Prompt Engineering for Automatic Short Answer Grading in Brazilian Portuguese

Rafael Ferreira Mello[6,7], Luiz Rodrigues[2], Luciano Cabral[2,3], Filipe Dwan Pereira[4,7], Cleon Pereira Júnior[5,7], Dragan Gasevic[6], Geber Ramalho[1]

[1]Centro de Informática – Universidade Federal de Pernambuco
Recife – PE – Brazil

[2]Instituto de Computação - Universidade Federal de Alagoas
Maceió - AL - Brasil

[3]Campus Jaboatão dos Guararapes - Instituto Federal de Pernambuco
Jaboatão dos Guararapes - PE - Brasil

[4]Campus Boa Vista - Universidade Federal de Roraima
Boa Vista - RR - Brasil

[5]Campus Iporá - Instituto Federal Goiano - Iporá - GO - Brasil

[6]Faculty of Information Technology - Monash University - Melbourne - Australia

[7]Centro de Estudos Avançados do Recife - Recife - PE - Brasil

rafael.mello@ufrpe.br

Abstract. *Automatic Short Answer Grading (ASAG) is a prominent area of Artificial Intelligence in Education (AIED). Despite much research, developing ASAG systems is challenging, even when focused on a single subject, mostly due to the variability in length and content of students' answers. While recent research has explored Large Language Models (LLMs) to enhance the efficiency of ASAG, the LLM performance is highly dependent on the prompt design. In that context, prompt engineering plays a crucial role. However, to the best of our knowledge, no research has systematically investigated prompt engineering in ASAG. Thus, this study compares over 128 prompt combinations for a Portuguese dataset based on GPT-3.5-Turbo and GPT-4-Turbo. Our findings indicate the crucial role of specific prompt components in improving GPT results and shows that GPT-4 consistently outperformed GPT-3.5 in this domain. These insights guide prompt design for ASAG in the context of Brazilian Portuguese. Therefore, we recommend students, educators, and developers leverage these findings to optimize prompt design and benefit from the advancements offered by state-of-the-art LLMs whenever possible.*

1. Introduction

Assessment plays a critical role in the learning process, offering insights into students' knowledge acquisition and comprehension of the material. It also aids in refining teaching methodologies and enhancing the feedback process [Nicol and Macfarlane-Dick 2006]. Educational assessments encompass a broad range of activities, from multiple-choice, straightforward questions to grading, to evaluating open-ended responses like essays or

DOI: 10.5753/sbie.2024.242424

short answers [Burrows et al. 2015]. However, the complexity and labor-intensive nature of evaluating individual student assessments, especially in large classroom settings, present substantial challenges for teachers [Putnikovic and Jovanovic 2023]. Thus, the implementation of efficient strategies to overcome this problem is needed. One strategy is partial grading automation, which aims to simplify the evaluation process without compromising quality, as highlighted in the literature [Mohler and Mihalcea 2009, Bonthu et al. 2021].

In previous works, significant focus has been placed on Automatic Short Answer Grading (ASAG) [Mohler and Mihalcea 2009, Chakraborty et al. 2023, Bonthu et al. 2021, Putnikovic and Jovanovic 2023, Condor et al. 2021]. ASAG evaluates concise, open-ended responses against standard answers or specific criteria. The complexity of ASAG arises from the variability in length and content of the responses, which mirror the diverse linguistic expressions individuals use to express similar meanings. The development of ASAG systems, even for a single subject, is demanding due to these variations. Extending these systems across multiple domains to create a universal ASAG system presents an even more challenging task [Camus and Filighera 2020a]. These systems must be capable of interpreting and accurately grading a wide range of responses, each potentially unique in its presentation and meaning [Patil and Adhiya 2022, Putnikovic and Jovanovic 2023].

