete and relevant examples from various fields of knowledge to support their arguments.
- **Present Proposals for Intervention:** The essay should analyze the theme and propose feasible and creative solutions for the issues presented.

2.3. Competency 3 (C3)

After that, Competency 3, titled *"Selecting, Relating, Organizing, and Interpreting Information, Facts, Opinions, and Arguments in Defense of a Point of View"*, focuses on constructing a cohesive and convincing argumentative essay, defending a viewpoint in a clear, consistent, and well-structured manner [INEP 2020c].

The assessment of Competency 3 seeks to identify the candidate's ability to:

- **Select, Relate, Organize, and Interpret Information:** The essay should present relevant data, facts, and opinions on the topic, logically and coherently connected, demonstrating a critical and analytical perspective on the subject.
- **Construct a Textual Project:** The essay must follow a clear and organized structure, with a well-defined introduction, development, and conclusion, each serving its specific role in the argumentation.
- **Use Cohesive Resources:** The text should include connectives and textual markers that ensure cohesion and the progression of ideas, making the argument easy to follow and understand.
- **Present a Clear Point of View:** The candidate must defend a clear and precise thesis on the proposed theme, using consistent and well-supported arguments to persuade the reader.

2.4. Competency 4 (C4)

Another fundamental stage is Competency 4, where the focus shifts to the ability to write clearly, concisely, and objectively, using language in a precise and efficient manner to convey the message to the reader [INEP 2020d]. The title of Competency 4 is *"Demonstrating Knowledge of the Linguistic Mechanisms Necessary for the Construction of Argumentation"*.

The assessment of Competency 4 goes beyond merely choosing elegant words. It seeks to identify the candidate's ability to:

- **Select Appropriate Vocabulary:** The text should feature rich and precise vocabulary that clearly conveys the intended meaning, avoiding colloquial terms, slang, or informal expressions that undermine the formality of the writing.
- **Construct Concise and Cohesive Sentences:** Sentences should be well-structured, with a clear subject, verb, and complement, avoiding long and complex sentences that hinder readability and comprehension.
- **Use Correct Punctuation:** Punctuation must be used appropriately to ensure clarity and accurate interpretation of the text, demonstrating knowledge of Portuguese orthographic and punctuation norms.
- **Avoid Redundancy and Wordiness:** The text should present ideas directly and objectively, avoiding unnecessary repetitions and irrelevant information that disrupt the reading flow.

2.5. Competency 5 (C5)

In the final stretch of the ENEM Essay, titled *"Developing a Proposal for Intervention for the Addressed Problem, Respecting Human Rights"*, Competency 5 is one of the most important, as it focuses exclusively on the ability to present a concrete and viable intervention proposal for the problem addressed in the given topic, demonstrating social commitment and proactivity in seeking solutions [INEP 2020e].

The main points evaluated by Competency 5 are:

- **Propose Feasible and Realistic Solutions:** The intervention proposal must be concrete and executable, with specific actions that can be implemented, considering society's resources and actual conditions.
- **Present a Detailed Proposal:** The text must describe the steps to be followed to implement the proposal, including the responsible parties for each stage, the necessary resources, and the expected execution deadlines.
- **Consider Human Rights and Sustainability:** The proposal must align with human rights and environmental sustainability principles, promoting social well-being and ecological preservation.
- **Link the Proposal to the Development of the Text:** The intervention proposal should be related to the proposed topic and the arguments presented throughout the essay, demonstrating coherence and textual cohesion.

3. Related Work

3.1. Automated Essay Scoring

The domain of automated essay scoring (AES) has been a focal point of research for half a century, with a diverse array of methodologies proposed to tackle this complex challenge. Traditional approaches relied on manually crafted features, such as lexical and syntactic elements, to discriminate between varying-quality essays [Phandi et al. 2015, Chen and He 2013]. However, recent years have witnessed a paradigm shift towards neural architectures, including pre-trained language models, which have exhibited promising results in this domain [Wu et al. 2023, Dong et al. 2017].

Impressive performance in AES has been attained through utilizing multi-layer perceptrons, which concurrently execute regression and ranking optimization [Xie et al. 2022]. These models have demonstrated superior capabilities in accurately assessing essay proficiency.

3.2. Large Language Models

Large Language Models (LLMs) have emerged as a revolutionary force in Natural Language Processing (NLP), demonstrating remarkable capabilities across various tasks. As highlighted by [Brown et al. 2020], these models, one of the primary abilities of LLMs is to generate high-quality text and perform well on tasks defined on the fly, underscoring their potential to transform various aspects of language understanding and generation. The power of LLMs lies in their capacity to be "prompted" to perform a multitude of NLP tasks, given just a few examples as input. This flexibility opens up numerous possibilities for applications in areas such as text summarization, question answering, and even creative writing. However, [Bommasani et al. 2021] showed that these models are not without their challenges. LLMs can sometimes exhibit unintended behaviors, such as generating factually incorrect information, producing biased or toxic content, or failing to adhere to user instructions. These issues highlight the importance of careful, prompt engineering and the need for robust evaluation frameworks to ensure the responsible deployment of LLMs in real-world applications.

Therefore, developing an automated solution for evaluating ENEM essays using large language models (LLMs) necessitates a well-structured and meticulously crafted prompt design. The importance of prompt engineering in this context cannot be overstated, as it fundamentally influences the efficacy and accuracy of the model's performance. Prompt engineering is the process of designing and refining input prompts that guide the LLMs in generating desired outputs. This section delves into the critical aspects of prompt design, drawing from various related works and methodologies to build a robust framework for our solution.

3.3. Prompt Engineering

One of the foundational studies in prompt engineering is presented by [Dong et al. 2023], highlighting the significant advantages of in-context learning. In-context learning allows LLMs to understand and generate responses based on the provided context, making it a critical component in designing effective prompts. By leveraging this approach, we can ensure that the model comprehends the nuances of ENEM essays and provides accurate corrections.

