Analysis of the Effectiveness of Large Language Models in Assessing Argumentative Writing and Generating Feedback

Daisy Cristine Albuquerque da Silva, Carlos Eduardo de Mello and Ana Cristina Bicharra Garcia
Federal University of Rio de Janeiro (UNIRIO), PPGI, Av. Pasteur, 458, Urca, RJ, Brazil

Keywords: Feedback Generation, Automated Feeback, Large Language Model, Feedback Effectiveness, GPT, LLM.

Abstract: This study examines the use of Large Language Models (LLMs) like GPT-4 in the evaluation of argumentative writing, particularly opinion articles authored by military school students. It explores the potential of LLMs to provide instant, personalized feedback across different writing stages and assesses their effectiveness compared to human evaluators. The study utilizes a detailed rubric to guide the LLM evaluation, focusing on competencies from topic choice to bibliographical references. Initial findings suggest that GPT-4 can consistently evaluate technical and structural aspects of writing, offering reliable feedback, especially in the References category. However, its conservative classification approach may underestimate article quality, indicating a need for human oversight. The study also uncovers GPT-4's challenges with nuanced and contextual elements of opinion writing, evident from variability in precision and low recall in recognizing complete works. These findings highlight the evolving role of LLMs as supplementary tools in education that require integration with human judgment to enhance argumentative writing and critical thinking in academic settings.

1 INTRODUCTION

Argumentative writing stands out for its persuasive expression on controversial topics, requiring authors to logically support their claims with evidence and counter potential objections (Blair, 2011), (Gage, 1987), (Ferretti and Lewis, 2018), (Kennedy, 1998). Unlike narrative or descriptive writing, it demands a strategic approach, balancing logical reasoning and compelling narrative crafting. This writing style involves anticipating and addressing opposing viewpoints, which adds to its complexity.

Creating argumentative texts is a dual-level cognitive challenge, combining the establishment of a coherent argumentative structure with the management of logical relations among discussion points (Kleemola et al., 2022), (Ferretti and Graham, 2019). The process is iterative, where content development and structural coherence are dynamically interlinked (Lovejoy, 2011). While tools like Sparks (Gero et al., 2022) and CoAuthor (Lee et al., 2022) offer assistance in specific writing stages, they often focus on isolated aspects of the writing process. Current technologies, despite being helpful, generally address either the planning, drafting, or revision phases, highlighting a need for more comprehensive solutions that support writers in structuring and refining their ideas across different phases and levels of abstraction (Almasri et al., 2019), (Wambsganss et al., 2021).

To address these gaps, we will undertake an effective evaluation of argumentative writing by employing Large Language Models (LLMs) such as GPT, guided by a detailed rubric. This rubric will provide clear descriptions of the required competencies at each stage of the article, ranging from the introduction to the final considerations.

This article investigates the integration of deep learning language models (LLMs), such as GPT, in enhancing argumentative writing. We use as a case study the Mario Travasso Project, a research encouragement initiative, where opinion articles written by military school students undergo a dual evaluation process: one by instructors and another by an LLM. In this process, authors receive constructive feedback based on a set of meticulously defined criteria, tailored to each stage of the development of an opinion article.

At the core of the investigation are the following research questions:

- **RQ1.** What is the effectiveness of Large Language Model (LLM) systems in evaluating argumentative writings?
- **RQ2.** What is the level of readability of the feed-

Albuquerque da Silva, D., Eduardo de Mello, C. and Garcia, A.
Analysis of the Effectiveness of Large Language Models in Assessing Argumentative Writing and Generating Feedback.
DOI: 10.5220/0012466600003636

In *Proceedings of the 16th International Conference on Agents and Artificial Intelligence (ICAART 2024) - Volume 2*, pages 573-582
ISBN: 978-989-758-680-4; ISSN: 2184-433X

back provided by Large Language Model (LLM)?

- **RQ3.** How does the feedback generated by Large Language Model (LLM) evaluations contribute to students' learning in developing more effective argumentative writing skills?

To answer these research questions, the study was structured into two distinct phases. Phase one involved a quantitative analysis, where the effectiveness of feedback from large language models (LLMs) was juxtaposed against instructors' evaluations in refining argumentative texts. This comparison was based on the predicted scores for various assessed categories within the texts. Additionally, a qualitative analysis was conducted in this phase, where instructors scrutinized the LLM-generated feedback. The objective was to explore the potential of incorporating this feedback into educational methodologies to enhance students' argumentative writing abilities.