A promising solution to the challenges in ASAG could be the integration of Large Language Models (LLMs). State-of-the-art LMMs are trained with massive data, enabling them to respond accurately to a diverse array of questions across various subjects [Ziyu et al. 2023]. In education, their versatility extends to numerous applications, including question-answering systems and interactive learning [Baidoo-Anu and Ansah 2023]. Furthermore, integrating LLMs into digital learning platforms drew significant interest in the educational technology community, as indicated by various studies [Zirar 2023, Yan et al. 2024]. Previous research has demonstrated the potential of LLMs in assessment and evaluation tasks with encouraging outcomes [Moore et al. 2022, Nguyen et al. 2023, Yancey et al. 2023, Zirar 2023]. These studies have focused on the proficiency of the LLMs in evaluating distinct activities, encompassing various topics, difficulty levels, and assessment criteria.

Although the initial results demonstrate the potential of using LLMs for ASAG, it is important to highlight that using different prompts could significantly change the outcome. Prompt engineering, defined as the strategic formulation of interactions with LLMs [Karmaker Santu and Feng 2023], is essential in optimizing their performance. As [Gao et al. 2023] and [Ziyu et al. 2023] have pointed out, the effectiveness of an LLM in a given task significantly depends on the quality of the prompts used. Researchers have invested considerable effort in developing methods for creating appropriate prompts [Wei et al. 2022, Karmaker Santu and Feng 2023, White et al. 2023, Eager and Brunton 2023]. Therefore, the research community has proposed various studies to assess different types of prompt components tailored for specific tasks [Short and Short 2023, Taylor et al. 2023]. To the best of our knowledge, however, no previous work has performed a detailed analysis of prompt engineering for ASAG.

This paper reports on a study that aimed to determine which component of a prompt engineering process could enhance the accuracy of LLMs in assessing open-ended

questions. Focusing on GPT models, specifically 3.5-turbo and 4-turbo, the study assessed their performance in ASAG tasks across distinct educational levels (high school and higher education) for the Brazilian Portuguese language. The study evaluated 128 prompt designs, incorporating different prompt components identified in existing literature. Our preliminary findings indicate that components such as *time to think* and asking the model to justify the final grade consistently improved the performance of GPT models in the assessment task.

2. Literature Review

This study aims to analyze different prompts for Automatic Short Answer Grading (ASAG) for a dataset in Brazilian Portuguese. In this regard, this section gives a general presentation of ASAG and prompt engineering. This is followed by a presentation of some works in the literature that already shown results of LLM applications for ASAG. Finally, after an explanation of the literature, we present the research questions that guide this study.

2.1. Automatic Short Answer Grading (ASAG)

ASAG is a critical and actively evolving area within Artificial Intelligence in Education (AIED). This domain is dedicated to the automated evaluation of short textual responses from open-ended questions [Chakraborty et al. 2023]. For instance, [Chakraborty et al. 2023] introduced an ASAG method that employs vectors to represent the knowledge content in students' responses, utilizing cosine similarity to calculate scores. A key aspect of this method is its reliance on the Universal Sentence Encoder for performance, which brings up important considerations regarding its adaptability and accuracy across various student answers. Despite the simplicity of the method, the results were inconclusive as the authors evaluated a relatively small sample. [Sahu and Bhowmick 2020] conducted a comprehensive and systematic analysis of various ASAG systems. They reported the evaluation of several machine learning algorithms, but the main novelty was the introduction of ensemble methods that improved the performance of ASAG across diverse datasets. This ensemble-based approach, incorporating linear regression, achieved better results than simple models using the University of North Texas dataset for ASAG and classification tasks with the SciEntsBank and Beetle corpora for question and answering task.

Despite using traditional machine learning algorithms, the current literature focuses on using transformer models to create new methods for ASAG. For instance, [Sung et al. 2019] adopted BERT for ASAG. They demonstrate its superior performance across multiple domains, reporting an up to 10% absolute improvement in macro-average-F1 on the SemEval-2013 benchmarking dataset compared to state-of-the-art results. In the same direction, [Camus and Filighera 2020b] performed a wide assessment of several pre-trained Transformer-based architectures. In general, the RoBERTa large language model reached the best values for the different datasets evaluated.

Moreover, other papers evaluated the generalizability of models for ASAG. [Condor et al. 2021] investigated the influence of ASAG model components on generalization beyond the training set. They employed diverse methods, including Sentence-BERT and traditional approaches like Word2Vec and Bag-of-words, to generate vec-

tor representations of student responses. In the best-case scenario, the Sentence-BERT reached 62.12% accuracy.

