Automatic Short Answer Grading in the LLM Era: Does GPT-4 with Prompt Engineering beat Traditional Models?

Rafael Ferreira Mello
Monash University
Melbourne, Australia
Rafael.deMello@monash.edu

Cleon Pereira Junior
Instituto Federal Goiano
Iporá, Brazil
cleon.junior@ifgoiano.edu.br

Luiz Rodrigues
Computing Institute - Federal University of Alagoas
Maceió, Brazil
luiz_rodrigues17@hotmail.com

Filipe Dwan Pereira
Federal University of Roraima
Boa Vista, Brazil
filipedwan@gmail.com

Luciano Cabral
CJBG-IFPE, NEES-UFAL, CESAR School
Jaboatão dos Guararapes, Brazil
lscabral@gmail.com

Newarney Costa
Instituto Federal Goiano
Iporá, Brazil
newarney.costa@ifgoiano.edu.br

Geber Ramalho
UFPE
Recife, Brazil
glr@cin.ufpe.br

Dragan Gasevic
Monash University
Melbourne, Australia
dgasevic@acm.org

Abstract

Assessing short answers in educational settings is challenging due to the need for scalability and accuracy, which led to the field of Automatic Short Answer Grading (ASAG). Traditional machine learning models, such as ensemble and embeddings, have been widely researched in ASAG, but they often suffer from generalizability issues. Recently, Large Language Models (LLMs) emerged as an alternative to optimize ASAG systems. However, previous research has failed to present a comprehensive analysis of LLMs' performance powered by prompt engineering strategies and compare its capabilities to traditional models. This study presents a comparative analysis between traditional machine learning models and GPT-4 in the context of ASAG. We investigated the effectiveness of different models and text representation techniques and explored prompt engineering strategies for LLMs. The results indicate that traditional machine learning models outperform LLMs. However, GPT-4 showed promising capabilities, especially when configured with optimized prompt components, such as few-shot examples and clear instructions. This study contributes to the literature by providing a detailed evaluation of LLM performance compared to traditional machine learning models in a multilingual ASAG context, offering insights for developing more efficient automatic grading systems.

LAK 2025, Dublin, Ireland
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0701-8/25/03
https://doi.org/10.1145/3706468.3706481

CCS Concepts

• **Applied computing → Education**; • **Computing methodologies → Natural language processing**.

Keywords

Automatic short answer grading, Natural Language Processing, Assessment, LLM, GPT

ACM Reference Format:
Rafael Ferreira Mello, Cleon Pereira Junior, Luiz Rodrigues, Filipe Dwan Pereira, Luciano Cabral, Newarney Costa, Geber Ramalho, and Dragan Gasevic. 2025. Automatic Short Answer Grading in the LLM Era: Does GPT-4 with Prompt Engineering beat Traditional Models?. In *LAK25: The 15th International Learning Analytics and Knowledge Conference (LAK 2025), March 03–07, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3706468.3706481

1 Introduction

Assessment is a critical yet resource-intensive component of the learning process, particularly when scaling formative assessments, which are intended to be diagnostic, enabling students and teachers to adjust their approaches to maximize learning outcomes [4]. While closed-response assessment questions, such as multiple-choice and true/false, are commonly used in formative assessment due to their efficiency in grading, they have several drawbacks, including the potential for students to rely on test-taking strategies, a lack of face validity, and the complexity in generating multiple answer options [44]. In contrast, open-ended questions, which require students to respond in their own words, offer a more in-depth assessment of student understanding but are substantially more challenging to grade at scale [32]. In this context, Automatic Short Answer Grading (ASAG) has emerged as a promising solution, offering rapid and more objective evaluations that could simplify the assessment process without compromising quality [1, 38].

Research on ASAG demonstrates its capability to evaluate concise, open-ended responses by comparing them against standard answers or specific criteria [1, 13, 38, 44]. However, the complexity of ASAG systems emerges from aspects such as the variability in the length and content of student responses, which reflect the diverse linguistic expressions used to communicate similar meanings Ferreira et al. [17]. Due to these variations, developing effective ASAG systems takes time and effort. Extending such systems across multiple domains and languages is an even more daunting task [5], requiring the ability to score a wide range of responses, each uniquely expressed in its presentation and meaning [42, 44].

