Automated Reformulation of Argumentative Essays to Improve Argument Organization and Development*

Naomi James Sutcliffe de Moraes *and* Denis Deratani Mauá
Institute of Mathematics, Statistics and Computer Science
University of São Paulo
São Paulo, Brazil
{nmoraes, ddm}@ime.usp.br

Abstract

This work presents a study of automated reformulation of argumentative essays written by college-bound native speakers of Brazilian Portuguese as a form of pedagogical feedback. We first evaluate the feasibility of using large language models (LLMs) to score argument quality with respect to three criteria: the defense of a point of view, organization, and development. We then employ an LLM to provide a reformulated version of the essay as feedback. As we discuss, the main challenge is to constrain the automated feedback to address only argument quality, rather than improving other aspects such as spelling or cohesion, and to modify the essay as little as possible. We achieve levels of agreement in automatic essay scoring comparable to human inter-rater agreement metrics, while increasing explainability. Instructing the LLM to add argument support (facts, examples, etc.) was the best way to get non-superficial changes to the arguments, and it was able to add true examples and facts to the essays even without being provided with background information on the topic.

1 Introduction

The ability to write argumentative texts is an important skill for college-bound students to master because it requires mastery of formal writing, argument structure, and critical thinking (Karbach, 1987). Indeed, argumentative essay writing is a common element of large-scale standardized assessments worldwide (e.g. ENEM, TOEFL, LSAT).

In order to improve student essay writing skills, instructors provide formative feedback designed to modify student thinking or behavior (Shute, 2008). Recent studies suggest that the type of feedback has a substantial effect on student learning (Santos et al., 2010; Tocalli-Beller and Swain, 2005). In corrective feedback, the instructor identifies errors and makes suggestions, whereas in reformulation the instructor rewrites the text to improve it and provides the reformulated text as feedback. Kim and Bowles (2019) found that when students received reformulation feedback they better comprehended their structural or conceptual errors, whereas when they received corrective feedback they focused more on superficial mistakes.

Instructor feedback on student writing is, however, time-consuming to produce, non-immediate, inconsistent and difficult to scale. Consequently, researchers have turned to automated feedback tools, with a strong focus on Automated Essay Scoring (AES) to assign grades based on predefined criteria such as grammar, style, and content (Klebanov and Madnani, 2022). Since human scoring is inherently subjective, AES model performance is typically evaluated by comparing the agreement between automated predictions and reference human scores to the inter-rater agreement between two human graders. Recent studies show that modern large language models (LLMs) achieve inter-rater agreement comparable to that of human graders, even without task-specific supervised training (Silveira et al., 2024; Barbosa et al., 2025).

While LLMs are now routinely used to reformulate texts of all kinds, providing relevant, hallucination-free output remains challenging (Ji et al., 2023). Existing LLMs also have a tendency to change texts drastically and not specifically in the way they are prompted to do so (Shu et al., 2024). Restricting an LLM to modifying just one characteristic of a text is not straightforward (Zhang et al., 2023; Sun et al., 2023).

Using an annotated corpus of argumentative essays written in Brazilian Portuguese, this work presents an investigation of both automated scoring and automated reformulation as coupled components of a feedback generation pipeline. The first

*This work was funded in part by FAPESP grant No. 2022/02937-9, CNPq grants No. 140211/2022-05 and 305136/2022-4 and CAPES Finance Code 001.

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 425–435

objective is to evaluate whether an LLM can assign numerical scores for three argument quality criteria in a manner consistent with human ratings, as measured by inter-rater agreement metrics.

While we use AES here as an internal measurement tool to evaluate Automated Essay Reformulation (AER) results—and not necessarily as an end goal—our results show that our three-criteria LLM-based scorer is competitive with state-of-the-art AES models. The second objective is to evaluate whether the LLM can be guided to provide a reformulated version of the essay as a reference—improving only the argument development and/or organization—without overwhelming the student with extraneous modifications. The empirical results show that while it is possible to guide the LLM to focus reformulation feedback on a given criterion, improvement in the score for one criterion usually correlates with improvement in the score for the other—they are not completely separable. Our human evaluator also noted differences in the way the LLM reformulated essays due to small changes in the prompt language that were not apparent from the score differences alone.

2 Related Work

We start by reviewing the most relevant related work.

2.1 Automated Essay Scoring

Recent work has explored the use of large language models for automated essay scoring (AES), either as standalone scoring systems or in combination with traditional machine learning methods. Mizumoto and Eguchi (2023) developed a hybrid AES system to score TOEFL essays written by non-native speakers of English. The system uses LLM-generated scores together with scores produced by feature-based supervised machine learning models. Notably, the authors did not perform prompt engineering or include examples in their prompts. Tang et al. (2024) investigated multi-dimensional scoring (organization, ideas, style and conventions) and the effects of changing LLM temperature for scoring narrative (non-argumentative) essays. More recently, researchers have explored the use of LLMs as direct scoring mechanisms for argumentative writing. For example, Wang (2024) used an LLM to calculate a holistic score for argumentative essays, investigating the effects of fine-tuning the LLM compared to zero-shot prompting.