The second phase of the study focused on the practical application of the LLM-generated feedback. Students received this detailed, personalized advice to revise their opinion articles, aiming to elevate the quality of their work based on the guidance provided. Post-revision, these articles underwent a second round of evaluation. This phase was pivotal in assessing the students' learning progression and the impact of the feedback they received. The overarching goal was to determine how the amalgamation of LLM's insights and instructor oversight can effectively foster students' argumentation and writing skills.

Additionally, it seeks to create guidelines based on human evaluators' perspectives to improve the relevance and acceptance of AI feedback. This approach aims to link AI advancements with their educational application, suggesting a method to effectively use automated feedback for enhancing argumentation and writing skills.

This paper is organized as follows: in the next section, presents a review of the literature on Generative AI in education by discussing the theoretical perspective adopted. After that, in Section 3, details the study's methodology, the opinion articles analyzed, and the competencies assessed, as well as the generation of automated feedback and the proposed evaluation method. Section 4 presents the findings in response to the research questions, highlighting the capabilities and limitations of LLMs in the evaluative context. Section 5 addresses the implications of the results, the limitations identified, and suggests future directions, concluding with a comprehensive view of the impacts and the need for advancement in integrating LLMs into educational practice, and finally, in Section 6, the conclusions are presented.

2 RELATED WORK

The significance of feedback in educational settings is well-established, prompting researchers to delve into the mechanisms by which feedback influences learning and to define what constitutes effective feedback. The model proposed by Hattie and Timperley (Hattie and Timperley, 2007) is notable in this area; it's widely recognized and has been employed extensively to scrutinize textual feedback in various studies (Cavalcanti et al., 2020), (Lin et al., 2023). This model (Hattie and Timperley, 2007) categorizes feedback into four levels — task-focused (FT), process-focused (FP), self-regulatory-focused (FR), and self-focused (FS) — each targeting key learning processes: setting learning objectives, evaluating current performance, and charting the course to achieve desired outcomes.

Within Automated Feedback Systems (AFSs), there is a tendency to rely on pre-defined rule sets created by domain experts (Pardo et al., 2018). While these systems offer some relief to educators by automating feedback, their utility is compromised when faced with open-ended tasks, such as student reports, which exhibit a wide range of possible responses and thus necessitate a large, complex set of rules (Jia et al., 2022). In light of recent advances in artificial intelligence, there is growing interest in leveraging pre-trained language models to provide textual feedback for more intricate tasks. The use of ChatGPT by Aydin and Karaarslan (Aydın and Karaarslan, 2022) in academic writing tasks illustrates this trend. Their findings indicate that ChatGPT is capable of generating literature reviews with minimal match rates when evaluated by plagiarism tools, showcasing the potential of such models. The encouraging performance of ChatGPT in this context makes it an intriguing subject for further investigation into its ability to offer elaborate feedback in argumentative writing.

Automated assessment is a long-standing endeavor in the field of educational technology. The initial automated assessment tools were geared towards solvable tasks, such as mathematics or programming assignments, where evaluation typically relies on unit tests or direct output comparisons (Hollingsworth, 1960); (Messer et al., 2023). These methods often overlook less quantifiable yet crucial indicators of learning and understanding, such as design quality, code maintainability, or areas that may confuse students. Modern tools like AutoGrader, which offers real-time assessment for programming exercises, still focus narrowly on output correctness and do not adequately account for documentation or maintainability (Liu et al., 2019).

Assessing students' understanding from natural language responses, however, poses different challenges and has undergone significant evolution. Early Automated Short Answer Grading (ASAG) models utilized statistical approaches or domain-specific neural networks (Heilman and Madnani, 2013); (Riordan et al., 2017). In recent years, Large Language Models (LLMs) have been shown to outperform domain-specific language models (Mann et al., 2020); (Chung et al., 2022). LLMs facilitate the grading of open-ended task responses without the need for task-specific fine-tuning (Cao, 2023); (Mizumoto and Eguchi, 2023); (Yoon, 2023). Yet, (Kortemeyer, 2023) disclosed that while LLMs like GPT-4 can be useful for preliminary grading of introductory physics tasks, they fall short for natural language responses required in comprehensive exam assessments. Furthermore, while LLMs such as GitHub Copilot streamline the code generation and review process, they may fall short on more nuanced programming tasks and open-ended evaluations (Finnie-Ansley et al., 2022). Therefore, in their current state, LLMs should be viewed as helpful but fallible tools, with final assessments still under the purview of human instructors.

It is also essential to consider how students perceive AI graders and the implementation of automated graders in educational settings (Zhu et al., 2022). Many discussions revolve around the socio-technical dynamics of automated assessment, including the potential for machine bias introduction (Hsu et al., 2021). Using NLP for short answer grading is not a trivial task and has been established as an assessment challenge in its own right.