A pivotal technique in enhancing the reasoning capabilities of LLMs is the "chain of thought" prompting, as explored in the study by [Wei et al. 2023]. This method provides a series of intermediate reasoning steps within the prompt, significantly improving the model's ability to perform complex reasoning tasks. For the automated correction of ENEM essays, incorporating chain of thought prompts can facilitate the model's understanding of the essay structure, coherence, and argumentation, leading to more precise and insightful feedback. In order to address highly complex tasks, one notable approach that has gained traction is prompt chaining, a technique that involves sequencing multiple prompts to guide LLMs through multi-step reasoning processes. This method has shown promise in addressing tasks too intricate for a single prompt to handle effectively. The work of [Wu et al. 2022] introduces PromptChainer, a visual programming interface that facilitates creating and debugging prompt chains. Their research highlights the importance of scaffolding for output transformation between nodes and the need for multi-granular debugging capabilities in chain authoring.

The efficacy of prompt chaining compared to alternative approaches has been a subject of investigation in text summarization. [Sun et al. 2024] conducted a comparative study between prompt chaining and stepwise prompting, two strategies designed to emulate the human-like iterative process of drafting, critiquing, and refining summaries. Their findings suggest that prompt chaining, which employs a series of discrete prompts for each phase, outperforms the stepwise prompt method that integrates all phases within a single prompt. This superiority is attributed to prompt chaining's ability to produce a more genuine refinement process instead of the potentially simulated refinement observed in stepwise prompting.

3.4. AES with LLMs

In tandem with these advancements, the advent of large language models (LLMs) has ignited substantial interest in their potential for AES. These transformer-based models possess remarkable linguistic proficiency and have exhibited exceptional performance across a wide range of natural language processing tasks. Consequently, researchers have explored the efficacy of LLMs in evaluating written essays. [Han et al. 2024] conducted a notable investigation into the impact of prompt engineering on LLM performance in AES. Their findings underscore the significance of providing comprehensive context, particularly through the incorporation of scoring rubrics, to enhance the accuracy of LLM-based essay assessments. Building upon this foundation, [Kim and Jo 2024] proposed a novel approach that combines LLMs with comparative judgment (CJ) techniques for AES. Their method outperforms traditional rubric-based scoring when utilizing LLMs by employing zero-shot prompting to select between pairs of essays. This research highlights the potential of leveraging LLMs with human-like judgment processes for improved AES outcomes. These studies collectively underscore the dynamic and evolving landscape of AES research. While traditional methods and neural architectures have made significant contributions, the emergence of LLMs has introduced novel opportunities and challenges in this field.

Most traditional automated essay scoring (AES) approaches rely on a single prompt, as noted in [Mansour et al. 2024]. However, this approach has limitations, particularly when assessing essays with multiple criteria, such as those required for ENEM. To address this, we explore the use of multiple prompts tailored to different aspects of the essay. This multifaceted approach can provide a more comprehensive evaluation by separately addressing content, structure, grammar, and coherence. Further research by [Naismith et al. 2023] illustrates the impact of different prompt conditions on the performance of LLMs on AET tasks. They investigate three distinct prompt structures: rating then rationale (rating-first), rationale then rating (rationale-first), and rating only (rating-only). The findings suggest that the order and nature of information presented in the prompt can significantly affect the model's output. For our application, experimenting with these different prompt structures can help identify the most effective approach for essay evaluation.

3.5. AES in Brazilian Portuguese language

In recent years, the field of automated essay scoring (AES) in the context of Brazilian Portuguese language assessment. One of the pivotal works in this domain is the Essay-BR: a Brazilian Corpus of Essays [Marinho et al. 2021]. This dataset comprises essays written by Brazilian high school students, which were meticulously graded by human professionals adhering to the criteria of the ENEM exam. The availability of such a corpus has been instrumental in facilitating the research and development of AES systems tailored to the Brazilian educational context. Building upon this foundation, recent research has explored the application of advanced natural language processing techniques to automate the essay grading process for the ENEM exam. [Mayer 2023] presents a notable contribution in this area, leveraging the BERTimbau model, a transformer-based language model adapted for Portuguese. This study developed five independent systems designed to evaluate essays based on a competency matrix, aligning with the ENEM's assessment criteria. The researchers employed specialized learning techniques and investigated optimal hyperparameter configurations to enhance model efficacy. Although the hyperparameter optimization process proved ineffective, the use of fixed architectures demonstrated promising results, approximating the quality of human evaluations. Complementing the efforts in AES, [Moreira and Moura 2023] addresses a crucial aspect of the learning process: providing interpretable feedback to students. This research investigates interpretability methods for machine learning models in the context of AES for ENEM-style essays. The study highlights the effectiveness of LIME (Local Interpretable Model-Agnostic Explanations) in identifying crucial terms and trends that the model associates with the AES task. These insights play a vital role in understanding the strengths and weaknesses of the developed models, potentially enhancing the quality of feedback provided to students.

These works collectively represent significant progress in the field of AES for Brazilian Portuguese, addressing various aspects from dataset creation to model development and interpretability. They demonstrate the potential of AI-driven approaches to streamline the evaluation process while maintaining alignment with established educational standards. Furthermore, these studies pave the way for future research in enhancing the accuracy, efficiency, and interpretability of automated essay scoring systems, particularly in the context of large-scale national examinations like ENEM.

4. Methodology

This section outlines the methodological framework and technical approaches employed in developing and assessing the LLM-based solution for automated ENEM essay correction. By elucidating the data acquisition, the evaluated LLM model, the prompt design strategy, and evaluation procedures, we aim to comprehensively understand the research process and its underlying rationale.

4.1. Data Acquisition and Preparation

The Essay-BR extended corpus [Marinho et al. 2022] serves as the primary dataset for this study. This dataset comprises ENEM-style essays with expert evaluations aligned with the exam's reference matrix. It is an extended version of the original dataset [Marinho et al. 2021], containing over 6,000 essays, their respective themes, evaluations by competency, and the final score of each essay.

The corpus was constructed through the automated extraction of essay texts, competency-specific assessments, and overall scores from individual evaluation pages. This methodically curated dataset provides a solid empirical basis for training and validating our multifaceted, prompt-based essay assessment system, enabling a granular analysis of ENEM essay components.