A separate line of research approaches AES indirectly by first performing argument mining and then estimating scores from the extracted argumentative structure. Nyugen (2018) used a machine learning model to identify basic argument features such as number of claims and number of support relations. They then used a logistic regression model using these and related argument features to estimate a holistic score. Wachsmuth (2016) also used machine learning to perform argument mining and automated scoring. They first used supervised machine learning to classify sentences, then used the sequences of types of argument discourse units to estimate scores for four argument quality dimensions: organization, thesis clarity, prompt adherence and argument strength.

For Brazilian Portuguese, some work has been done on automated grading of the *Exame Nacional do Ensino Médio* (ENEM), which is a nationwide standardized exam used as part of the admission process by many universities in Brazil. An argumentative essay is a core component of the exam. The essay is evaluated with respect to five criteria: fluency, style, argumentation, cohesion and persuasiveness.

Several studies have explored AES methods for ENEM essays using feature-based machine learning and deep learning approaches (Fonseca et al., 2018; Silveira et al., 2024). Although these systems achieved reasonable inter-rater agreement scores, training data imbalance led to diminished accuracy for infrequent scores. Very recently, Barbosa et al. (2025) analyzed the possibility of using LLMs to score ENEM essays. For the argument quality trait, the best performing method (OpenAI GPT 4o model) achieved a QWK value of 0.57, which is considered moderate agreement with human scores. While the authors experimented with several prompting strategies, they did not engineer prompts for a specific trait (e.g. argument quality).

2.2 Automated Essay Reformulation

While LLMs are used widely to reformulate texts, only recently have they been considered as pedagogical tools for essay reformulation. Chen (2025) evaluated the potential of ChatGPT to reformulate TOEFL essays in a general manner, polishing the essay in terms of cohesion, syntax, and vocabulary. The study tested only three simple prompts and used ROUGE scores to evaluate meaning retention in the reformulated essay. They also calculated indices to measure syntactic complexity, clausal

complexity, lexical sophistication, lexical diversity and cohesion. In another study, Song et al. (2024) developed both AES and AER systems to evaluate the essays of Chinese third graders writing in English. The AER prompt was very simple: "Please revise the essay written by a student." Similar to our work here, they employ an AES system to measure score changes due to reformulation and they use the cosine similarity of text embeddings to evaluate to what extent the meaning of the original essays had been maintained. Both of the studies described above requested general revision of essays written by non-native speakers. To the best of our knowledge, no prior work restricts LLM-based reformulation specifically to argument quality.

Despite these advances, relatively little work has explored how prompt engineering can guide LLMs to evaluate specific rubric traits, such as argument quality, rather than producing only holistic or loosely defined multi-trait scores. In addition, the extent to which LLMs can generate controlled essay reformulations that improve a particular argumentative trait remains unclear. In this work, we address these questions by developing trait-focused prompts for AES and by evaluating LLM-generated reformulations using both automated scores and human qualitative assessment.

3 Methodology

We now present the data, prompt engineering techniques and evaluation protocol we use.

3.1 Performance Evaluation for AES

Quadratic Weighted Kappa (QWK) is the most widely adopted metric for measuring inter-rater agreement in automated essay scoring, as it accounts for the ordinal nature of score levels and penalizes large disagreements more strongly than small ones. It is computed as a chance-corrected normalization of the squared difference of levels, such that a QWK of 0 denotes agreement purely by chance, and 1 denotes complete agreement. QWK scores above 0.2 are considered above-chance agreement, scores in the range of 0.4 to 0.6 are considered moderate agreement, and scores higher than 0.6 are considered substantial agreement (Cohen, 1960). The Mean Absolute Error (MAE) is calculated as the average of absolute differences between ratings; hence it provides a direct linear interpretation in terms of score point deviation. The Root Mean Squared Error (RMSE) instead computes the average squared difference between ratings, and thus emphasizes larger differences. For this reason, we pay particular attention to the per-level MAE and RMSE values, preferring models that perform equally well for all score levels. Since all three metrics are sensitive to score distribution and the number of possible score levels (Doewes et al., 2023), comparisons across different datasets are in general unreliable unless the underlying score distributions are comparable.