3 METHOD

3.1 Opinion Article

The article examines a collection of 33 opinion pieces written by students from Brazilian Army military schools, focusing on National Defense topics. These articles aim to persuade readers through strong argumentative writing, drawing on viewpoints supported by citations from esteemed authors, relevant experiences, and thorough analysis. Authors are required to present well-reasoned perspectives and demonstrate deep understanding of the subject matter. They must take clear stances, knowing their work will face scrutiny and debate. This practice of argumentation showcases the author's analytical and critical skills and aids in shaping a unified military discourse, fostering a culture of informed and reflective national defense.

3.2 Skills Assessed

In this study, the evaluation of articles was conducted across multiple categories, each with specific criteria and scoring systems:

Choice of Topic. This category examined the relevance, timeliness, originality, and the author's understanding of the research topic. Each aspect was allocated a score of 0.25 points, totaling a maximum of 1 point.

Presence and Relevance of Keywords. The assessment in this category was based on the quantity of keywords used, with the scoring ranging from 0 to 1 point.

Quality of the Introduction. Articles were evaluated on their introductory presentation, clarity of objectives, research justification and importance, as well as the underlying hypotheses and methodologies used. Scores in this category varied from 0 to 2 points.

Development. This comprehensive category focused on the structured and detailed presentation of the topic, the theoretical basis, research methodologies, normative and typographical aspects, and the overall structure of the article, including pre-textual, textual, and post-textual elements. Writing quality, encompassing cohesion, coherence, spelling, and grammar, as well as adherence to ABNT standards, was also scrutinized. The score in this category ranged from 0 to 3 points.

Final Considerations. Here, the conclusion was evaluated based on how well it interpreted the arguments or elements discussed in the text, addressed the objectives and hypotheses, and summarized the research results and key information or arguments. The scoring in this category varied from 0 to 2 points.

References. The final category assessed the accuracy and completeness of the bibliographical references, including the correct identification and alphabetical ordering of cited works. The possible score ranged from 0 to 1 point.

In summary, the evaluation criteria included relevance and originality in topic choice, keyword inclusion, introduction presentation, formulation of the research problem, objectives, military perspective, hypothesis, methodology, and theoretical foundation. The development section focused on content organization, data collection and analysis, result discussion, and adherence to presentation standards, including writing quality aspects like cohesion and grammar. The final considerations assessed the author's ability to interpret arguments and summarize results. References were checked for comprehensiveness, ac-

curacy, and adherence to formatting rules. This thorough process aimed to ensure the academic quality of the opinion pieces.

3.3 Generative Pre-Trained Transformer

The Generative Pre-trained Transformer (GPT) series, particularly GPT-4, stands out among Large Language Models (LLMs), marking a significant evolution from its first iteration, GPT-1. OpenAI's ChatGPT, a specialized version of GPT tailored for conversational interactions, is based on the GPT-3 model and operates with 175 billion parameters, facilitating dynamic, interactive conversations. ChatGPT's popularity stems from its diverse capabilities in content generation and a user-friendly question-answering interface. Its public accessibility and practical applications across various domains have led to its widespread adoption. In education, ChatGPT has been explored for enhancing teaching and learning, offering personalized learning experiences, aiding concept comprehension, generating and explaining code, and supporting educational assessments (Biswas, 2023), (Firat, 2023), (Lo, 2023), (Rospigliosi, 2023).

In the article evaluation process, we employed the GPT-4 model. For each competency detailed in section 3.2., we developed specific prompts, aligned with the nuances and requirements of each of these skills. These prompts were carefully designed to interact with the GPT-4, allowing for a detailed and contextualized analysis of the articles.

3.4 Evaluation Method

The proposed method for analyzing the effectiveness of large language models (LLM) in assessing argumentative writing and generating feedback consists of two distinct phases.

In the first phase, both quantitative and qualitative analyses were conducted, where opinion articles were evaluated by both an instructor and the LLM.

According to the quantitative analysis, the evaluation resulted in scores for various categories, ranging from the choice of topic, the number of keywords cited, the construction of the introduction, the development of the argument, the crafting of the concluding remarks, and finally, the arrangement of the references. Depending on the score received, the article was classified as complete, partially complete, or incomplete. Moreover, at this stage, instructors examined the feedback generated by the LLM, which included a score for each category and a guidance text to help writers improve their writing. Figure 1 presents an example of this feedback provided by the LLM model.