Due to the costs associated with using LLMs, we utilized a random sample of 500 essays from this dataset, covering different themes.

4.2. Model

Given the complexity and nuances of essay evaluation, we selected GPT-4o-mini as the core LLM architecture for our solution. This lightweight variant of the state-of-the-art GPT-4o model excels in understanding and generating human-quality text, making it an ideal candidate for the task. GPT-4o-mini is a streamlined iteration of GPT-4o, which builds upon the GPT-4 architecture initially introduced by gpt4. GPT-4 demonstrated human-level performance on various professional and academic benchmarks, scoring in the top 10% of a simulated bar exam.

GPT-4o-mini maintains the robust foundation of GPT-4o while offering optimizations that make it more resource-efficient without compromising its text, reasoning, and coding intelligence capabilities. It continues to set high standards in multilingual, audio, and vision capabilities, ensuring comprehensive and nuanced essay evaluation [OpenAI 2024].

4.3. Prompt design

Prompt engineering and chaining were employed to develop a series of interconnected prompts that mimic the ENEM evaluation process. The prompt engineering process was primarily based on analyzing the ENEM essay correction manuals, closely examining the rules, examples, and guidelines, and then adapting them to the context of prompt creation. Drawing inspiration from these established correction manuals, we iteratively refined various prompt engineering techniques to construct concise prompts targeting specific sub-steps of the evaluation. These granular prompts were then sequenced to mirror the holistic evaluation flow. We utilized PromptFlow as a robust platform for orchestrating the intricate interplay of prompts that underpin the essay evaluation process.

The resulting system generates a comprehensive score encompassing global and competency-based assessments and provides detailed feedback aligned with the evaluation criteria.

4.4. Evaluation Framework

To assess model performance, we utilize:

- **Proportion of Exact Agreement (PEA)**: The proportion of essays where the model and human raters assign identical scores.

  $$PEA = \frac{\sum_{i=1}^{n} 1(y_i = \hat{y}_i)}{n}$$

  where y_i is the human score, $\hat{y}_i$ is the model score, n is the total number of essays, and 1 is the indicator function.

- **Proportion of Adjacent Agreement (PAA)**: The proportion of essays where the model and human raters assign scores differing by a small, predefined value (here, up to 100 for the final grade and up to 80 for the competency scores, as in the official ENEM essay evaluation criteria).

  $$PAA = \frac{\sum_{i=1}^{n} 1(|y_i - \hat{y}_i| \leq \delta)}{n}$$

  where δ is the predefined threshold (100 for final grade, 80 for competency scores).

- **Pearson Correlation Coefficient (PCC)**: A measure of the linear relationship between model and human rater scores, indicating the strength and direction of the association.

  $$PCC = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}}$$

  where $\bar{y}$ is the mean of human scores and $\bar{\hat{y}}$ is the mean of model scores.

These metrics enable a comprehensive evaluation of the model's capacity to replicate human rater judgments, providing insights into precision and consistency. It is important to note that in the context of ENEM essay scoring, where human raters are permitted a degree of discrepancy, Adjacent Agreement (AA) holds particular significance. This metric reflects the model's ability to generate scores within the acceptable range defined by ENEM's guidelines, aligning with the real-world evaluation practices of the exam.

These metrics enable a comprehensive evaluation of the model's capacity to replicate human rater judgments, providing insights into precision and consistency.

5. Proposed framework

This section delineates the architectural framework of the LLM-based system designed for the automated correction of ENEM essays. The overarching task was meticulously decomposed into 24 distinct prompts to address the multifaceted nature of the ENEM essay assessment. These prompts, crafted to leverage the power of chain-of-thought reasoning, in-context learning, and the principle of prompt chaining and other prompting thechniques collaborate synergistically to provide a granular analysis of the essay.

Our framework employs six independent components, each dedicated to evaluating a specific aspect of the ENEM essay. Five components assess individual competencies (C1-C5), while a sixth component aggregates the results into a final score and comprehensive feedback. Each component utilizes a prompt chain tailored to its specific task. The number of prompts and their key functions are summarized in Table 1.

Table 1. Competency Evaluator Components: Prompt Structure and Function

Competency	Number of Prompts	Key Functions of Prompts
C1	3	Grammatical error identification, in-depth error analysis, overall grammatical competency evaluation
C2	4	Analysis of knowledge areas, adherence to textual type, topic comprehension, overall competency evaluation
C3	4	Analysis of structural flow, textual plan evaluation, argument scrutiny, overall competency evaluation
C4	5	Inter-paragraph cohesion analysis, word repetition analysis, key opinion extraction, transition word assessment, overall competency evaluation
C5	7	Multi-dimensional intervention proposal analysis (action, agent, means/mode, effect, detail), identification of the most comprehensive proposal, overall competency evaluation

Figure 1 illustrates a global view of the architectural framework of the LLM-based system, providing an overview of the interrelations among its components. In contrast, Figure 2 offers a detailed examination of each individual component, highlighting their specific roles and functionalities in the automated correction process of ENEM essays. This distinction emphasizes how the overarching architecture integrates various assessment dimensions while allowing for a granular analysis of each competency.

This modular and hierarchical framework enables a thorough, multifaceted evaluation of ENEM essays, leveraging the capabilities of LLMs through strategically designed prompts and their orchestrated interaction. This approach enhances comprehension of the framework by revealing the intricate web of connections, thus providing a holistic view of how each prompt contributes to the overall system's functionality.