3.2 Dataset

We use the corpus of 381 argumentative essays in Brazilian Portuguese organized by Silveira et al. (2024). The essays were taken from mock Brazilian ENEM exams held monthly on the UOL web portal over many years (Universo Online (UOL), 2020). Actual ENEM essays are not publicly available. Each essay was annotated by two expert human raters (who they refer to as rater A and rater B) with respect to each of five traits (sub-scores) on a six-point ordinal scale, following the official ENEM guidelines and rubrics (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), 2019).

We focus here solely on trait/sub-score 3, which we will refer to as the **argument quality score**. It measures how well the author "selects, relates, organizes, and interprets information, facts, opinions, and arguments in support of a point of view" (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), 2019). In addition to the quality of the argument, the official ENEM grading guidelines contain rules unrelated to argumentation that result in an essay having all sub-scores zeroed. Since these cases are unrelated to argument quality, we discarded the four zeroed essays from the dataset. We call the resulting dataset EA-AES (for Essay Argument-AES). None of the resulting essays received a score of 0 for argument quality.

The confusion matrix comparing the argument quality scores given by the two raters (named A and B) for the EA-AES dataset is shown in Table 1. The most common scores are 3 or 4, with few scores of 5. The mean absolute error (MAE) for rater B versus rater A is 0.73, the root mean squared error (RMSE) is 1.05 and the Quadratic Weighted Kappa (QWK) is 0.57 (considered a moderate agreement rate).

To avoid the ambiguity in evaluating the reformulation models caused by the discrepancy between the scores of raters A and B, we selected only the

essays given identical argument quality scores by both raters. We call this dataset Essay Argument Consensus AES (EAC-AES). As can be seen in the diagonal of the matrix in Table 1, the set contains 168 essays. We also selected a smaller subset of 20 essays from EA-AES (all of the essays from September, 2019) to be used in the AER investigation, to mitigate the burden of human supervision. This particular subset was selected because of its relatively flat score distribution.

The UOL web portal posts a topic each month, an introduction including an explanation of its controversial aspects, and an invitation for readers to write an argumentative text and submit it for grading. For each topic, background reading material similar to that provided on the real ENEM exam is provided (see Brazilian Ministry of Education (2024) for details on the real exam and the rubric). The background material consists of extracts from news articles or government webpages and, usually, links to the full article or webpage from which the short text was extracted. The background material provided on the same page as the topic will be referred to as the "short background" (average length 496 words) and the complete background material found by following the links on the topic page to read additional text will be referred to as the "long background" (average length 1,152 words).

3.3 Large Language Models

Autoregressive (a.k.a. causal) large language models (LLMs), such as the GPT family, are trained to predict the next token and therefore are naturally suited for text generation and controlled rewriting (Shu et al., 2024). In this study we use OpenAI's GPT-4o via zero-shot prompt engineering for both AES and AER. GPT-4o's generation capabilities make it especially suitable for producing reformulated essays and we mitigate known risks (hallucination, over-zealous rewriting) by using restrictive instructions in prompts (OpenAI, 2024). Furthermore, GPT-4o was shown to match the performance of supervised learning-based AES models when prompts are carefully designed (Barbosa et al., 2025). We used the OpenAI model "text-embedding-3-large" for the embeddings used to calculate cosine similarity.

		Rater A					
		0	1	2	3	4	5
Rater B	0	**0**	0	0	0	0	0
	1	0	**38**	6	12	2	0
	2	0	11	**9**	34	2	0
	3	0	11	4	**69**	24	1
	4	0	11	1	52	**45**	4
	5	0	3	0	3	33	**6**

Table 1: Confusion matrix comparing Rater A and Rater B argument quality scores in the EA-AES dataset.

Score	Description
0	Presents information, facts, and opinions unrelated to the topic and without defending a point of view.
1	Presents information, facts, and opinions that are only loosely related to the topic or are inconsistent and without defending a point of view.
2	Presents information, facts, and opinions related to the topic in defense of a point of view.

Table 2: Rubric for Stance Criteria

3.4 Human-Like AES

For the first objective, determining if an LLM can accurately assign a numerical argument quality score following human grader rubrics, we considered three independent argument quality criteria inspired by the official ENEM grader's guideline:

- *Stance*: the presence and defense of a position
- *Organization*: the organization of supporting content in a logical, hierarchical manner
- *Development*: the development of the arguments, including facts, examples, and explanations of how they are related to the stance

We prompt an LLM to provide scores for each of the three criteria using the rubrics and levels described in Tables 2, 3 and 4 (The exact wording of the prompts is available on GitHub).[1] Similarly to the official ENEM rubric, the argument quality score is then computed as a function of the provided criterion-based scores, as follows. If the stance criterion is scored 0 or 1, that is the overall score. If it is 2, then the overall score is taken as the lowest of the organization and development criteria scores. We name this LLM-based strategy **3C**, for Three-Criteria Scoring Prompt.