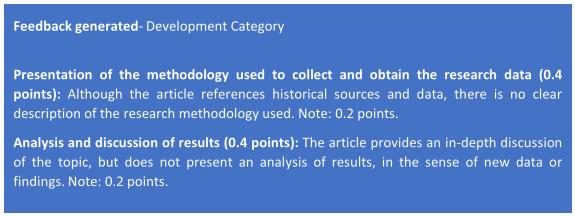

Figure 1: Feedback generated.

And according to the qualitative analysis, The study employed readability as a key measure to assess the text quality produced by the model, a common metric in evaluating written material. An instructor was enlisted to rate the feedback generated by the Large Language Model (LLM) using a detailed five-point scale. The scale is defined as follows: 0 for Incomprehensible; 1 for Not fluent and incoherent; 2 for Somewhat fluent but incoherent; 3 for Fluent but somewhat incoherent; and 4 for Fluent and coherent. The instructor applied this scale across various categories, including topic selection, keyword definition, introduction evaluation, development, final considerations, and references, assigning scores to each. Consequently, a readability score for each feedback piece provided in the opinion article evaluations was determined based on this systematic approach (Jia et al., 2022); (van der Lee et al., 2021).

The purpose of this phase was to explore how this feedback could be integrated into educational methodologies to improve students' skills in argumentative writing.

In the second phase, the students got involved in the practical application of the feedback they had received. They used the detailed and personalized feedback generated by the Broad Language Model (LLM) to refine their opinion pieces, with the aim of raising the quality of their work in line with the guidance they had received.

To evaluate the effectiveness of the feedback provided by GPT-4, we used the theoretical feedback framework developed by Hattie and Timperley. This framework facilitated the analysis of the feedback components to ensure that they were constructive and beneficial according to the standards set out in the model. In addition, students were tasked with evaluating feedback through the lens of the four-level model proposed by Hattie and Timperley, such as task, process, regulation and self. This allowed for a comprehensive understanding of the impact of such feedback on the learning process (Hattie and Timperley, 2007).

The four-level model of feedback, as proposed by Hattie and Timperley, provides a structured approach

to evaluating feedback across task relevance, process clarity, regulatory impact, and personal affirmation. It scrutinizes how feedback aligns with the task objectives, aids in information processing and learning, bolsters the learner's self-regulation, and supports their self-esteem and motivation. This comprehensive assessment ensures feedback is not only informative and task-specific but also constructive and empowering for the student (Hattie and Timperley, 2007).

The overall goal of the study was to determine how the combination of LLM knowledge and teacher supervision can effectively improve students' argumentation and writing skills.

4 RESULTS

The results achieved in this study were organized based on the three research questions previously established.

In response to the **RQ1**: What is the effectiveness of Large Language Model (LLM) systems in evaluating argumentative writings?, the chart in Figure 2 illustrates the distribution of scores assigned to opinion articles by two different evaluators: the Instructor and GPT-4.

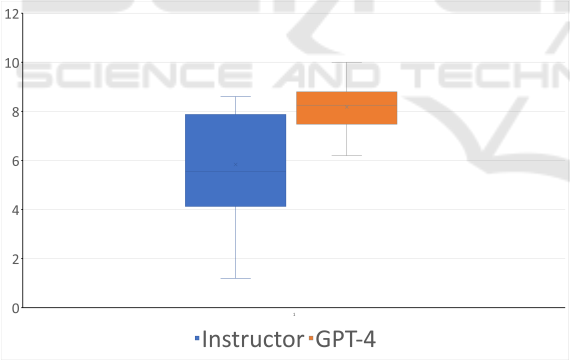

Figure 2: The distribution of article evaluation score.

The evaluations carried out by both the instructor and GPT-4 demonstrate the subjectivity inherent in the process of evaluating argumentative writing. There is a significant variation in the scores awarded by the instructor, which is clearly illustrated in the boxplot by longer whiskers and a larger box, indicating a greater dispersion of the data. This may reflect individual differences in the interpretation of the evaluation criteria or in the perception of the quality of the articles. On the other hand, the evaluations conducted by GPT-4 show less variation, with shorter whiskers and a narrower box in the boxplot, which suggests a more uniform application of the evaluation criteria. The higher median of the GPT-4 scores also implies a tendency of the language model to assign generally higher evaluations, perhaps due to a more objective analysis or a different sensitivity to quality elements in the writing.

The consistency observed in the GPT-4 evaluations highlights the potential of large-scale language models to provide uniform evaluations of argumentative writing. This tendency towards a more consistent evaluation may stem from the model's ability to apply evaluation criteria in a systematic way, without the variations that can arise from human subjective judgments. However, GPT-4's tendency to assign higher average scores suggests that it may be less critical or interpret the quality criteria differently compared to the human instructor. Furthermore, the presence of outliers in the instructor's evaluations underscores the importance of human judgment, particularly in cases where contributions are atypical or exhibit unique characteristics that may not be fully captured by an automated language model.