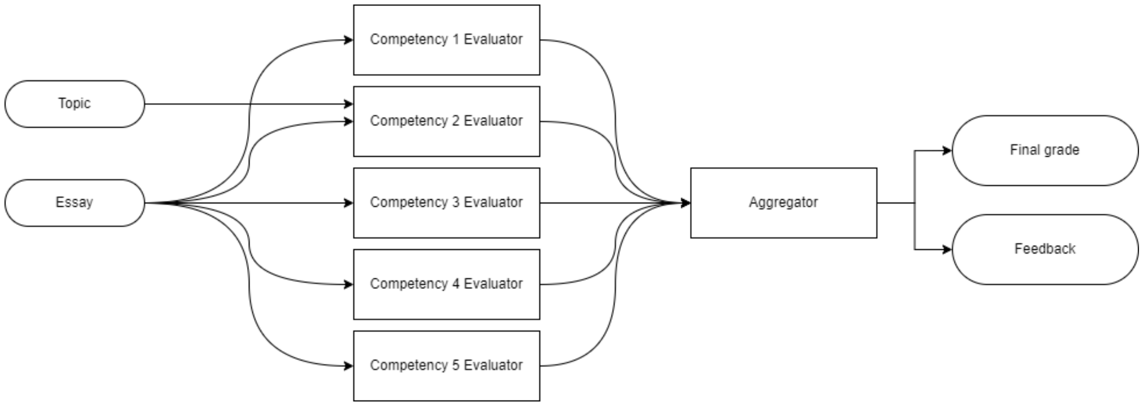
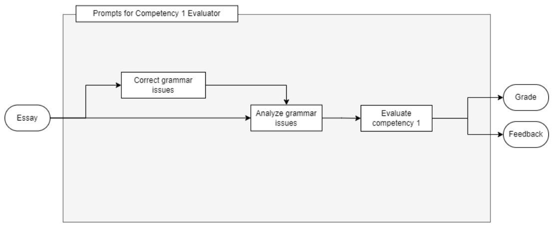
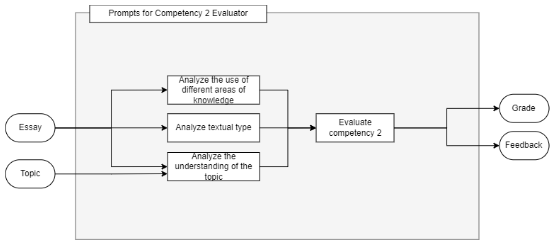


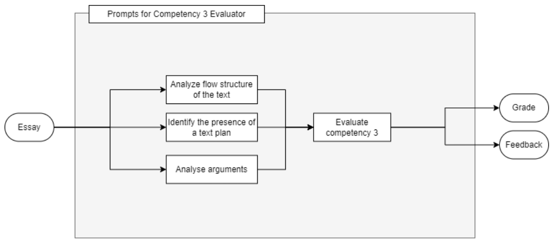
Figure 1. Broad view of the proposed framework



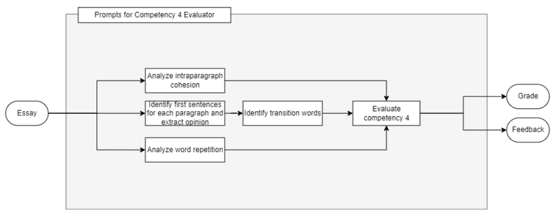
(a) Detailed view of the Competency 1 Evaluator



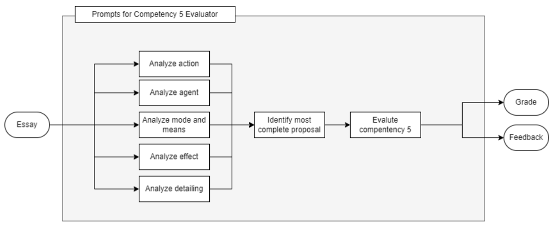
(b) Detailed view of the Competency 2 Evaluator



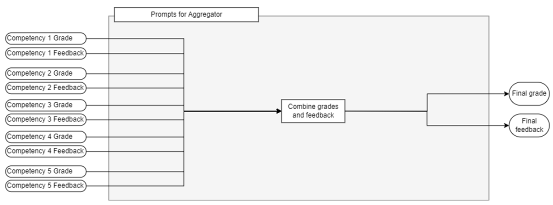
(c) Detailed view of the Competency 3 Evaluator



(d) Detailed view of the Competency 4 Evaluator



(e) Detailed view of the Competency 5 Evaluator



(f) Detailed view of the Aggregator

Figure 2. Overview of the framework components

5.1. The Art of Building the Correct Prompts

In the realm of developing an LLM-based solution for automated correction of ENEM essays, crafting effective prompts is akin to a delicate art form. This process requires continuous experimentation and a profound understanding of the context in which one is working.

While the INEP's Participant's Handbook provides a six-level performance scale, the descriptors used are often subjective and open to interpretation. Terms like "very well" or ”detailed” lack objective definitions, making it challenging to create precise prompts. To address this ambiguity, we turn to the Correction Manual for Competency 5, which offers a more structured approach. By combining these two sources of information and structuring prompts around these points, we can guide the LLM to perform a more objective and consistent evaluation. This approach bridges the gap between the subjective language of the performance levels and the more concrete elements outlined in the correction manual. The art of prompt crafting in this context involves finding the right balance between providing specific instructions and allowing the model enough flexibility to handle the diverse range of student responses. It requires iterative refinement, testing with various essay samples, and continuous adjustment based on the model's performance. All the prompts used in our prompt chaining flow combine various prompt engineering techniques and largely follow a well-defined structure. Below, we will illustrate this work using one of the prompts from the Competency 5 Evaluator as an example: the 'Analyze agent' prompt, which is responsible for identifying the presence of an explicit agent in the candidate's written intervention proposal. [1]

5.1.1. Deep dive into an example: 'Analyze agent' prompt

1. **Role and Task Definition:** The architecture initiates with a precise definition of the model's role and the specific task at hand. This foundational step establishes the context and frames the model's approach for subsequent analytical processes.

   > *”You are a specialist in grading ENEM essays, and you are evaluating an essay from a candidate.”*

2. **Control Flow Management:** The design incorporates multiple control flow statements embedded within the prompt, directing the model's attention and the sequence of processing steps. This strategic approach ensures that the model adheres to a structured path of analysis.

   > *”You will be provided with this essay.”*

3. **Context Structuring:** A comprehensive context is provided to the model, encompassing relevant background information and specific evaluation criteria necessary for the task.