Integrating Large Language Models (LLMs) into ASAG presents a promising solution to the challenges faced by traditional ASAG systems, which generally depend on large annotated datasets [34]. LLMs, such as GPT-4, are trained on massive datasets and have demonstrated proficiency in responding to a wide range of questions across various subjects, making them suitable for educational applications like ASAG [58]. Unlike earlier state-of-the-art models, which relied on transfer learning and fine-tuning for specific tasks and required technical expertise, LLMs can often generalize better across different tasks with minimal prompt engineering, thereby reducing the technical barriers to their implementation in educational contexts [60]. Previous research has shown that LLMs can perform evaluation tasks on novel datasets with encouraging outcomes, even matching the performance of expert human raters in assessing short answer reading comprehension questions [40].

While preliminary findings suggest that LLMs have significant potential to enhance ASAG, it is crucial to consider several factors when evaluating their effectiveness [9, 24]. Techniques such as prompt engineering, few-shot learning, and fine-tuning are particularly important and can influence the performance of LLMs in specific educational tasks [7, 26]. Additionally, language- and domain-specific factors are crucial in determining the most suitable model for ASAG [17]. However, a limited understanding remains of how specific subjects, languages, and response types impact the performance of LLMs [47].

To the best of our knowledge, no previous work has evaluated the performance of LLMs across multiple languages or performed a detailed comparison with traditional machine learning and deep learning models in the context of ASAG, considering the accuracy and the generalizability of the models, as done in other tasks [11, 25]. This lack of comprehensive analysis represents a significant gap in the literature, particularly given the increasing availability and ubiquitous use of LLMs across various technologies, including educational tools. Compared to traditional machine learning or BERT-based approaches, their ease of integration makes them more likely to be adopted in ASAG systems, even when they may not outperform conventional methods. As a result, evaluating the strengths and limitations of LLMs in diverse linguistic environments is crucial, making this study especially timely and necessary.

Therefore, this paper presents a comprehensive comparative study of machine learning and deep learning models with LLMs applied to ASAG across two distinct datasets, one in Portuguese and the other in English. The study evaluates models utilizing various textual representation techniques, ranging from the TF-IDF [49] to the word embedding models like GloVe [43] and advanced transformers such as BERT [15]. Additionally, the study explores the impact of different prompt engineering strategies on model performance for LLMs. This research aims to provide valuable insights into the efficacy of these models across diverse linguistic contexts, addressing a gap in the existing literature regarding the application of LLMs and traditional models for ASAG in multiple languages.

2 Background

2.1 Automatic Short Answer Grading (ASAG)

ASAG is a research field that leverages Natural Language Processing (NLP) techniques to evaluate short textual responses to open-ended questions automatically [4, 17]. The most used methods to ASAG have centered on exploring various machine learning models for automatic grading [10, 19, 54]. For instance, Sahu and Bhowmick [48] conducted a comprehensive and systematic evaluation of multiple ASAG systems, comparing several machine learning algorithms. Their key innovation was the introduction of ensemble methods, which significantly enhanced ASAG performance across different datasets. Specifically, their ensemble-based approach outperformed simpler models, reaching 0.95 RMSE on the University of North Texas dataset [48].

Recent advancements in this field have increasingly focused on pre-trained language models (PLMs). For instance, Sung et al. [53] 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 [6] 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.

Several studies evaluated the generalizability of models for ASAG. Condor et al. [13] 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 vector representations of student responses. In the best-case scenario, the Sentence-BERT reached 62.12% accuracy. In a non-English context, del Gobbo et al. [14] introduced GradeAid, an ASAG framework. 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. Using a Transformer-based Hebrew (aleph-BERT).Despite these achievements, 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 Large Language Models for ASAG

LLMs for ASAG has become the focus of an increasing number of studies. Much of this research has concentrated on key aspects of grading, including prompt engineering, few-shot learning, and fine-tuning, which are critical factors that greatly influence the performance and effectiveness of these models in educational contexts [9, 59]. Given the potential of LLMs for automatic grading, research has also explored issues such as generalizability and fairness, challenges commonly encountered in both traditional machine learning models and human evaluations [51].

In a recent study, Naismith et al. [39] evaluated GPT-4's ability to assess discourse coherence, demonstrating that it can generate ratings closely aligned with those of human evaluators, achieving up to 0.40 in Cohen's Kappa and 0.97 in adjacent agreement. These results indicate significant promise for enhancing Automated Writing Evaluation technology in educational contexts. Similarly, Nguyen et al. [40] examined open-ended self-explanation responses from the Decimal Point learning game to evaluate ChatGPT's performance in solving exercises, determining accuracy, and delivering meaningful feedback. The study found that while ChatGPT is effective at handling conceptual questions, it encounters difficulties with tasks involving decimal place values and number lines. Nonetheless, ChatGPT achieved a 75% accuracy rate in evaluating student responses.