Also in response to the **RQ1**: What is the effectiveness of Large Language Model (LLM) systems in evaluating argumentative writings?, table in figure 3 shows the distribution of instructor and GPT-4 generated feedback.

Categories	Classified	Instructor	GPT-4	Precision	Recall
1 - Choice of topic	Complete	16	33	0,53	1
	Partially Complete	12	0	0	0
	Incomplete	5	0	0	0
2 - Keywords	Complete	8	8	1	1
	Partially Complete	3	3	1	1
	Incomplete	22	22	1	1
3 - Introduction	Complete	7	12	0,6	0,75
	Partially Complete	23	21	0,8	0,57
	Incomplete	3	0	0	0
4 - Development	Complete	0	3	0	0
	Partially Complete	33	30	1	0,9
	Incomplete	0	0	0	0
5 - Final considerations	Complete	6	26	0,23	1
	Partially Complete	23	6	0,66	0,017
	Incomplete	4	1	1	0,25
6 - References	Complete	21	12	0,91	0,052
	Partially Complete	11	18	0,044	0,72
	Incomplete	1	3	0,33	1

Figure 3: The distribution of instructor and GPT-4 generated feedback.

The table presents a comparison between the assessments of a human instructor and those conducted by GPT-4, a large-scale language model, across six distinct categories of writing evaluation criteria. Precision and recall metrics for the GPT-4's evaluations are also provided.

Regarding the category of topic choice, GPT-4 identified all 33 topics as complete, while the instructor classified only 16 as complete, with 12 being partially complete and 5 incomplete. This suggests that GPT-4 may have a broader or less stringent criterion

for considering a topic as complete. The precision is 0.53, indicating that more than half of GPT-4's complete classifications were correct, and the recall is 1, indicating that GPT-4 identified all the complete cases that the instructor classified.

In analyzing the evaluation of the introduction's quality, a significant discrepancy was observed, with GPT-4 classifying more introductions as complete (33) compared to the instructor (7). Both agree on 21 cases of partially complete introductions, but they differ on incomplete introductions. A precision of 0.6 for complete and a recall of 0.57 for partially complete indicate a moderate accuracy of GPT-4 aligned with the instructor's classifications.

On the other hand, in the development category, GPT-4 and the instructor are almost aligned in the development assessments, with GPT-4 marking none as incomplete. Precision is perfect (1) for complete and partially complete categories, while recall is 0.9 for partially complete, which is very high.

The biggest discrepancy was in the final considerations category, where GPT-4 classified a large majority as complete (26) as opposed to the instructor (6). Recall is perfect (1) for complete, but precision is very low (0.23), and recall is practically nil (0.017) for partially complete, which suggests that GPT-4 has difficulty evaluating this criterion properly compared to the instructor.

Overall, GPT-4 exhibits a tendency towards a conservative approach when confirming items as Complete, preferring a more lenient stance towards Partially Complete classifications. This is further corroborated by high recall scores across several categories, suggesting a strong likelihood of accuracy when Complete is assigned. Contrarily, the fluctuating precision rates imply that GPT-4 may prematurely or inaccurately assign Complete or Partially Complete statuses in certain scenarios.

In response to the **RQ2**: What is the level of readability of the feedback provided by Large Language Model (LLM)?, the chart in Figure 4 shows the human instructor's assessment of feedback texts generated by GPT-4 across various evaluated competencies.

The line graph depicted in figure 4 provides a detailed visualization of the evaluation scores for each of the 33 articles across multiple categories. The Choice of topic category consistently achieves the top score, indicating that the Large Language Model (LLM) consistently assesses the topics' relevance and appropriateness very positively.

In the categories of Keywords and Development, the feedback from the LLM also receives high scores, although there are minor fluctuations. These variations suggest that there might be slight inconsistencies in the LLM's feedback regarding these specific aspects of the articles.

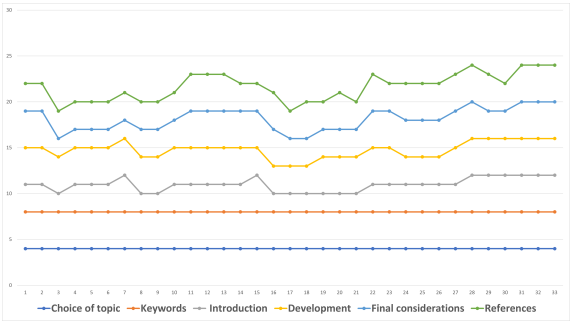

Figure 4: The human instructor's assessment of feedback texts generated by GPT-4 across various evaluated competencies.