   > *”In the guidelines for this essay exam, there is a specific section on evaluating the proposed intervention presented. For a proposal to be considered concrete, it must include five mandatory elements: action, agent, means/mode, effect, and details.”*

4. **Explicit Instruction:** Clear and specific instructions are included to focus the model's attention on identifying the particular element under evaluation, namely the *agent*.

[1] The prompts here are written in English. However, the original prompts were written and tested in Portuguese.

> *"Your task is solely to identify one of these agents: the AGENT. The question to be answered to identify the agent of the proposed action is 'Who performs it?'"*

5. **In-Context Learning through Exemplification:** The architecture employs exemplification to illustrate both valid and invalid responses, facilitating the model's learning process and application of patterns to new instances.

> *"In the two texts below, examples of valid agents will be highlighted by triple quotes:*
> *1. In this way, """content providers""" should use filters to control the spread of false information.*
> *2. Furthermore, it is important that the """State, in partnership with schools""" promote informative lectures for the entire population."*

6. **Handling Edge Cases:** Provisions are made for potential edge cases, such as null agents, with explicit guidelines for addressing these scenarios to ensure robustness in the model's responses.

> *"Important: consider only valid agents (non-null agents). When the agent is expressed by terms that do not allow for the precise identification of the social actor responsible for executing the action, it should be considered a "null element."*
>
> *Examples of null agents:*
> *1. Someone, no one, some, a few, some and others, you;*
> *2. Verbs in the imperative mood - as long as there is no vocative."*

7. **Instruction Reinforcement:** To counteract recency bias, key instructions are reiterated at the end of the prompt, reinforcing critical directives to prevent oversight.

> *"Based on this context, your task is to identify the presence of the ACTION element in the candidate's text."*

8. **Chain of Thought Prompting:** The architecture encourages a step-by-step reasoning process, promoting transparency and aiding in error detection through explicit reasoning requirements.

> *"Think step-by-step and explain the process before answering."*

9. **Multiple Intervention Handling:** The design accounts for the possibility of multiple intervention proposals within a single text, offering instructions for managing such cases to ensure thorough evaluation.

> *"NOTE: There may be more than one intervention proposal in the paragraph; in this case, each one may or may not have a specific agent."*

10. **Structured Output Format:** A structured JSON format is specified for the model's output, ensuring consistency and facilitating subsequent data processing and interpretation.

”Respond in the following format:
{
”first_proposal”: {
”rationale”: ”...”,
”agent”: ”...”
},
”second_proposal”: {
”rationale”: ”...”,
”agent”: ”...”
}
}
If there is no valid agent, the ”agent” field should have the value ’-’”.

5.1.2. ’Analyze agent’ prompt in practice

To further illustrate the functionality of our prompt chain, we provide an example of the system’s interaction with user input, specifically the ’Analyze agent’ prompt in the Competency 5 evaluator. The developed prompts are used as system roles, defining the interaction’s context, rules, and constraints. This system role serves to guide the model by establishing task-relevant guidelines and domain-specific knowledge.

system

You are a specialist in grading ENEM essays, and you are evaluating an essay from a candidate.
You will be provided with this essay.
In the guidelines for this essay exam, there is a specific section on evaluating the proposed intervention presented. For a proposal to be considered concrete, it must include 5 mandatory elements: action, agent, means/mode, effect, and details.
Your task is solely to identify one of these agents: the AGENT. The question to be answered to identify the agent of the proposed action is ’Who performs it?’
In the two texts below, examples of valid agents will be highlighted by triple quotes:
1. In this way, ”””content providers””” should use filters to control the spread of false information.
2. Furthermore, it is important that the ”””State, in partnership with schools””” promote informative lectures for the entire population.
Important: consider only valid agents (non-null agents). When the agent is expressed by terms that do not allow for the precise identification of the social actor responsible for executing the action, it should be considered a null element.

Examples of null agents:
1. Someone, no one, some, a few, some and others, you;
2. Verbs in the imperative mood - as long as there is no vocative.
Based on this context, your task is to identify the presence of the ACTION element in the candidate’s text.
Think step-by-step and explain the process before answering.
NOTE: There may be more than one intervention proposal in the paragraph; in this case, each one may or may not have a specific agent.

”Respond in the following format:
{
”first_proposal”: {
”rationale”: ”...”,
”agent”: ”...”
},
”second_proposal”: {
”rationale”: ”...”,
”agent”: ”...”
}
}
If there is no valid agent, the ”agent” field should have the value ’-’.”

The user role represents the human interacting with the language model, initiating queries, providing input, and steering the conversation. As defined in the system prompt and as seen in Figure 2 (e), the user role is composed entirely of the essay to be evaluated.

Below we will use as an example a translated version into English of one of the essays that make up the dataset used. It is expected that the assistant correctly identifies the agent, which, in this essay, sent by the user prompt, is the Ministry of Health.

user

The information and psychological support should constitute a fundamental exercise of citizenship for the development of a society based on collective well-being. However, the stigmas associated with mental illnesses in Brazilian society demonstrate the fragility and inefficiency of exercising this principle in Brazil. Such inefficiency occurs both due to the state's negligence regarding mental health and the lack of a socio-emotional perspective in education.

Faced with this scenario, it is important to highlight the state's ignorance in the psychological aspects of its citizens. According to article 6 of the Federal Constitution, the most important legal document in the country, every citizen has the right to health. However, the absence of action promoting this constitutional right by the national government is undeniable. This situation is evidenced in the precarious psychological support offered by public health and, consequently, guarantees the condition of sub-citizenship for many individuals afflicted by psychopathologies.

Furthermore, added to this unempathetic state behavior, there is also the absence of educational models that deconstruct the prejudices related to mental health problems in Brazil. Understanding this assertion means considering that the current educational model perpetuates the atrocious and labeling thinking about the psychological particularities of each individual. In light of this, it is regrettable the absence of a humanistic perspective, as evidenced by the Brazilian philosopher, Augusto Cury, reinforcing a culture of ignorance and disrespect for the socio-emotional, psychological, and mental difficulties of others, making it essential to dissolve the common sense regarding mental illnesses in the Brazilian reality.