Some studies indicate that the challenges encountered by LLMs in the context of ASAG can be attributed to the fact that these sophisticated NLP models are not only concerned with providing an answer but also with addressing issues related to spelling and text structure [51]. In this regard, Urrutia and Araya [55] conducted a comparative analysis of several LLMs and traditional machine learning models within a specific mathematical context. Even after configuring the parameters and testing the prompts for LLMs, the traditional model demonstrated superior performance, reaching an f-score up to 94.49 against 85.27 achieved by GPT. The study highlighted that recursive questions or answers with spelling issues were significant factors that impeded the effectiveness of LLMs.

Henkel et al. [24] explored the potential of GPT models (3.5 and 4) in grading short-answer questions. The study compared the performance of these models using zero-shot and few-shot settings. The results indicated that GPT-4 significantly outperformed GPT-3.5. Although few-shot settings improved the models' performance, the transition from GPT-3.5 to GPT-4 was the most impactful feature. The best result, obtained with GPT-4 using few-shot prompting, achieved a Cohen's kappa score of 0.70.

The research by Chamieh et al. [9] presents another study on using LLM in the context of ASAG. What distinguishes this work is the comparison of two LLMs (GPT and LLaMA) in zero-shot and few-shot settings. In addition to these settings, the researchers also analyzed a fine-tuned model and a supervised upper bound. The study employed three datasets to conduct the experiments. The results demonstrate strong performance of the zero-shot and few-shot models in general knowledge tasks. GPT-4 performed (0.87 Quadratic Weighted Kappa (QWK)) close to the BERT (0.94 - QWK) and outperformed the SVM model (0.80 - QWK). The LLaMA models showed promising results (0.77 - QWK); although they did not reach the performance levels of GPT-4, their results remained consistent across different shot numbers. In contrast, GPT-3.5 appeared to overfit as more shots were introduced. This highlights the potential of few-shot LLMs for short answer scoring, particularly in tasks involving general knowledge questions.

In contrast to studies that examined LLMs for ASAG, Carpenter et al. [8] assesses the performance of several LLMs in comparison to FLAN-T5 models. This aimed at evaluating student explanations in real-time during classroom settings. The study used various prompt combinations to compare the performance of three LLMs: GPT-3.5, GPT-4, and LLaMA 2. Overall, the results demonstrated that fine-tuning FLAN-T5 and using few-shot learning with GPT-4 were viable approaches for evaluating explanations reaching 0.798 and 0.779 of the F1-weighted score, respectively.

Despite the promising results, the papers exploring LLMs for ASAG, in general, do not present a detailed analysis of how they compare to the wide range of traditional machine learning models often used in ASAG. Additionally, these studies do not make extensive use of the various prompt engineering techniques available. They often apply relatively straightforward prompts, potentially limiting the models' effectiveness in complex tasks such as ASAG. However, research increasingly recognizes that the performance of LLMs across a wide range of tasks can be substantially improved through the careful design and optimization of prompts. By refining the way tasks are framed and presented to these models, it is possible to better align their responses with desired outcomes, thus enhancing accuracy, consistency, and overall effectiveness [26, 56].

2.3 Prompt Engineering

Prompt engineering involves designing and refining input prompts to effectively communicate with LLMs, with the goal of optimizing the model's responses for accuracy and relevance [26]. Accordingly, research has demonstrated that prompt engineering can significantly enhance the performance LLMs across various tasks [35, 56].

For instance, Wei et al. [56] investigated the effectiveness of incorporating a series of intermediate reasoning steps (i.e., "chain-of-thought" prompting) to enhance the performance of LLMs on complex reasoning tasks. Their experiments, conducted on three language models, demonstrated that chain-of-thought prompting significantly improves performance on tasks such as arithmetic calculations, commonsense reasoning, and symbolic reasoning. The empirical results showed notable gains, particularly when a PaLM 540B model was prompted with eight chain-of-thought exemplars.