The Introduction and Final considerations categories exhibit more significant score variations, with noticeable dips in the 'Introduction' scores at certain intervals. This pattern suggests that the LLM may struggle with consistently capturing an effective beginning of an argument or providing a convincing conclusion.

The References category displays the most significant score fluctuations, highlighting potential inconsistencies in how the LLM evaluates the adequacy and relevance of the references within the articles. This could indicate that the LLM's ability to provide feedback on references is less reliable and may require further refinement.

The high readability scores awarded by the instructor to the feedback from the Large Language Model (LLM) indicate that language models like GPT-4 hold promise for text evaluation tasks. Despite this, the observed variability in the scores for the 'Introduction' suggests that the LLM's processing and feedback generation capabilities for certain text elements could be enhanced.

While the LLM's feedback generally received high scores across most categories, this could either demonstrate its competence in producing coherent and pertinent content or suggest a possible inclination toward favorable evaluations. To ensure the feedback's accuracy and prevent potential bias, a discerning approach is warranted.

For a thorough and nuanced assessment of argumentative writing, it may be beneficial to refine the LLM's algorithms or to employ a hybrid approach that combines the model's automated evaluation with human review. This would help in maintaining the integrity and applicability of the feedback provided by the LLM, especially in areas like the introductory quality of articles where personalized and context-aware analysis is crucial.

Table 1: The distribution of four levels in the feedback provided by GPT-4.

Levels	Quantity	Frequency
Task	33	100%
Process	21	64%
Regulation	0	0
Self	0	0

In response to the **RQ3**: How does the feedback generated by Large Language Model (LLM) evaluations contribute to students' learning in developing more effective argumentative writing skills?, the feedback generated was assessed using Hattie and Timperley's four-level model. In this model, the feedback is evaluated in relation to each of the four levels: task, processing, regulatory, and self (Hattie and Timperley, 2007).

At the task level, the focus is on evaluating the relevance of the feedback to the specific activity, checking the accuracy and usefulness of the information provided to improve the student's performance in that specific task. At the processing level, the assessment concentrates on the feedback's ability to enhance the student's understanding and processing of information, considering the clarity, specificity, and relevance of the feedback to the learning process.

At the regulatory and self levels, the assessment of feedback takes a broader perspective. At the regulatory level, the feedback is analyzed for its effectiveness in improving the student's ability to regulate their own learning, including aspects such as self-assessment, motivation, and confidence. At the self level, the focus is on the impact of the feedback on the student's self-esteem and self-image, assessing whether the feedback provides appropriate recognition and encouragement. This four-level approach to feedback assessment aims to carefully analyze its effectiveness across various dimensions of student performance and learning.

Task-level feedback was consistently present, as indicated by the 100% figure in the table. This means that the Large Language Model (LLM), in this case GPT-4, provided relevant feedback for the task in all evaluations. This is in line with the model's objective to assess the accuracy and utility of information to enhance the student's performance on the specific task.

Process-level feedback was generated 64% of the time, indicating that more than half of the reports received comments aimed at improving the student's comprehension and information processing. This suggests a significant capability of the model to provide clear and specific guidance, which is crucial for the learning process.

Notably, there was no feedback generated at the regulatory and self levels. This may point to a limitation of the model in addressing aspects of self-regulation, motivation, confidence, self-esteem, and self-image. The absence of feedback at these levels might indicate that the LLM is not equipped to recognize or generate comments that directly influence these more subjective and personal areas of student learning.

The table 1 suggests that the LLM is effective in providing feedback related to tasks and processes, which is positive for helping students improve their argumentative writing skills. However, for a more holistic development of student skills, it would be beneficial if the model could also generate feedback at the regulatory and self levels, which requires future attention to enhance the model or to combine its use with the supervision of a human instructor.

5 DISCUSSION

5.1 Implications

The implications of this study are significant for the application of Large Language Models (LLM) like GPT-4 in educational settings, particularly in the context of evaluating argumentative writing.

Firstly, the study demonstrates the effectiveness of LLMs in providing consistent feedback on task-level elements of writing. This indicates that LLMs can potentially be used to augment the evaluation process, offering students immediate, objective, and consistent feedback on certain technical aspects of their writing. Such support could be instrumental in helping students to understand and meet the basic requirements of argumentative writing tasks.

The variance in scoring between GPT-4 and human instructors highlights another important implication: while LLMs might provide a uniform evaluation, they may do so with a bias towards leniency or may not align with human instructors' more critical and nuanced assessment of quality. This discrepancy underscores the need for human oversight to ensure the quality and criticality of evaluations.