In order to provide a baseline for argument quality scoring, we also consider two more holistic

[1] https://github.com/ndemoraes/AutoReformArgEssays

Score	Description
2	No development or development of only one piece of information or fact.
3	Development of some information or facts.
4	Development of most of the information and facts included in the essay.
5	Development of information and facts throughout the text. The argumentation must be articulate, clear, and coherent.

Table 3: Rubric for Development Criteria

Score	Description
2	Disorganized or contradictory argumentation.
3	Poorly organized argumentation.
4	Organized argumentation.
5	Consistent, organized argumentation.

Table 4: Rubric for Organization Criteria

prompting strategies that ask the LLM directly for an overall score. The **OS** (Overall Scoring) prompt uses the official ENEM summary rubric verbatim to describe the levels. Since two of the levels described in the summary rubric appear to overlap, we rephrased the wording to improve clarity. We name this prompt **OS-I**, for OS with Improved wording.

As per standard practice (Karmaker Santu and Feng, 2023; White et al., 2023), the AES prompts consisted of several building blocks: Persona (instructs the LLM to act like a certain type of person), Task, Reasoning (describes the reasoning that the LLM should use when carrying out the task), Background, and Essay (essay to be graded).

3.5 Automated Essay Reformulation

The main goal of this study is to analyze to what extent an LLM can be guided to provide specific feedback on argument organization and development. Since the "best" way to provide feedback in the form of a reformulated essay depends on many factors—and since the pedagogical utility of a specific type of reformulation will probably depend on the quality of the original essay—we seek to understand how to "dial" the reformulation-as-feedback process to obtain a variety of results.

Accordingly, we designed several series of prompts that progressively relax the degree of control imposed on the LLM, moving from highly constrained pedagogical reformulation toward unconstrained rewriting. We measure the effectiveness of each reformulation prompt using LLM-based multidimensional AES with Prompt 3C. This allows us to characterize the extent to which a reformulation prompt improves text with respect to each criterion.

We also ask a single human judge (a native speaker of Brazilian Portuguese with a degree in education, who works as a professional editor for publishing houses, and who is not a co-author of this paper) to evaluate the reformulated essays and to tally the number of argument elements (new claims, facts, examples, etc.) added to the texts. The human evaluator was not given any information about the prompts or their objectives. As previously described, we limit the analysis to a set of 20 essays on a single topic.

4 Results

We now report empirical results targeting answering a number of research questions (RQs).

4.1 Automatic Essay Scoring

We first investigate variations of each prompt for AES, such as including the short or long background information, using tags to delimit the student essay from other information, and prompt organization into system and user fields. Our experiments showed that using the short background (average 496 words) led to slightly higher inter-rater agreement (QWK) than using the long background (average 3068 words) (see Table 5 for an example). The use of delimiting tags (strings like [START OF ESSAY] and [END OF ESSAY]) led to an improvement of e.g. 7% in the MAE metric. In another experiment, we varied which prompt components were placed in the "system" or "user" fields of the LLM. This variation had a negligible effect on performance. Last, we experimented with a variation of Prompt 3C that also asked the LLM to provide a justification; the results were significantly worse (e.g. a 27% drop in MAE). This is particularly interesting because Wang and Gayed (2024) included a request for justification in their zero-shot prompt and did not try a version without justification, which might have performed better.

Using the best configurations, we then proceed to evaluate the validity of the three-criteria prompt. Recall that our interest in an LLM-AES model is to obtain automatic, human-like scoring (that can then provide a more fine-grained evaluation of the effects of reformulation). We interpret "human-like"

Prompt	Bkgd.	QWK	MAE	RMSE
3C	short	**0.6262**	0.5868	0.8494
3C	long	0.6165	**0.5808**	**0.8411**

Table 5: Comparison Across Background Lengths (all but essay in "system" field of prompt)

to mean that it follows the rubric for the argument quality score based on the three criteria discussed. Accordingly, we ask:

RQ 1: Can an LLM provide interpretable, accurate human-like automated scoring? To answer the question we verify if the overall argument quality scores obtained by Prompt 3C match the performance of the scores obtained by Prompts OS and OS-I, as well as the best performing method for this trait according to Barbosa et al. (2025), which was a 14.7B parameter Phi4 LLM fine-tuned on 28-56M parameters with LoRA (see the referenced paper for details).

Table 6 presents inter-rater agreement metrics for the argument quality scoring methods. The prompts are given the short background texts, as discussed in Section 3.2. One sees that Prompt 3C, which arguably more closely follows human-like reasoning, not only matches but surpasses the performance of the other prompts for all three metrics. Note that the improved wording in Prompt OS-I indeed leads to an increase in agreement over the original wording, but Prompt 3C has even higher performance.