Finally, [del Gobbo et al. 2023] introduced GradeAid, an ASAG framework. Unlike previous research, GradeAid accommodated non-English datasets, underwent a comprehensive validation and benchmarking, and was tested on diverse publicly available datasets, including a newly accessible dataset for researchers. Using advanced regressors for joint lexical and semantic feature analysis, GradeAid performed comparable to existing systems, demonstrating root-mean-squared errors as low as 0.25 for specific dataset-question pairs.

Despite the achievements of previous research in the field, a standard limitation has been the restricted generalizability of these models across various domains and languages. In this context, the recent and rapid advancements in LLMs present a promising solution to this challenge.

2.2. Prompt Engineering

As mentioned before, the design of a prompt could highly affect the performance of an LLM. Thus, the literature shows several attempts to include components to improve the quality of the prompt in a process called prompt engineering [Short and Short 2023, Karmaker Santu and Feng 2023, Wei et al. 2022]. In short, prompt engineering is the practice of designing and refining input prompts to communicate with AI language models effectively, optimizing the model's responses for accuracy and relevance [Karmaker Santu and Feng 2023].

For instance, [Wei et al. 2022] explored the effectiveness of including a series of intermediate reasoning steps in the prompt (called chain-of-thought) to enhance the capabilities of LLM in tackling complex reasoning tasks. Through experiments conducted on three substantial language models, the application of chain-of-thought prompting improves tasks such as arithmetic calculations, commonsense questions, and symbolic reasoning. The empirical findings reveal gains, illustrated by prompting a PaLM 540B with eight chain-of-thought exemplars. In [Brown et al. 2020], GPT-3's few-shot capabilities, another prompt component, are evaluated across various NLP tasks, including translation, question-answering, and cloze tasks, revealing competitive outcomes compared to prior state-of-the-art fine-tuning methods. GPT-3's few-shot learning excelled on different NLP datasets. However, challenges are identified on specific datasets, indicating areas where improvement is needed.

A broad examination of various components of prompt engineering was undertaken in the study by [Karmaker Santu and Feng 2023]. Their research offers suggestions on how specific components of a prompt, such as directly expressing the goal, using bullet lists for instructions, incorporating few-shot examples, integrating information from external resources, seeking explanations or justifications, and assigning roles, can significantly influence the performance of a LLM. Building upon these insights, they proposed a Taxonomy for Prompt Crafting (TELeR - Turn, Expression, Level of Details, Role), which serves as a structured framework for designing and optimizing prompts. However, a limitation of this study is its lack of empirical evaluation. While the proposed taxonomy provides a theoretical foundation for prompt crafting, the absence of practical, data-driven validation means that these suggestions' real-world effectiveness and applicability remain

untested.

In terms of education, [Eager and Brunton 2023] examined the potential of incorporating an LLM model into teaching and learning practices. Their study presented guidance on formulating instructional text using prompt engineering. They also illustrated the application of this AI technology in assessment design through a case study. Their standpoint highlights the importance of prompt engineering in an LLM as a valuable technique that complements other methodologies, enhancing teaching and learning outcomes in higher education. Although the recommendations for prompts in the educational context, the authors did not analyze the impact of different components in the outcome of the model.

2.3. LLM for ASAG

Given the relatively recent advent of LLMs, there is a limited but growing body of research evaluating these models in the context of ASAG. However, it is important to recognize and highlight the significant contributions of the studies conducted in this area.

In recent work, [Naismith et al. 2023] evaluated the performance of GPT-4 for evaluating discourse coherence in English, revealing its potential to produce ratings comparable to human assessments, up to 0.40 of Cohen's Kappa and 0.97 of adjacent agreement. The results suggest significant potential for enhancing Automated Writing Evaluation (AWE) technology in the learning and assessment domain.

Similarly, [Nguyen et al. 2023] examined more open-ended self-explanation responses from the Decimal Point learning game. This study evaluates how ChatGPT solves exercises, determines accuracy, and delivers meaningful feedback. Findings reveal ChatGPT's effectiveness in handling conceptual questions, yet it faces difficulties with decimal place values and number line problems. Nevertheless, ChatGPT reached an accuracy of 75% in assessing the student answers, and it generates high-quality feedback comparable to that of human instructors.