Therefore, it is concluded that the exclusionary reality surrounding mental illnesses is intrinsically linked to problematic state action and educational apathy. For this reason, measures must be taken to reverse this situation. That being said, it is up to the Ministry of Health, the regulatory body for social guarantees in the field of public health, to create a National Psychological Support Program, which will disseminate information and provide care in a free and effective manner - such as psychoclinical and psychiatric support - through investments that foster quality socio-emotional education, so that this stigmatizing society is transformed into a democratic community, as elucidated by Cury, based on human understanding and social well-being.

The assistant role, embodied by the language model itself, interprets the user's input in light of the system-defined parameters and generates appropriate responses. This role aims to provide helpful, accurate, and contextually relevant information or perform requested tasks within the boundaries set by the system prompt, which in this case is a JSON containing the rationale behind the decision and also the identified agent.

assistant

```
{
    "first_proposal": {
        "rationale": "The candidate suggests that the Ministry of Health create a National Psychological Support Program to disseminate information and provide free and effective care, in addition to promoting investments in quality socio-emotional education.",
        "agent": "Ministry of Health"
    }
}
```

This example shows how the system correctly identifies the agent responsible for the proposed intervention in the provided essay. As noticed, the model correctly identified 'Ministry of Health' as the agent, as expected. In addition, the rationale field provides context for the identified agent, demonstrating the model's capacity not only to identify the agent but also to understand the context of its action within the proposal. The structured JSON output ensures easy parsing and integration into the broader evaluation framework.

6. Results and Discussion

Our LLM-based model demonstrated promising performance in predicting ENEM essay scores, both for individual competencies and the overall final score. The results reveal varying degrees of accuracy across different evaluation metrics and competencies.

Table 2. Performance Metrics for Each Competency and Final Score

Competency	Exact Agreement	Adjacent Agreement	Pearson Correlation
C1	0.4444	1.0000	0.2431
C2	0.6111	1.0000	0.5684
C3	0.5370	1.0000	0.5613
C4	0.5370	1.0000	0.4591
C5	0.3333	1.0000	0.6357
Final Score	0.2407	1.0000	0.9048

6.1. Overall Performance

As shown in Table 2, our model achieved varying levels of performance across different metrics and competencies.

Remarkably, our model achieved a 100% Adjacent Agreement rate for both the final score and all individual competencies. This result holds particular significance in the context of ENEM essay scoring, where human raters themselves are allowed a discrepancy of up to 100 points in their evaluations. Mirroring this real-world evaluation practice, we prioritize Adjacent Agreement (AA) as a key indicator of the model's ability to produce scores that align with the acceptable range defined by ENEM's guidelines. While Proportion of Exact Agreement (PEA) remains a valuable metric for assessing precision, the high tolerance for variation in human scoring necessitates prioritizing AA as a primary measure of the model's practical relevance and ability to generate consistent and reliable scores.

For the final essay score, our model achieved an exact agreement rate of 24.07% with human raters. While this may seem modest, it's important to consider it in the context of the perfect adjacent agreement rate. The Pearson correlation coefficient between the model and human rater scores for the final score was 0.9048, revealing a very strong positive linear relationship. This high correlation indicates that the model effectively captures the underlying rating criteria and aligns closely with human judgment in assessing overall essay quality.

6.2. Comparison with Mayer 2023

To contextualize our results within the field of automated essay scoring for ENEM, we compared our model's performance with [Mayer 2023]. Table 3 presents a comparison of Proportion of Adjacent Agreement (PAA) and Proportion of Exact Agreement (PEA) across different competencies for both solutions.

Table 3. Comparison of Our Solution and Essay-br across different competencies

Competency	Method	PAA	PEA
C1	Our solution	**1.000**	0.444
	Mayer 2023	0.992	**0.683**
C2	Our solution	**1.000**	**0.611**
	Mayer 2023	0.984	0.572
C3	Our solution	**1.000**	0.537
	Mayer 2023	0.990	**0.593**
C4	Our solution	**1.000**	**0.537**
	Mayer 2023	0.990	0.513
C5	Our solution	**1.000**	0.333
	Mayer 2023	0.970	**0.494**

As evident from Table 3, our solution consistently outperforms [Mayer 2023] in terms of Adjacent Agreement (PAA) across all competencies. Our model achieves a perfect PAA score of 1.000 for all five competencies, compared to Essay-br's scores ranging from 0.970 to 0.992. This superior performance in PAA is particularly significant given that PAA is considered the principal metric in essay scoring, as discussed earlier.

In terms of Proportion of Exact Agreement (PEA), the results are mixed. Our solution performs better in competencies C2 and C4, while Essay-br shows higher PEA for competencies C1, C3, and C5. Notably, our model's PEA for C2 (0.611) is significantly higher than Essay-br's (0.572), indicating a particular strength in assessing this competency.

It's important to highlight the fundamental differences in methodology between our approach and Essay-br. The Essay-br methodology employs a fine-tuning approach, which, while effective, may inadvertently learn and perpetuate biases present in the human-scored training data. In contrast, our approach utilizes prompt engineering techniques that are entirely based on the official ENEM essay scoring rubric. This distinction is crucial, as our method aims to adhere strictly to the established scoring criteria, potentially reducing the risk of perpetuating human biases in the scoring process.

6.3. Feedback

Furthermore, our solution offers a significant advantage beyond just providing scores. As part of the evaluation process, our model generates meaningful feedback for students, explaining the rationale behind each score and offering specific suggestions for improvement. This feature transforms the automated scoring system from a mere assessment tool into a valuable learning aid, providing students with actionable insights to enhance their writing skills.

The ability to provide detailed feedback, combined with our model's strong performance in Adjacent Agreement, positions our solution as a comprehensive tool for ENEM essay evaluation. While [Mayer 2023] demonstrates strengths in exact agreement for certain competencies, our approach offers a more holistic solution that not only scores essays accurately but also supports student learning and improvement.