We also see that Prompt 3C performs better than the Phi4 model, which was previously shown to beat the performance of zero-shot LLM-based prompting. We credit the improved performance to the fact that Prompt 3C reasoning instructions are shorter and include only aspects relevant to argument quality, while the prompts by Barbosa et al. (2025) use excerpts from the full ENEM rubric, which tend to be more verbose. Note that, although Prompts OS, OS-I and 3C are evaluated using all 168 instances of the EAC-AES dataset, Phi4 is evaluated only on the 28 instances that were not used for its training, to avoid an overly optimistic estimate. The change in dataset may render direct comparison of numbers unreliable and we cannot rule out the effect of the dataset on the differences observed between Prompts OS, OS-I and 3C and Phi4.

Further evidence that Prompt 3C performs scoring comparable to human scoring is the similarity

Approach	QWK	MAE	RMSE
Prompt OS	0.5632	0.7186	0.9439
Prompt OS-I	0.5930	0.6826	0.9128
Prompt 3C	**0.6283**	**0.5928**	**0.8374**
Phi4 (Test Set)	0.5748	0.6071	0.8660
A vs. B (EA-AES)	0.5700	0.7323	1.0491

Table 6: Comparison of AES Performance. Phi4 is evaluated only on the 28 instances of the EAC-AES dataset that were not used for its training.

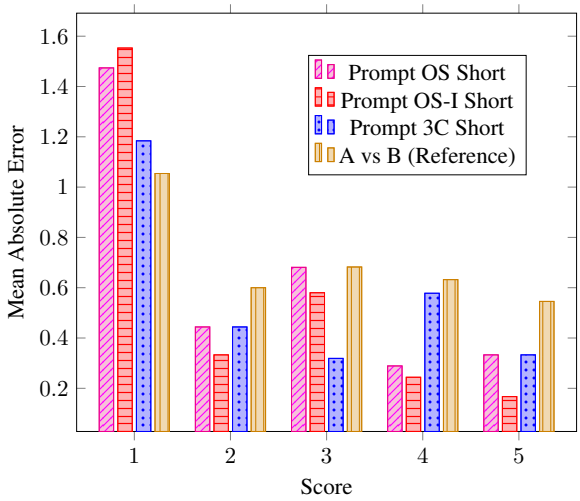

Figure 1: Mean Absolute Error per score level

of the values of the QWK, MAE and RMSE metrics (computed on the EAC-AES dataset) to those of the two raters (computed on the EA-AES dataset). Again, we recognize that direct comparison is complicated by the difference in datasets.

Our intended use for AES models is to evaluate the score change of reformulated essays. As such, it is important that the model performs well across different score levels, particularly at lower values (which have more room for improvement). Figure 1 shows per-level MAE score distributions of the the different AES models. One sees that Prompt 3C follows the per-level profile of the human raters closely, and its MAE errors are distributed more evenly across levels than those of the other prompts.

Data Leakage Since the essays and a set of human rater scores (ranging from 0 to 200) are publicly available on the internet, there is a risk that the results of any LLM-based AES are heavily influenced by memorization rather than generalization or reasoning. That risk was mitigated by two strategies.

First, the original annotated dataset followed the

ENEM guidelines and was graded on a 40-point, six-level scale (i.e., 0, 40, ..., 200). Our prompts, on the other hand, instruct the LLM to score on a scale from 0 to 5. Second, criteria scores like those generated by Prompt 3C are not on the UOL web portal or included in publicly available datasets. It is unlikely that criteria scores of this type are available online.

The essays on the UOL web portal are actually annotated with scores (which are not used in our experiments, due to their lack of reliability). To check whether the LLM might have looked up these scores despite our mitigation strategies, we compared the scores generated by the LLM to the UOL scores. The highest percentage of matching (where the LLM-generated argument quality score matched the UOL web portal's argument quality score) was about 38%. In contrast, the best percentage of matching between the LLM and Rater A's scores was about 55%. This is evidence that the model's agreement with Rater A and Rater B's scores is reflective of LLM reasoning abilities rather than data memorization.

4.2 Automated Essay Reformulation

We now report the results of the focus of this work: the extent to which LLMs can be controlled to produce reformulated versions of argumentative essays with respect to stance, development and organization criteria. We analyze LLM prompts progressing from highly constrained pedagogical reformulations toward unconstrained rewriting.