In another case, [Li et al. 2023] presented a framework utilizing ChatGPT for student answer scoring and feedback generation in automated assessment. Through diverse template prompts, they extracted rationales, refining inconsistent outputs to align with marking standards. The refined ChatGPT outputs are employed to fine-tune a smaller language model, resulting in an 11% improvement in the overall Quadratic Weighted Kappa (QWK) score compared to ChatGPT. The generated rationales from the method closely match those of ChatGPT, presenting an alternative solution for achieving explainable automated assessment in education. However, their prompt engineering approach faces limitations in testing due to the expansive search space for generating automated prompt text.

Applied to Finnish, [Chang and Ginter 2024] presented a study for ASAG with ChatGPT. the research used a dataset with 2000 student answers in Finnish from ten undergraduate courses. In this case, they created a prompt with the context, compared with zero-shot and one-shot settings, and put details about the expected output. The best results found were in GPT-4 with one-shot settings. In the same study, they observed a negative association between student answer length and model performance. The results encourage investigating the impact of the addition of few-shot.

Initial studies leveraging LLMs, including the ones mentioned in this section, often do not extensively utilize various prompt engineering techniques. In contrast, it is recognized that the effectiveness of LLMs in diverse tasks can be significantly enhanced through optimized prompt design [Karmaker Santu and Feng 2023, Wei et al. 2022].

2.4. Research Question

The initial research exploring the use of LLMs for ASAG primarily centered on assessing the performance capabilities of various models in this specific task. Nevertheless, to the best of our knowledge, no prior research addresses the evaluation of the prompt engineering process tailored for ASAG applications. This study aims to bridge this gap in the literature by systematically assessing the impact of different components of prompt design on the effectiveness of LLMs in the context of ASAG. As such, our first research question is:

> **RESEARCH QUESTION 1 (RQ1):**
> *What specific components of prompt design can enhance the effectiveness of LLMs when applied to ASAG?*

Furthermore, different LLMs may perform differently when applied to the same task. While much of the existing research has concentrated on utilizing models such as GPT-3.5 or open-source alternatives, recent developments have highlighted the significant potential of GPT-4 in a range of NLP tasks [OpenAI 2023]. Therefore, our second research question is:

> **RESEARCH QUESTION 2 (RQ2):**
> *To what extent can GPT-4 models surpass the performance of their predecessors in the context of ASAG?*

3. Method

As noted in the previous section, this study aims to answer two research questions involving the LLM model and prompt engineering. Based on this, this section presents the dataset used to answer the research questions, applying the Portuguese language context. Next, based on the literature review, we survey the relevant components in prompt construction that will be analyzed in this research. Finally, we present how the results should be evaluated.

3.1. Dataset

This work assessed the prompt engineering components using the dataset proposed by [Galhardi et al. 2020], which includes the traditional elements for ASAG evaluation: the question, instructor answer, student answer, and the final score for the answer of each student. This dataset (in this paper referred to as the PT_ASAG dataset), encapsulates 7,473 answers made by 659 students to 15 questions. The topic of this data was related to biology at the 8th grade of elementary school level, written in Brazilian Portuguese. In this case, 14 senior undergraduate biology students, all from the same class, evaluated the responses using a predetermined scale from 0 (lowest score) to 3 (highest score). These students were in their final year of college. At least two students scored each answer with Cohen's kappa of 0.43. Due to the cost associated with running the GPT models, we

randomly selected 30% of the data from this dataset to evaluate the prompt components. The selection included 30% of the answers for each question and stratified by scores in order to keep a similar behavior with the original data. Table 1 summarizes the total number of instances in the dataset and the ones utilized in the experimentation. It is important to note the dataset's inherent imbalance, which predominantly features scores in the 0 and 1 categories. Table 2 shows the distribution of the original dataset.

Table 1. Statistics of the dataset evaluated.