In another study, Brown et al. [3] evaluated GPT-3's few-shot learning capabilities, a prompt-based approach, across various NLP tasks, including translation, question-answering, and close tasks. Their findings revealed that GPT-3 achieved competitive results, often matching or surpassing previous state-of-the-art methods that relied on fine-tuning. While GPT-3 improved on multiple NLP datasets, the study also identified challenges on certain datasets, highlighting areas where further improvement is needed.

Karmaker Santu and Feng [26] conducted a comprehensive analysis of various elements involved in prompt engineering. Their study offers valuable insights into how specific prompt components — such as clearly stating the goal, using bullet points for instructions, incorporating few-shot examples, integrating external information, requesting explanations or justifications, and assigning roles — can significantly enhance the performance of LLMs. Building on these findings, they proposed a Taxonomy for Prompt Crafting, which includes Turn, Expression, Level of Detail, and Role, providing a structured framework for designing and optimizing prompts. However, the study's main limitation is the lack of empirical validation. While their taxonomy offers a solid theoretical foundation for prompt crafting, its real-world effectiveness and applicability remain untested due to the absence of data-driven evaluation.

Specifically for ASAG, Li et al. [30] introduced a framework that leverages ChatGPT for scoring student answers and generating feedback in automated assessments. By using a variety of template prompts, they extracted rationales from ChatGPT, refining inconsistent outputs to better align with marking standards. These refined outputs were then used to fine-tune a smaller language model, which led to an 11% improvement in the overall result score compared to ChatGPT alone, reaching 64.36 QWK. The rationales generated by this method closely resemble those produced by ChatGPT, offering a promising approach for explainable automated assessment in education. However, their prompt engineering strategy is limited by the vast search space required for generating effective automated prompts.

Overall, these studies demonstrate the increasing body of research concerning prompt engineering for LLMs. However, this review highlights that these studies rarely focus on ASAG and that the limited research in this context does not involve a wide range of prompt alternatives or compare LLMs to standard machine learning models, besides failing to investigate how these outcomes vary among different languages.

2.4 Research Questions

In summary, the literature highlights numerous efforts to enhance LLMs' performance through prompt engineering [26, 56]. Previous work has also evaluated different prompts for ASAG [8, 9, 24, 35]. However, to the best of our knowledge, no prior research has extensively evaluated different combinations of prompt elements to optimize LLM results for ASAG in a multilingual context. As such, our first Research Question (RQ) is:

> **Research Question 1 (RQ1):**
> *Which elements of prompt engineering can improve the performance of LLMs in the context of ASAG for Brazilian Portuguese and English?*

Another gap in the literature concerns the comparison of LLMs with traditional machine learning models, in the context of ASAG, to assess the potential advantages of using LLMs. To address this, we also sought to answer the following:

> **Research Question 2 (RQ2):**
> *To what extent can LLM models outperform traditional machine learning models in the context of ASAG?*

Additionally, the generalizability of ASAG models, particularly in handling unseen questions, remains underexplored. This gap encompasses both traditional machine learning models and LLM-based approaches, especially when confronted with question sets not present in their training data. Therefore, our third research question is as follows:

> **Research Question 3 (RQ3):**
> *How do traditional machine learning models and LLMs compare in classification performance when applied to unseen questions?*

3 Method

3.1 Dataset

We used two datasets [18, 37] and for each of them, we created two configurations of training and test sets. This section details each dataset and describes the methods used for stratification. The *first dataset*, created by Mohler et al. [37] and referred to as the Texas dataset in this paper, comprises 80 questions and 2,273 students' answers and has been used as the English dataset of this study. These questions were generated with assignments from the Data Structures course at the University of North Texas, with the answers of first-year undergrad students. Two human judges graded the answers independently, utilizing an integer scale from 0 (completely incorrect) to 5 (perfect answer). Both judges, graduate students in the computer science department, were listed as teaching assistants in the course. The average grade from both annotators served as the gold standard for system output comparison. According to the Pearson correlation coefficient, the inter-annotator agreement between the two grades was 0.586. The *second dataset*, proposed by Galhardi et al. [18] and referred to as the PT_ASAG dataset in this paper, has 7,473 answers 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). At least two students scored each answer with an overall agreement of 0.43, according to Cohen's kappa.

Stratifying the datasets was conducted in two distinct stages. In the *first stage*, referred to as split 1, approximately 30% of the data for each question was allocated to the test set, while the remaining 70% was used for training. The balancing factor was the score assigned to each student's answer. For questions with insufficient records across all score levels (fewer than two records), these cases were included in the training set, with the remaining records then used to achieve proper stratification. In the *second stage* (split 2), we allocated 30% of the questions for testing and 70% for training. To ensure a balanced distribution between the test and training datasets, we first calculated the median score for each question and then ranked the questions based on their median scores and the number of responses. Following this ranking, we systematically selected every third question until 30% of the questions were allocated to the test set.