RQ 2: How can an LLM be guided to add facts and examples to improve development? To answer that question, we designed three prompts, all including background information, which asked the LLM to improve argument quality while avoiding changing text characteristics related to the other text features such as grammar, spelling, style and linguistic cohesion. Prompt P1 instructs the LLM to "minimally" reformulate the text in order to improve the argument. Prompt P2 instructs the LLM explicitly to add at least two ideas from the background material. Prompt P3 instructs the LLM to maximize the argument quality score of the reformulated essay, including by adding information from the background material. The results are summarized in the boxplot diagram in Figure 2, which shows both the score increase and the number of argument elements added.

The average argument quality score for Prompt P1 increased by 0.7 points. Despite the background material being provided and the LLM being told that it could use information from it to shore up the development, it did not. On average, it added 0.8 and subtracted 0.35 argument elements per essay. The results for Prompt P2 show that the number of argument elements added and the argument quality score both increased relative to Prompt P1.

While the average argument quality score of essays reformulated by Prompt P3 increased significantly (by 1.5 points), most if this increase is credited to an increase in the organization criteria score, which Prompt P2 (that was instructed to add argument elements) had not increased as much. Also, fewer argument elements were added by Prompt P3 than by Prompt P2, indicating that the LLM improved both organization and development in order to increase the argument quality score (recall that the argument quality score, when higher than one, is the lowest of the organization and development criteria scores).

We computed the cosine similarity of embeddings of the original and reformulated versions and observed that differences ranged from 0.9 to 0.96, suggesting that the meaning of the reformulated essays was not altered significantly. In contrast, the unrestricted revision prompt used by Song et al. (2024) resulted in cosine similarity values of around 0.86.

RQ 3: Can the LLM be guided to improve only organization or only development? To answer that question, we designed yet another two prompts. Prompt P4 guided the LLM to improve just organization (maintaining development) and Prompt P5 just development (maintaining organization). They instructed the LLM to maximize the organization criterion score and the development criterion score, respectively. As can be seen from Figure 2, both prompts failed to focus improvement on a single criteria. Note, however, that some of the imbalance was because the organization criterion score happened to be greater than or equal to the development criterion score for all essays, so the prompts focusing on improving just organization had less to improve.

It is interesting to compare Prompt P3 (asked to maximize the argument quality score) and Prompt P5 (asked to maximize only the development criteria score). Surprisingly, the latter resulted in higher scores on average (even for organization) than the former. It also had a larger Levenshtein distance be-

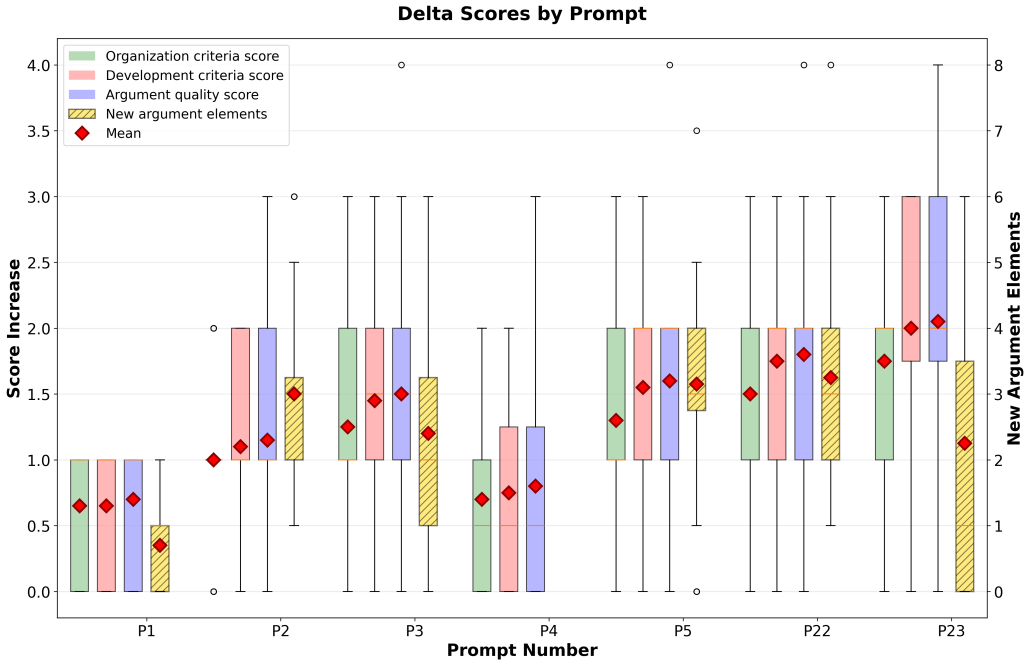

Figure 2: Prompts including background information

tween reformulated and original essays (not shown) and more argument elements added to the essay. The word count of the essays reformulated using Prompt P5 increased significantly: almost twice as many words were added as by Prompt P2 and almost 5 times as many as by Prompt P3. The human evaluator noted that Prompt P3 "had a will of its own" and for poor essays simply rewrote the text. The evaluator judged that Prompt P5 followed the original text more closely, which was verified by its higher cosine similarity value. When a poorly written essay was sarcastic, Prompt P3 rewrote it completely, creating a much more objective and politically correct version—this type of essay is more likely to get a maximum score on the ENEM exam, so this could be an unintended consequence of telling the LLM to maximize the argument quality score for an ENEM essay. All in all, Prompt P5 would arguably be a better pedagogical tool, since it improved the argument quality score the most while straying less from the original.