	Entire Data	Sample Used
Questions	15	15
Student Answer	7,473	2,242

Table 2. Distribution of the original dataset among classes.

Label	number	percentage
0	2354	31.50%
1	2227	29.80%
2	1775	23.75%
3	1117	14.94%

3.2. LLM model

The GPT models, including GPT-3.5, have gained significant attention in academic discussions for their ability to tackle various challenges, as highlighted in previous studies [Kasneci et al. 2023, Ziyu et al. 2023]. GPT-3.5 exhibits proficiency in learning new tasks through different prompting designs. It includes zero-shot learning (where no examples are provided), few-shot learning (which involves a small number of examples), and in-context learning (where learning is based on contextual information within the model's input limit), as described by Brown et al. [Brown et al. 2020]. It is important to highlight that GPT-3.5 has shown superior performance over the GPT-3 model in specific tasks like question-answering [Brown et al. 2020].

Currently, OpenAI released GPT-4, an enhanced language model that surpassed GPT-3.5 in numerous NLP tasks, as documented in the technical report [OpenAI 2023]. GPT-4 has performed better, ranking in the top 10% in various professional and academic examinations, including the Uniform Bar Exam in the USA and tests in disciplines like physics and psychology [OpenAI 2023].

Despite these achievements, it is important to acknowledge the limitations of GPT models, which include tendencies to produce inaccurate or fictional content (called hallucinations), constraints in processing extended contexts, and challenges in learning from extensive historical data [OpenAI 2023]. Considering GPT models' versatility in handling various tasks, we have integrated it into our experimental setup. More specifically, we used the OpenAI API to access the GPT-3.5-Turbo and GPT-4-Turbo Models.

3.3. Prompt Engineering

As mentioned before, the formulation of well-structured prompts is critical when employing large language models. The design of these prompts substantially affects the model's

performance in producing relevant and precise results [White et al. 2023]. There are many recommendations to write efficient prompts, including writing (i) clear instructions, (ii) delimitation of the context, and (iii) indication of output format [Giray 2023]; give time to the model think [Kojima et al. 2022]; define a list of actions to address the problem (as known as Chain-of-Though prompt) [Wei et al. 2022]; introduce the model role [Karmaker Santu and Feng 2023]; include examples of correct interactions (also known as few-shot prompt) [Brown et al. 2020]; provide additional information for the model [Gao et al. 2023]; ask the model to justify the outcome [Karmaker Santu and Feng 2023]; among others.

Based on the identified relevant prompt components, we crafted various segments of the to-be-evaluated prompt, as summarized in Table 3. This table provides an organized overview of each prompt component, illustrating our comprehensive approach to developing the prompt structure for the proposed experimentation.

Table 3. Prompt Components Assessed

Component	Text
Instruction	Assess the students' answer on a scale from 0 (completely incorrect) to 5 (perfect answer).
Context	Type of activity assessed.
Role	Act as a specific topic teacher.
Think	Think step by step.
Step by step	Explicitly list of steps to follow.
Few shot	include the instructor's answer as example
Rubric	Include the detailed rubric for the task
Justification	Ask the model to justify the final score suggested.
Output	Details about the expected output

Our research systematically assessed every possible combination of these prompt components to identify the most influential components in prompt effectiveness. We invariably included the *instruction* and *output* elements for each prompt variation, as these components are fundamental to providing a clear understanding of the task and the expected output format. In total, we evaluated 128 unique prompts.

3.4. Evaluation methodology

To assess the performance of the proposed prompts, we employed Cohen's κ [Cohen 1960] and Quadratic Weighted κ (QWK) [Vanbelle 2016], metrics widely recognized in the field of AIED. More specifically, to address RQ1, we conducted an assessment of the 128 prompts associated with each dataset. Initially, we ranked them based on their Cohen's κ scores for the ASAG task. In this case, a better-ranked prompt has a higher κ score when compared to the instructor's assessment of students' answers. Then, we analyzed the occurrence frequency of various prompt components within different tiers of ranked prompts - specifically, those in the top-5, top-10, and top-20 categories. This analysis enabled us to identify trends in the most prevalent components in the highest-scoring prompts. For this analysis, we employed GPT-3.5, primarily due to cost constraints.