Below is a translated example of the feedback generated by competency for one of the essays in our evaluation database:

- **Competency 1 - 120**
  The candidate presented significant flaws in the grammatical structure of the text, with truncation of periods and juxtaposition of clauses, as well as deviations related to punctuation, agreement, and verb tense. The score of 120 is appropriate, as it reflects the number of errors and deviations observed, which fit the conditions established for this score.
- **Competency 2 - 120**
  The text's argumentation is considered superficial, with an attempt to address the topic of functional illiteracy, but without the necessary depth. The lack of a clear introduction, a well-defined thesis, and a conclusion that synthesizes the arguments compromises the effectiveness of the text. The score of 120 is appropriate, as it demonstrates a medium-level mastery of argumentative writing, but without achieving the required depth.
- **Competency 3 - 120**
  Although the text presents a clear idea in addressing the topic, the articulation of the ideas is deficient. The introduction does not provide a clear statement of the points to be discussed, and the development is more expository than argumentative. These flaws weaken the candidate's defense of their point of view, justifying the score of 120.
- **Competency 4 - 120**
  The internal cohesion of the text is reasonable, with cohesive elements present, but the absence of argumentative operators between paragraphs limits the articulation of the argument. The presence of an inappropriate cohesive element is also a negative point. The score of 120 reflects the regularity in internal cohesion, but the lack of connectors between ideas compromises the text's flow.
- **Competency 5 - 160**
  The proposal presented is robust, with four valid elements, although the absence of an agent is a limitation. The proposal does not violate human rights and is relevant to the topic, which justifies the score of 160, based on the first proposal, which is the most complete. The second proposal, on the other hand, is considered weak and did not contribute to the final score.

7. Conclusion

This study has made significant strides in addressing the challenge of automated essay scoring (AES) for the Brazilian National High School Exam (ENEM), filling a crucial gap in LLM-based AES research for Portuguese datasets. Our novel approach, leveraging advanced LLM architectures and innovative prompt engineering techniques, has demonstrated remarkable effectiveness in evaluating ENEM essays according to established criteria.

The results of our study are particularly promising. We achieved 100% adjacent agreement across all competencies and the final score, outperforming existing methods in this critical metric. The strong Pearson correlation of 0.9048 with human raters for the final score further validates the reliability of our model. While exact agreement rates varied across competencies, our approach showed comparable or superior performance to existing methods in several areas.

A key innovation of our research lies in the systematic chain of prompts we developed, which not only evaluates essays but also provides meaningful feedback to students. This feature transforms our solution from a mere scoring tool into a comprehensive learning aid, addressing a significant limitation of previous AES approaches.

Our methodology's adherence to the official ENEM essay scoring rubric through prompt engineering techniques offers a distinct advantage over fine-tuning approaches. By basing our evaluation strictly on established criteria, we potentially reduce the risk of perpetuating biases that may be present in human-scored training data.

Another advantage of the proposed solution is the provision of feedback alongside the essay evaluation. This feature transforms the system into more than just a grading tool, offering students a valuable resource to enhance their writing skills based on clear and detailed criteria.

However, we acknowledge that there is room for improvement, particularly in terms of exact agreement rates for some competencies. Future research could focus on enhancing these aspects while maintaining the strengths of our current model.

In conclusion, this study contributes significantly to the field of AES by demonstrating the effectiveness of prompt engineering in adhering to established scoring criteria, providing valuable feedback, and achieving high adjacent agreement rates. Our approach offers a promising solution for large-scale educational assessments in Portuguese, with potential applications beyond the ENEM context. As natural language processing technologies continue to advance, the integration of such systems into educational assessment promises to enhance efficiency, consistency, and educational value in essay evaluation processes.

8. Limitations

While our proposed LLM-based solution for automated ENEM essay evaluation demonstrates promising results, it is essential to acknowledge its limitations:

- **Limited Dataset Size:** Our evaluation was conducted on a sample of 500 essays from the Essay-br dataset due to the computational cost associated with LLMs. While this sample size allowed for a comprehensive analysis, utilizing the entire dataset could potentially yield more robust and generalizable results. Future work should explore evaluating the model on a larger scale.
- **Focus on Adjacent Agreement:** Our primary evaluation metric was Proportion of Adjacent Agreement (PAA), prioritizing the model's ability to score within an acceptable range of human raters. While this aligns with ENEM's scoring guidelines, focusing solely on AA might overshadow potential discrepancies in exact score prediction. Further investigation into improving Proportion of Exact Agreement (PEA) rates is crucial.
- **Generalizability to Other Essay Genres:** While promising for ENEM essays, our solution's generalizability to other essay genres requires further investigation. The current prompt engineering approach is tailored to the specific structure and criteria of ENEM essays. Adapting to different genres, such as persuasive or even another essay style, might necessitate modifying the prompt structure, incorporating genre-specific instructions, or even exploring alternative LLM architectures better suited for diverse writing styles.

- **Dependence on Prompt Engineering:** Our approach heavily relies on meticulously crafted prompts to guide the LLM's evaluation process. While this allows for transparency and control over the evaluation criteria, it also introduces a level of subjectivity in the prompt design itself. Different prompt formulations could potentially lead to variations in model performance.
- **Lack of Real-World Implementation:** Our study focused on evaluating the model's performance on a pre-existing dataset. Implementing this solution in a real-world ENEM scoring scenario would require addressing additional challenges, such as scalability, security, and ethical considerations related to automated assessment.

Addressing these limitations will be crucial for refining the model and ensuring its practical applicability in large-scale educational assessments. Future research should focus on expanding the dataset, exploring techniques to improve EA rates, and investigating the model's generalizability to other essay genres. Additionally, real-world implementation studies are necessary to assess the feasibility and ethical implications of deploying such a system in a high-stakes testing environment.

References

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N. S., Chen, A. S., Creel, K. A., Davis, J., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L. E., Goel, K., Goodman, N. D., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T. F., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M. S., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X. L., Li, X., Ma, T., Malik, A., Manning, C. D., Mirchandani, S. P., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J. C., Nilforoshan, H., Nyarko, J. F., Ogut, G., Orr, L., Papadimitriou, I., Park, J. S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y. H., Ruiz, C., Ryan, J., R'e, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K. P., Tamkin, A., Taori, R., Thomas, A. W., Tramèr, F., Wang, R. E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S. M., Yasunaga, M., You, J., Zaharia, M. A., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., and Liang, P. (2021). On the opportunities and risks of foundation models. *ArXiv*.