RQ 4: Is background information needed? To answer that, we tested a second series of prompts that differed from the first only by not including background material, just the topic introduction. The change in argument quality score and in the two criterion scores was similar to the first series for similar prompts. This seems to indicate that the LLM was able to strengthen argument development on its own, based on its training texts, but this finding is based on a small dataset on a single topic and may not be generalizable. Pedagogically speaking, adding evidence from the material provided to the student is probably more valuable, since the reformulation would appear more achievable. However, not providing background material in the prompt is certainly less expensive (fewer tokens).

RQ 5: Must the LLM be restricted to modifying only argument quality? We designed a third series of prompts. P22 is similar to P2, and P23 is similar to P3, but in P22 and P23 the LLM is not restricted to improving only argument quality. Although the number of argument elements added was about the same for the corresponding prompts in the first and third series (refer to Figure 2), the development and organization criteria scores were much higher for the third series. Five of the essays reformulated by Prompt P22 received a score of 4, while the remaining 15 received a score of 5. All but two of the Prompt P23 reformulations had argument quality scores of 5, while the others were scored as 4s.

According to the human evaluator, Prompt P23, which was asked to maximize the argument quality score, rewrote the texts with a significantly more elegant style (compared to the correspond-

ing Prompts P1 to P5), even when not necessary. Sometimes Prompt P23 made the essay more politically correct and in one case even changed the stance of the author to something less critical. Another difference noted by the human evaluator was the addition in most essays of a similar introductory sentence and conclusion sentence to many of the essays. This occurred not only in the essays with low scores, but even in the essay that originally had the top score for argument quality. This is apparently its standard method to improve the organization of the essay. It is more common in English writing than in Brazilian Portuguese writing, and is probably a side-effect of the LLM having been trained on English texts.

The Levenshtein distance between original and reformulated essays was noticeably larger for the new prompts (the distance for P22 was 63% higher than for P2; distance for P23 was 37% higher than for P3) and the cosine similarity was about 6% smaller. Even without human assessment, these metrics indicate that these new prompts changed the texts more extensively.

5 Conclusions

Through a series of controlled experiments, we demonstrated that LLMs can achieve levels of agreement in automated essay scoring comparable to human inter-rater agreement while following the argument quality criteria guidelines used by human graders. This facilitated detailed evaluation of targeted forms of argument quality reformulation for pedagogical purposes under strict constraints.

AES Our results provided evidence that the performance of LLM-based AES systems that simulate human scoring is high enough to be practical, with QWK values as high as 0.63 and MAE values as low as 0.58. Careful prompt engineering here is critical. A basic prompt, using the ENEM rubric summary wording, no clarifying tags and no background material was 31% worse in terms of MAE than the best prompt, which broke the scoring into three criteria and provided background material. An advantage of our three-criteria method is that the more fine-grained scores produced lead to greater interpretability than a holistic argument quality score. For reference, the best result by Barbosa et al. (2025) using GPT 4o was a QWK of 0.38, and their best model (a fine-tuned Phi4) obtained a QWK of 0.57, although these metrics were based on slightly different data.

AER Regarding Automatic Essay Reformulation, our results showed that simply instructing the LLM to improve an essay is not sufficient for pedagogically relevant feedback. One must specifically ask the LLM to add information (facts, examples, etc.) to strengthen argument quality or it will only make superficial changes. Our tests indicate that the LLM is unable to completely separate development improvement from organization improvement. Nonetheless, it can shift its focus to a certain extent when asked to alter only one of the two. Asking it to "maximize" the argument quality score tends to cause it to rewrite weaker essays, straying very far from the original, and is unlikely to be an effective pedagogical approach for weaker students.

Even when not provided with background information, the LLM is able to add true examples and facts to the essays to strengthen development. This approach would be less expensive (fewer tokens), but arguably less pedagogical, since the student reading the reformulation could reasonably claim that they did not have access to that information when writing.

Allowing the LLM to change features other than argument quality might be a good approach for stronger essays, which likely require less support regarding their arguments. These students would possibly benefit from reading the resulting polished essays. In general, the essays were rewritten—rather than corrected—so grammatical errors disappeared without being fixed in an obvious manner.