To address RQ2, we utilized the GPT-4-Turbo model to evaluate the top-5 prompts identified in RQ1 for each dataset. This approach enabled us to directly compare the performance of GPT-4 with that of GPT-3.5, thereby providing insights into the efficacy of the newer model in the ASAG task.

4. Results

Since our study aims to answer two RQ, this section is organized according to these questions. Thus, the first part provides an overview of which prompt components are considered most relevant to the research context, and the second part presents a comparison between GPT-3.5 and GPT-4.

4.1. RQ1: Relevant Prompt Components

Table 4 displays the frequencies of the assessed prompt components within the top-5, top-10, and top-20 highest-performing prompts, as determined using GPT-3.5, ranked according to Cohen's κ values. The analysis excludes instruction and output elements, as these were present in all created prompts. The analysis reveals that distinct prompt components were relevant. The components 'few shoot' and 'role' proved to be most significant. In contrast, 'context' and 'rubric' were identified as the least effective components.

Table 4. Frequency of each prompt component.

Component	Top-5	Top-10	Top-20
Context	00	02	06
Role	04	**09**	**14**
Think	**05**	06	13
Step by step	02	02	08
Few shot	**05**	**10**	**20**
Rubric	04	06	11
Justification	02	02	03
Total	22	37	75

4.2. RQ2: Performance of GPT-4 for ASAG

Table 5 presents the outcomes for the top-5 prompts identified in the preceding RQ, now extended to include results from GPT-4. This table is ranked according to Cohen's κ values for GPT-3.5. The table highlights a consistently superior performance of the GPT-4 model across all cases. Considering Cohen's κ, GPT-4 attained a κ value of up to 0.7040, which falls within the range of substantial agreement. Moreover, there is a relevant observation emerging from the analysis. Due to the dataset's imbalance, the QWK scores are significantly higher than κ, given that QWK accounts for the frequency of each category relative to the number of instances.

5. Discussion

This paper analyzed the role of prompt engineering within the context of ASAG. For this, we investigated all combinations of seven prompt components (i.e., context, role, think, step by step, few shot, rubric, and justification), in terms of their ability to grade answers written in Brazilian Portuguese, for two advanced LLMs: GPT-3.5-Turbo e GPT-4-Turbo.

Table 5. Performance of GPT models for ASAG

	GPT-3.5		GPT-4	
Prompt Components	κ	QWK	κ	QWK
role, think, few shot, rubric	0.335	0.820	0.493	0.812
role, think, step by step, few shot, rubric	0.326	0.752	0.704	0.898
think, step by step, few shot, rubric	0.312	0.749	0.623	0.868
role, think, few shot, justification	0.312	0.749	0.620	0.842
role, think, few shot, rubric, justification	0.308	0.768	0.627	0.835

The results of this study reveal two primary insights. First, the concepts of 'time to think' and 'few shoot' emerge as significant factors, featuring prominently among the top-5 impactful components. 'Role' and 'Rubric' also appeared consistently in the top-5. These observations are consistent with existing research in the field of prompt engineering [Kojima et al. 2022]. Additionally, our analysis revealed that certain elements, such as 'context' and 'rubric,' demonstrated small to no effect in improving the results. This finding serves as an initial step toward establishing a systematic approach to prompt engineering for ASAG applications [Eager and Brunton 2023]. Therefore, we recommend students, educators, and researchers consider these prompt components while designing prompts.

Furthermore, our findings reinforce previous results where GPT-4 typically achieves better outcomes, even when compared with results obtained through prompt engineering on GPT-3.5 [OpenAI 2023]. Our results demonstrated that GPT-4 overcame GPT-3.5 across all cases, supporting prior research on the evolution of GPT-based models [OpenAI 2023]. Moreover, the substantial agreement achieved in the PT_ASAG dataset and consistently higher QWK scores reflect GPT-4's robustness in correctly assessing elementary-level students' answers with simple prompts. Notably, our top-performing result with GPT-4 on the PT_ASAG dataset outperformed the best outcomes reported in the existing literature by 26.5% and 38.3% for Cohen's $\kappa$and QWK [Galhardi et al. 2020], respectively.