Moreover, the study suggests that LLMs could be refined to better address the full spectrum of feedback categories, perhaps by incorporating more advanced natural language understanding and generation capabilities. Until such advancements are achieved, a hybrid evaluation approach that combines LLM feedback with human review might offer a more balanced and comprehensive assessment strategy.

This study suggests that while LLMs have the potential to significantly enhance the educational assessment process, particularly for argumentative writing, they currently require human collaboration to provide a holistic and nuanced evaluation. The integration of LLMs into educational practices should therefore be approached with caution, ensuring that the technology is used to complement and not replace the invaluable insights provided by human instructors.

5.2 Limitations

This study reveals several limitations in the use of Large Language Models (LLMs) like GPT-4 for evaluating argumentative writings and generating feedback.

The first limitation identified is GPT-4's inability to provide feedback on the regulatory and self levels, which are essential for students' comprehensive development. The absence of feedback in these areas suggests that GPT-4 may not be fully equipped to handle the more subjective aspects of learning, such as self-regulation, motivation, and self-confidence. This indicates that LLMs still need to evolve to provide a holistic assessment that goes beyond technical content and addresses the personal and emotional dimensions of learning.

Another limitation observed is the variation in the precision and recall of the feedback from GPT-4, suggesting that the model may be less critical and have a tendency to prematurely or inaccurately classify works as complete. This could lead to overly positive evaluations that do not adequately challenge students to improve their writing skills.

Moreover, the discrepancy between the evaluations of GPT-4 and human instructors highlights the variability and subjectivity in the assessment of argumentative writing. This underscores the need for human review to ensure critical and detailed evaluations, particularly in atypical cases or those that exhibit unique characteristics not fully captured by the automated model.

The research also points to the need to refine the algorithms of LLMs to improve feedback generation on specific text elements, such as the introduction and conclusion, where there was greater variability in evaluation scores.

While GPT-4 and other LLMs show considerable potential to assist in educational assessment, this study highlights several areas that require attention and future development. Human collaboration is still essential to ensure the quality and relevance of the feedback provided to students, especially in learning aspects that go beyond the current capabilities of LLMs.

6 CONCLUSION

The exploration of Large Language Models (LLMs) like GPT-4 in the assessment of argumentative writing within educational settings has yielded both promising potentials and notable limitations. The study confirms that LLMs are effective in providing systematic and consistent feedback on task-level elements, which is crucial for the development of students' argumentative writing skills. The ability of these models to offer immediate and objective feedback can greatly enhance the learning experience by providing students with clear guidance on technical aspects of their writing tasks.

However, the study also highlights critical limitations that need to be addressed. The LLMs, in their current state, lack the capability to engage with the more nuanced areas of learning, particularly self-regulation and personal development, which are vital for a student's holistic growth. The discrepancy in evaluation between human instructors and GPT-4 underscores the subjectivity and complexity of argumentative writing assessment, indicating the necessity for human insight to achieve a critical and nuanced evaluation.

Furthermore, variations in the precision and recall of feedback from LLMs point to a possible leniency in their evaluations, which could potentially lead to an inadequate challenge for students to improve their writing. The study also calls attention to the need for improving LLMs' algorithms, particularly in generating feedback for specific text elements that require more sophisticated analysis, such as introductions and conclusions.

In conclusion, while LLMs present a significant step forward in supporting educational assessments, the study suggests that they should not replace but rather complement human instruction. Human expertise remains indispensable in providing quality feedback, especially for aspects that transcend the current capabilities of LLMs. Future developments in LLM technology should aim for a more holistic assessment tool that can adequately address all levels of feedback and be integrated seamlessly into educational methodologies. As we advance, a collaborative approach, combining the strengths of LLMs with human oversight, appears to be the most effective strategy for enriching student learning and enhancing argumentative writing skills.

REFERENCES

Almasri, A., Ahmed, A., Almasri, N., Abu Sultan, Y. S., Mahmoud, A. Y., Zaqout, I. S., Akkila, A. N., and Abu-Naser, S. S. (2019). Intelligent tutoring systems survey for the period 2000-2018.

Aydın, Ö. and Karaarslan, E. (2022). Openai chatgpt generated literature review: Digital twin in healthcare. *Available at SSRN 4308687*.

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

Blair, J. A. (2011). *Groundwork in the theory of argumentation: Selected papers of J. Anthony Blair*, volume 21. Springer Science & Business Media.

Cao, C. (2023). Leveraging large language model and story-based gamification in intelligent tutoring system to scaffold introductory programming courses: A design-based research study. *arXiv preprint arXiv:2302.12834*.