Note the inherent imbalance in both datasets. The Texas dataset, for instance, shows a higher concentration of scores in the 3 and 4 categories, indicating a bias towards these ratings. Conversely, the PT_ASAG dataset predominantly features scores in the 0 and 1 categories. Table 1 summarizes the statistics of the datasets for each split described.

3.2 Machine and Deep Learning Models

In this work, we categorize traditional models as those that require a distinct training phase and are generally applied to specific domains. To facilitate a comprehensive comparison, we explored a range of models, from classic approaches like decision trees and support vector machines to advanced deep learning algorithms. Next, we detail the processing steps in training these models, including feature extraction.

Table 1: Statistics of the datasets evaluated.

	Split 1				Split 2			
	Texas Dataset		PT_ASAG Dataset		Texas Dataset		PT_ASAG Dataset	
	Train	Test	Train	Test	Train	Test	Train	Test
Questions	87	87	15	15	61	26	11	4
Student Answer	1690	752	6895	2970	1703	739	7220	2642

3.2.1 Feature Extraction. A traditional text mining technique, bag-of-words vector, was initially adopted in the present study. This method transforms from the textual document (e.g., online discussion messages) into an array consisting of the weights of the terms [33]. This study adopted the TF–IDF (term frequency-inverse document frequency) technique. A limitation of TF–IDF is the high dimensionality of the final vector generated. This happens because the TF–IDF uses the entire vocabulary of the documents analyzed to produce the final output. To minimize the effects of this problem, several pre-processing methods were applied[33]:

- *Removal of stopwords*: it removes the words with little significance in a text, such as articles, conjunctions, and prepositions.
- *Lemmatization*: It replaces the words contained in the corpus with their root forms, known as lemma. For instance, the verbs "is" and "been" are converted to "be."
- *Restriction of grammatical classes*: It applies a POS Tagger to analyze each word or term contained in a sentence and then assigns each item a grammatical class. In this study, only adjectives, adverbs, nouns, and verbs were considered relevant.

For the deep learning algorithms, we utilized Global Vectors for Word Representation (GloVe), FastText, and Bidirectional Encoder Representations from Transformers (BERT). GloVe transforms each word into a 300-dimensional vector, capturing the semantic relationships between words [43]. FastText, an extension of Word2Vec, constructs embeddings by considering subword information, which helps in representing rare words and morphologically rich languages [36]. Finally, BERT is a neural network designed for a deep understanding of language [15]. It pre-trains bidirectional representations from unlabeled text, considering the context of both sides of a word in all layers.

3.2.2 Model Selection. To evaluate the proposed features, we considered different models available for NLP applications. Initially, we adopted Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), and eXtreme Gradient Boosting (XGBoost) models due to their robustness for text classification problems [23, 31, 46]. DT and SVM are popular models that learn decision boundaries from labeled data, while RF and XGBoost are ensemble models that combine many basic models to improve the generalization on new data. We used these classifiers with both TF-IDF and BERT vector representations.

Additionally, we aimed to evaluate an approach that uses deep learning for the ASAG task. For this, we applied the BERT model and Bidirectional Long Short-Term Memory (BiLSTM), which have been used to outperform traditional classifiers in several NLP tasks [15]. LSTM networks are designed to learn from historical data, featuring an architecture that enables retaining previous information for use with current data [52]. In a standard LSTM, data flows in a single direction and is processed once. Conversely, a BiLSTM model processes the data twice, analyzing it in both forward and reverse directions, enhancing its ability to capture context [52]. As we did not aim to use parameter tuning techniques, we used the default parameters for training the BiLSTM model.

BERT is a contextual language model based on a deep neural network with bidirectional processing. It produces embeddings that change depending on the textual context of each lexicon occurrence, enabling the capture of meaning variations [15]. BERT is pre-trained on the Next Sentence Prediction (NSP) task, which helps it generate deep bidirectional representations of unlabeled text. Fine-tuning BERT requires adding only a single output layer and optimizing it for various NLP tasks. In this work, we utilize BERT-Multilingual for the Portuguese language. It is important to highlight that a substantial body of research consistently provides strong evidence for BERT's superior performance compared to other deep learning models [22, 27, 50]. This trend of superior performance also holds in specialized fields such as education, as shown by significant studies like Clavié and Gal [12] on EduBERT and Lee and Koh [29] on its use in teamwork dynamics.