Limitations The main limitations of this study were the use of a single LLM (OpenAI's GPT 4o model) and the limited dataset size (168 essays for the AES experiment and 20 essays for the AER experiment). A third limitation was that the essays and LLM prompts were all written in Brazilian Portuguese. A forth limitation was the lack of human scores for the three criteria—stance, organization and development—for comparison. We also note that the essay scores concentrate around the average score, which biases comparisons. As future work, we intend to ask a human annotator to directly validate the fine grained scores provided by the LLM for the three criteria: stance, organization and development.

References

André Barbosa, Igor Caetano Silveira, and Denis Deratani Mauá. 2025. An empirical analysis of large language models for automated cross-prompt essay trait scoring in Brazilian portuguese. *Journal of the Brazilian Computer Society*, 31(1):858–871.

Brazilian Ministry of Education. 2024. Exame Nacional de Ensino Médio. Accessed on October 12, 2025.

Yingzhao Chen. 2025. Evaluating the potential of ChatGPT-reformulated essays as written feedback in L2 writing. *Computers and Education: Artificial Intelligence*, page 100500.

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Afrizal Doewes, Nughthoh Kurdhi, and Akrati Saxena. 2023. Evaluating quadratic weighted kappa as the standard performance metric for automated essay scoring. In *Proceedings of the 16th International Conference on Educational Data Mining, EDM 2023*, pages 103–113. International Educational Data Mining Society (IEDMS).

Erick Fonseca, Ivo Medeiros, Dayse Kamikawachi, and Alessandro Bokan. 2018. Automatically grading Brazilian student essays. In *Computational Processing of the Portuguese Language. PROPOR 2018*, pages 170–179. Springer International Publishing.

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). 2019. Enem redações - material de leitura, módulo 5, competência III. Accessed on 2025-10-01.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Compututing Surveys*, 55(12).

Joan Karbach. 1987. Using toulmin's model of argumentation. *Journal of Teaching Writing*, 6(1):81–92.

Shubhra Kanti Karmaker Santu and Dongji Feng. 2023. TELeR: A general taxonomy of LLM prompts for benchmarking complex tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14197–14203. Association for Computational Linguistics.

Ha Ram Kim and Melissa Bowles. 2019. How deeply do second language learners process written corrective feedback? Insights gained from think-alouds. *TESOL Quarterly*, 53(4):913–938.

Beata Beigman Klebanov and Nitin Madnani. 2022. *Automated Essay Scoring*. Springer Nature, Switzerland.

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

OpenAI. 2024. GPT-4o system card. Accessed on April 12, 2025.

Maria Santos, Sonia López Serrano, and Rosa M Manchón. 2010. The differential effect of two types of direct written corrective feedback on noticing and uptake: Reformulation vs. error correction. *International Journal of English Studies*, 10(1):131–154.

Lei Shu, Liangchen Luo, Jayakumar Hoskere, Yun Zhu, Yinxiao Liu, Simon Tong, Jindong Chen, and Lei Meng. 2024. Rewritelm: An instruction-tuned large language model for text rewriting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18970–18980.

Valerie J Shute. 2008. Focus on formative feedback. *Review of educational research*, 78(1):153–189.

Igor Cataneo Silveira, André Barbosa, and Denis Deratani Mauá. 2024. A new benchmark for automatic essay scoring in Portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 228–237, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Yishen Song, Qianta Zhu, Huaibo Wang, and Qinhua Zheng. 2024. Automated essay scoring and revising based on open-source large language models. *IEEE Transactions on Learning Technologies*, 17:1880–1890.

Jiao Sun, Yufei Tian, Wangchunshu Zhou, Nan Xu, Qian Hu, Rahul Gupta, John Wieting, Nanyun Peng, and Xuezhe Ma. 2023. Evaluating large language models on controlled generation tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3155–3168, Singapore. Association for Computational Linguistics.

Xiaoyi Tang, Hongwei Chen, Daoyu Lin, and Kexin Li. 2024. Harnessing LLMs for multi-dimensional writing assessment: Reliability and alignment with human judgments. *Heliyon*, 10(14).

Agustina Tocalli-Beller and Merrill Swain. 2005. Reformulation: The cognitive conflict and L2 learning it generates. *International Journal of Applied Linguistics*, 15(1):5–28.

Universo Online (UOL). 2020. Uol Educação Banco de Redações. Accessed June 11, 2025.

Qiao Wang and John Maurice Gayed. 2024. Effectiveness of large language models in automated evaluation of argumentative essays: finetuning vs. zero-shot prompting. *Computer Assisted Language Learning*, pages 1–29.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with ChatGPT. *arXiv preprint arXiv:2302.11382*.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2023. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Compututing Surveys*, 56(3).