Cavalcanti, A. P., Diego, A., Mello, R. F., Mangaroska, K., Nascimento, A., Freitas, F., and Gašević, D. (2020). How good is my feedback? a content analysis of written feedback. In *Proceedings of the tenth international conference on learning analytics & knowledge*, pages 428–437.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al. (2022). Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Ferretti, R. P. and Graham, S. (2019). Argumentative writing: Theory, assessment, and instruction. *Reading and Writing*, 32:1345–1357.

Ferretti, R. P. and Lewis, W. E. (2018). Argumentative writing. *Best practices in writing instruction*, 135.

Finnie-Ansley, J., Denny, P., Becker, B. A., Luxton-Reilly, A., and Prather, J. (2022). The robots are coming: Exploring the implications of openai codex on introductory programming. In *Proceedings of the 24th Australasian Computing Education Conference*, pages 10–19.

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

Gage, J. T. (1987). The shape of reason: Argumentative writing in college. *(No Title)*.

Gero, K. I., Liu, V., and Chilton, L. (2022). Sparks: Inspiration for science writing using language models. In *Designing interactive systems conference*, pages 1002–1019.

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

Heilman, M. and Madnani, N. (2013). Ets: Domain adaptation and stacking for short answer scoring. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 275–279.

Hollingsworth, J. (1960). Automatic graders for programming classes. *Communications of the ACM*, 3(10):528–529.

Hsu, S., Li, T. W., Zhang, Z., Fowler, M., Zilles, C., and Karahalios, K. (2021). Attitudes surrounding an imperfect ai autograder. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–15.

Jia, Q., Young, M., Xiao, Y., Cui, J., Liu, C., Rashid, P., and Gehringer, E. (2022). Insta-reviewer: A data-driven approach for generating instant feedback on students' project reports. *International Educational Data Mining Society*.

Kennedy, M. (1998). *Theorizing composition: A critical sourcebook of theory and scholarship in contemporary composition studies*. Bloomsbury Publishing USA.

Kleemola, K., Hyytinen, H., and Toom, A. (2022). The challenge of position-taking in novice higher education students' argumentative writing. In *Frontiers in education*, volume 7, page 885987. Frontiers.

Kortemeyer, G. (2023). Can an ai-tool grade assignments in an introductory physics course? *arXiv preprint arXiv:2304.11221*.

Lee, M., Liang, P., and Yang, Q. (2022). Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pages 1–19.

Lin, J., Dai, W., Lim, L.-A., Tsai, Y.-S., Mello, R. F., Khosravi, H., Gasevic, D., and Chen, G. (2023). Learner-centred analytics of feedback content in higher education. In *LAK23: 13th International Learning Analytics and Knowledge Conference*, pages 100–110.

Liu, X., Wang, S., Wang, P., and Wu, D. (2019). Automatic grading of programming assignments: an approach based on formal semantics. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering Education and Training (ICSE-SEET)*, pages 126–137. IEEE.

Lo, C. K. (2023). What is the impact of chatgpt on education? a rapid review of the literature. *Education Sciences*, 13(4):410.

Lovejoy, A. O. (2011). The great chain of being: A study of the history of an idea. new brunswick.

Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., et al. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Messer, M., Brown, N. C., Kölling, M., and Shi, M. (2023). Automated grading and feedback tools for programming education: A systematic review. *arXiv preprint arXiv:2306.11722*.

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

Pardo, A., Bartimote, K., Buckingham Shum, S., Dawson, S., Gao, J., Gašević, D., Leichtweis, S., Liu, D., Martínez-Maldonado, R., Mirriahi, N., et al.

(2018). Ontask: Delivering data-informed, personalized learning support actions.

Riordan, B., Horbach, A., Cahill, A., Zesch, T., and Lee, C. (2017). Investigating neural architectures for short answer scoring. In *Proceedings of the 12th workshop on innovative use of NLP for building educational applications*, pages 159–168.

Rospigliosi, P. a. (2023). Artificial intelligence in teaching and learning: what questions should we ask of chatgpt?

van der Lee, C., Gatt, A., van Miltenburg, E., and Krahmer, E. (2021). Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151.

Wambsganss, T., Kueng, T., Soellner, M., and Leimeister, J. M. (2021). Arguetutor: An adaptive dialog-based learning system for argumentation skills. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–13.

Yoon, S.-Y. (2023). Short answer grading using one-shot prompting and text similarity scoring model. *arXiv preprint arXiv:2305.18638*.

Zhu, X., Wu, H., and Zhang, L. (2022). Automatic short-answer grading via bert-based deep neural networks. *IEEE Transactions on Learning Technologies*, 15(3):364–375.