3.3 LLM model

In this study, we integrated a GPT model to examine the potential of using LLM for ASAG. We utilized the gpt-4o-2024-08-06 model through the OpenAI API[1], which was the most advanced version available during our research. As mentioned in the background section, a crucial aspect of effectively using LLMs lies in constructing well-designed prompts. The quality and structure of these prompts play a significant role in the model's ability to generate accurate responses [57]. For the current study, we crafted prompts tailored to the specific context of the task, using straightforward instructions and clearly defining the expected output format [21]. In addition to these initial elements, we applied the following strategies: (i) allowing the model time to process the information [28], (ii) outlining a sequence of steps to solve the problem (known as the Chain-of-Thought approach) [56], (iii) defining the model's role [26], (iv) incorporating examples of correct responses (referred to as few-shot prompts) [3], (v) offering additional context [20], and (vi) requesting the model to explain its reasoning [26]. Table 3 explains the final prompt elements employed in this research. It is important to mention that we operationalized the prompts in the original language of the dataset.

[1] https://platform.openai.com/docs/api-reference/authentication

Table 2: Prompt Elements for Texas Dataset

Element	English text
Instruction	Assess the students answer in a scale from 0 (completely incorrect) to 5 (perfect answer).
Context	You are assessing computer science students in the first year.
Role	Act as a computer science professor.
Think	Think step by step.
Step by step	Follow the steps: 1. formulate your own correct answer to the question. 2. check if the relevant elements you identified are contained in the student answer. 3. elaborate the rationale to justify the quality of the student answer. 4. give the final score
Few-shot	It follow an example of a correct answer. Question: question_instructor Answer: answer_instructor
Rubric	The final score should evaluate content completeness.
Justification	Reasoning about the justification for your response by explaining why you made the choices you actually made.
Output	Your answer should be just the final score: "0", "1", "2", "3", "4" or "5".

In the current study, we evaluated every possible combination of these prompt components to determine which one had the most significant impact on prompt effectiveness. We consistently included the "Instruction" and "Output" elements in each prompt variation, recognizing their essential role in clearly describing the task and expected output format. Altogether, we assessed 128 distinct prompts for each dataset.

3.4 Evaluation methodology

To assess the performance of the proposed models, we employed Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE), metrics widely recognized for regression problems in the field of Text Mining [2] and specifically for ASAG [45]. We chose to use MAE and RMSE despite employing a classification approach because these metrics do more than simply count a prediction as correct when it matches the exact category. They provide a better measure of how closely the predicted scores align with the true values, capturing proximity and neighborhood-level accuracy more effectively [41]. In short, both RMSE and MAE give a sense of how far off predictions are from the actual values on average [2, 45]. RMSE represents the square root of the average squared differences between the predicted and actual values. In this case, larger errors have a more significant impact due to the squaring of the differences. In contrast, MAE measures the average magnitude of errors in a set of predictions without considering their direction. It is calculated by taking the average of the absolute differences between predicted and actual values.

To address RQ1, we assessed 128 prompts from each dataset, ranking them by their RMSE and MAE scores for the ASAG task, where higher-ranked prompts had better alignment with the ground truth. We then analyzed the frequency of various prompt components within the top-5, top-10, and top-20 ranked prompts to identify trends among the most prevalent components in the highest-scoring prompts elements.

For RQ2 and RQ3, we compared traditional machine learning algorithms with the top-3 GPT prompt configurations for dataset splits 1 and 2, respectively, as shown in tables 5 and 6. As outlined in Section 3.1, split 1 divided the datasets based on student answers, while split 2 divided them based on teacher questions. This means that in split 1, the same questions were used in both the training and test sets, whereas in split 2, the questions differed between the training and test sets.