In conclusion, these findings contribute valuable guidance for optimizing ASAG systems, emphasizing the importance of prompt design and demonstrating the potential advancements offered by state-of-the-art LLMs like GPT-4. As practical implications, students, educators and developers can leverage these insights to refine prompt design strategies, tailoring them to the particular needs of the ASAG task. Additionally, the observed performance boost of GPT-4 suggests that upgrading to newer LLMs can substantially enhance ASAG outcomes.

6. Limitation and Future Directions

We acknowledge the following limitations of this study. First, while various prompt components were evaluated, the precise composition of specific components (such as context, step-by-step instructions, and rubrics) could also influence performance. Our focus in this study was on simple texts, aimed at assessing the overall significance of each component. For future research, we plan to implement our methods in a real-world setting, involving course instructors to design more detailed and specific prompt components. This ap-

proach is expected to provide a deeper understanding of the impact of each component's articulation on performance.

The second limitation of our study concerns the dataset used. While we employed a previously used dataset in ASAG literature, our experimentation was restricted to 30% of its data due to cost constraints. However, it is noteworthy that multiple prior studies in the field of AIED and NLP have conducted evaluations with even smaller data sets, and our results are consistent with existing literature. In future research, we aim to assess a larger sample size, potentially encompassing various languages and contexts, to enhance the robustness and applicability of our findings.

Finally, our analysis was exclusively focused on GPT models, widely recognized for their strong performance. However, this approach does limit the scope of our study. In future research, we plan to broaden our analysis to include other LLMs, particularly those that are open-source. This expansion will enable a more comprehensive comparison and understanding of the capabilities of various LLMs for ASAG.

Artifacts Availability

The artifacts generated from this study are available from the corresponding author.

Acknowledgments

This paper was partially funded by Conselho Nacional de Desenvolvimento Científico e Tecnológico (310888/2021-2) and Open AI research grant. We acknowledge the use of generative artificial intelligence tools, such as ChatGPT (3.5), Grammarly, and Google Translate, to aid in writing and revising this paper. The authors conducted a thorough review of the text and assume full responsibility for its content.

References

Baidoo-Anu, D. and Ansah, L. O. (2023). Education in the era of generative artificial intelligence (ai): Understanding the potential benefits of chatgpt in promoting teaching and learning. *Journal of AI*, 7(1):52–62.

Bonthu, S., Rama Sree, S., and Krishna Prasad, M. (2021). Automated short answer grading using deep learning: A survey. In *Machine Learning and Knowledge Extraction: 5th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2021, Virtual Event, August 17–20, 2021, Proceedings 5*, pages 61–78. Springer.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., 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. M., 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.

Burrows, S., Gurevych, I., and Stein, B. (2015). The eras and trends of automatic short answer grading. *International journal of artificial intelligence in education*, 25:60–117.

Camus, L. and Filighera, A. (2020a). Investigating transformers for automatic short answer grading. In *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part II 21*, pages 43–48. Springer.

Camus, L. and Filighera, A. (2020b). Investigating transformers for automatic short answer grading. In Bittencourt, I. I., Cukurova, M., Muldner, K., Luckin, R., and Millán, E., editors, *Artificial Intelligence in Education*, pages 43–48, Cham. Springer International Publishing.

Chakraborty, C., Sethi, R., Chauhan, V., Sarma, B., and Chakraborty, U. K. (2023). Automatic short answer grading using universal sentence encoder. In Auer, M. E., Pachatz, W., and Rüütmann, T., editors, *Learning in the Age of Digital and Green Transition*, pages 511–518, Cham. Springer International Publishing.

Chang, L.-H. and Ginter, F. (2024). Automatic short answer grading for finnish with chatgpt. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23173–23181.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Condor, A., Litster, M., and Pardos, Z. A. (2021). Automatic short answer grading with sbert on out-of-sample questions. In *Educational Data Mining*.

del Gobbo, E., Guarino, A., Cafarelli, B., and Grilli, L. (2023). Gradeaid: a framework for automatic short answers grading in educational contexts—design, implementation and evaluation. *Knowledge and Information Systems*, 65(10):4295—4334.