Brasil, M. d. E. 3,9 milhões estão inscritos no enem 2023 — gov.br. `https://www.gov.br/inep/pt-br/assuntos/noticias/enem/3-9-milhoes-estao-inscritos-no-enem-2023`. [Accessed 06-05-2024].

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Chen, H. and He, B. (2013). Automated essay scoring by maximizing human-machine agreement. In Yarowsky, D., Baldwin, T., Korhonen, A., Livescu, K., and Bethard, S., editors, *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1741–1752, Seattle, Washington, USA. Association for Computational Linguistics.

Dong, F., Zhang, Y., and Yang, J. (2017). Attention-based recurrent convolutional neural network for automatic essay scoring. In Levy, R. and Specia, L., editors, *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 153–162, Vancouver, Canada. Association for Computational Linguistics.

Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B., Sun, X., Xu, J., Li, L., and Sui, Z. (2023). A survey for in-context learning. *ArXiv*, abs/2301.00234.

Han, J., Yoo, H., Myung, J., Kim, M., Lim, H., Kim, Y., Lee, T. Y., Hong, H., Kim, J., Ahn, S.-Y., and Oh, A. (2024). Llm-as-a-tutor in efl writing education: Focusing on evaluation of student-llm interaction.

INEP (2020a). Competência 1 - demonstrar domínio da modalidade escrita formal da língua portuguesa. `https://download.inep.gov.br/educacao_basica/enem/downloads/2020/Competencia_1.pdf`.

INEP (2020b). Competência 2 - compreender a proposta de redação e aplicar conceitos das várias áreas de conhecimento para desenvolver o tema, dentro dos limites estruturais do texto dissertativo-argumentativo em prosa. `https://download.inep.gov.br/educacao_basica/enem/downloads/2020/Competencia_2.pdf`.

INEP (2020c). Competência 3 - selecionar, relacionar, organizar e interpretar informações, fatos, opiniões e argumentos em defesa de um ponto de vista. `https://download.inep.gov.br/educacao_basica/enem/downloads/2020/Competencia_3.pdf`.

INEP (2020d). Competência 4 - demonstrar conhecimento dos mecanismos linguísticos necessários para a construção da argumentação. `https://download.inep.gov.br/educacao_basica/enem/downloads/2020/Competencia_4.pdf`.

INEP (2020e). Competência 5 - elaborar proposta de intervenção para o problema abordado, respeitando os direitos humanos. `https://download.inep.gov.br/educacao_basica/enem/downloads/2020/Competencia_5.pdf`.

INEP (2023). A redação no enem 2023: Cartilha do participante. `https://download.inep.gov.br/publicacoes/institucionais/avaliacoes_e_exames_da_educacao_basica/a_redacao_no_enem_2023_cartilha_do_participante.pdf`.

Kim, S. and Jo, M. (2024). Is gpt-4 alone sufficient for automated essay scoring?: A comparative judgment approach based on rater cognition. *arXiv*.

Mansour, W. A., Albatarni, S., Eltanbouly, S., and Elsayed, T. (2024). Can large language models automatically score proficiency of written essays? In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2777–2786, Torino, Italia. ELRA and ICCL.

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

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

Mayer, J. L. S. (2023). Avaliação automática de redações no modelo do enem por meio do fine-tuning do bertimbau.

Moreira, G. and Moura, R. (2023). Explicabilidade de modelos para avaliação automática de redações. In *Anais do XVI Encontro Unificado de Computação do Piauí*, pages 113–120, Porto Alegre, RS, Brasil. SBC.

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@ACL 2023, Toronto, Canada, 13 July 2023*, pages 394–403. Association for Computational Linguistics.

OpenAI (2024). Gpt-4o mini: Advancing cost-efficient intelligence. `https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/`. Accessed: 2024-08-28.

Phandi, P., Chai, K. M. A., and Ng, H. T. (2015). Flexible domain adaptation for automated essay scoring using correlated linear regression. In Màrquez, L., Callison-Burch, C., and Su, J., editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 431–439, Lisbon, Portugal. Association for Computational Linguistics.

Prado, D. d. F. and Morato, R. A. (2017). A redação do enem como gênero textual-discursivo: uma breve reflexão. *Cadernos CESPUC de Pesquisa Série Ensaios*, (29):205–219.

Silva, A. and de Lima Cavalcante, F. (2023). O gênero redação do enem. *Revista Eletrônica de Estudos Integrados em Discurso e Argumentação*, 23(2):51–70.

Sun, S., Yuan, R., Cao, Z., Li, W., and Liu, P. (2024). Prompt chaining or stepwise prompt? refinement in text summarization. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Findings of the Association for Computational Linguistics ACL 2024*, pages 7551–7558, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. (2023). Chain-of-thought prompting elicits reasoning in large language models. `https://arxiv.org/abs/2201.11903`.

Wu, T., Jiang, E., Donsbach, A., Gray, J., Molina, A., Terry, M., and Cai, C. J. (2022). Promptchainer: Chaining large language model prompts through visual programming. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI EA '22, New York, NY, USA. Association for Computing Machinery.

Wu, Y., Henriksson, A., Nouri, J., Duneld, M., and Li, X. (2023). Beyond benchmarks: Spotting key topical sentences while improving automated essay scoring performance with topic-aware bert. *Electronics*, 12(1).

Xie, J., Cai, K., Kong, L., Zhou, J., and Qu, W. (2022). Automated essay scoring via pairwise contrastive regression. In Calzolari, N., Huang, C.-R., Kim, H., Pustejovsky, J., Wanner, L., Choi, K.-S., Ryu, P.-M., Chen, H.-H., Donatelli, L., Ji, H., Kurohashi, S., Paggio, P., Xue, N., Kim, S., Hahm, Y., He, Z., Lee, T. K., Santus, E., Bond, F., and Na, S.-H., editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2724–2733, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.