Eager, B. and Brunton, R. (2023). Prompting higher education towards ai-augmented teaching and learning practice. *Journal of University Teaching & Learning Practice*, 20(5):02.

Galhardi, L., de Souza, R. C. T., and Brancher, J. (2020). Automatic grading of portuguese short answers using a machine learning approach. In *Anais Estendidos do XVI Simpósio Brasileiro de Sistemas de Informação*, pages 109–124. SBC.

Gao, L., Madaan, A., Zhou, S., Alon, U., Liu, P., Yang, Y., Callan, J., and Neubig, G. (2023). Pal: Program-aided language models. In *International Conference on Machine Learning*, pages 10764–10799. PMLR.

Giray, L. (2023). Prompt engineering with chatgpt: A guide for academic writers. *Annals of Biomedical Engineering*, pages 1–5.

Karmaker Santu, S. K. and Feng, D. (2023). TELeR: A general taxonomy of LLM prompts for benchmarking complex tasks. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14197–14203, Singapore. Association for Computational Linguistics.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., et al. (2023). Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Li, J., Gui, L., Zhou, Y., West, D., Aloisi, C., and He, Y. (2023). Distilling chatgpt for explainable automated student answer assessment. *arXiv preprint arXiv:2305.12962*.

Mohler, M. and Mihalcea, R. (2009). Text-to-text semantic similarity for automatic short answer grading. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 567–575.

Moore, S., Nguyen, H. A., Bier, N., Domadia, T., and Stamper, J. (2022). Assessing the quality of student-generated short answer questions using gpt-3. In *European conference on technology enhanced learning*, pages 243–257. Springer.

Naismith, B., Mulcaire, P., and Burstein, J. (2023). Automated evaluation of written discourse coherence using gpt-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 394–403.

Nguyen, H. A., Stec, H., Hou, X., Di, S., and McLaren, B. M. (2023). Evaluating chatgpt's decimal skills and feedback generation in a digital learning game. In *European Conference on Technology Enhanced Learning*, pages 278–293. Springer.

Nicol, D. J. and Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in higher education*, 31(2):199–218.

OpenAI (2023). Gpt-4 technical report.

Patil, S. and Adhiya, K. P. (2022). Automated evaluation of short answers: A systematic review. *Intelligent Data Communication Technologies and Internet of Things: Proceedings of ICICI 2021*, pages 953–963.

Putnikovic, M. and Jovanovic, J. (2023). Embeddings for automatic short answer grading: A scoping review. *IEEE Transactions on Learning Technologies*.

Sahu, A. and Bhowmick, P. K. (2020). Feature engineering and ensemble-based approach for improving automatic short-answer grading performance. *IEEE Transactions on Learning Technologies*, 13(1):77–90.

Short, C. E. and Short, J. C. (2023). The artificially intelligent entrepreneur: Chatgpt, prompt engineering, and entrepreneurial rhetoric creation. *Journal of Business Venturing Insights*, 19:e00388.

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

Taylor, N., Zhang, Y., Joyce, D. W., Gao, Z., Kormilitzin, A., and Nevado-Holgado, A. (2023). Clinical prompt learning with frozen language models. *IEEE Transactions on Neural Networks and Learning Systems*.

Vanbelle, S. (2016). A new interpretation of the weighted kappa coefficients. *Psychometrika*, 81(2):399–410.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., and Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., and Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, n/a(n/a).

Yancey, K. P., Laflair, G., Verardi, A., and Burstein, J. (2023). Rating short l2 essays on the cefr scale with gpt-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 576–584.

Zirar, A. (2023). Exploring the impact of language models, such as chatgpt, on student learning and assessment. *Review of Education*, 11(3):e3433.

Ziyu, Z., Qiguang, C., Longxuan, M., Mingda, L., Yi, H., Yushan, Q., Haopeng, B., Weinan, Z., and Liu, T. (2023). Through the lens of core competency: Survey on evaluation of large language models. In Zhang, J., editor, *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 2: Frontier Forum)*, pages 88–109, Harbin, China. Chinese Information Processing Society of China.