4 Results

4.1 RQ1: Relevant Prompt Components

In Table 3, the analysis of the top-performing prompt components for the Texas and PT_ASAG datasets highlights the highest-performing prompts and the key differences between these two datasets. For the *Texas dataset*, which contained English answers, the most impactful component within the prompts was 'few-shot', showing a consistent presence across all rankings (top-5, top-10, and top-20). Other components like 'context', 'role', and 'rubric' showed a lower but notable impact in the top-20 prompts, with 'think' and 'justification' components appearing less frequently. This suggests that when grading English-language answers, providing the model with few examples (few-shot) significantly enhanced its ability to assess the answers correctly. The key components of the *PT_ASAG dataset*, which consisted of Portuguese answers, differ. 'Think' and 'Rubric' emerged as more influential, especially within the prompts. Additionally, 'Justification' and 'Role' components contributed significantly across the top rankings, particularly in the top-10 and top-20 prompts. This indicates that for Portuguese responses, providing detailed reasoning steps (think) and clear rubrics may have helped the model understand and grade the short answers more effectively, pointing to the importance of adapting the prompt design to different languages.

Table 3: Frequency of each prompt component.

	Texas data			PT_ASAG data		
Component	Top-5	Top-10	Top-20	Top-5	Top-10	Top-20
Context	01	05	11	02	04	09
Role	02	02	05	03	06	08
Think	01	03	05	05	10	16
Step by step	00	00	05	00	00	03
Few-shot	05	10	20	00	00	04
Rubric	01	03	07	04	07	14
Justification	01	02	07	04	07	13

Table 4 presents the detailed performance of the top-5 prompts for both datasets. The table shows that for the *Texas dataset*, the configuration with just the "few-shot" component achieved the lowest MAE of 1.02 and RMSE of 1.48. Combinations like "think + few-shot + rubric" yielded slightly higher errors, highlighting that additional components do not always enhance model accuracy for this dataset. For the *PT_ASAG dataset*, the best performance was observed with the "think + rubric + justification" prompt combination, with a remarkable reduction in error values–an MAE of 0.51 and RMSE of 0.94. GPT demonstrated an enhanced ability to handle straightforward tasks, as reflected by the lower error values achieved in PT_ASAG, which consisted of simpler questions. Complex configurations like "role + think + justification" had slightly higher errors but still outperformed the Texas dataset, which can be attributed to the 8th-grade level of the questions in PT_ASAG and their relative ease as compared to those in the Texas dataset. These results suggest that prompt design can have a notable impact on the model's performance, particularly when aligned with the complexity of the dataset.

4.2 RQ2: Performance of Traditional Models and GPT-4 for ASAG

Table 5 provides a detailed performance comparison between traditional machine learning algorithms and GPT-based models in terms of MAE and RMSE for the Texas and PT_ASAG datasets, using the split 1. As mentioned in section 3.1 split 1 used data from the same questions in training and test sets. For the *Texas dataset*, the BERT classifier was the top-performing algorithm in this analysis, reaching 0.59 and 1.02 for MAE and RMSE, respectively. Moreover, TF-IDF models, in combination with SVM and RF, achieved relatively low errors, with TF-IDF SVM showing an MAE of 0.60 and RMSE of 1.15. The performance of the BERT model on the *PT_ASAG dataset* was superior compared to that on the Texas dataset. In this case, BERT achieved 0.34 and 0.67 for MAE and RMSE, respectively. The TF-IDF models, particularly TF-IDF RF, obtained the lowest MAE of 0.37 and RMSE of 0.73. GPT4o-based models demonstrated higher error rates for both datasets, with MAE values reaching 1.02 and 1.28 for Texas and PT_ASAG datasets, respectively. This suggests that while BERT and traditional models managed to reach good results, especially with simpler tasks and lower grade-level questions, GPT4o capabilities may still require further prompt tuning to improve grading accuracy.

4.3 RQ3: Generalizability of Traditional Models and GPT-4 for ASAG

Table 6 summarises the the results of the generalizability analysis of of the evaluated models. For the *Texas dataset*, the best performance observed was for the BERT and TF-IDF SVM models, with an MAE of 0.64 and 0.65. Other models, such as BERT embeddings with traditional models, showed moderate performance with slightly higher MAE and RMSE values. For instance, BERT SVM achieved an MAE of 0.71 and an RMSE of 1.19. For the *PT_ASAG dataset*, the BERT and traditional TF-IDF DT models perform similarly but with slightly lower error rates when compared to the results on the Texas dataset - MAE of 0.58 (BERT) and 0.63 (TF-IDF DT). The GPT-based models achieved superior results, outperforming BERT on the PT_ASAG dataset and delivering comparable performance on the Texas dataset. These findings suggest that GPT models offer strong generalizability across different tasks and languages compared to traditional machine learning models.

5 Discussion

This paper presents the findings of a comprehensive comparative study of how various kinds of machine learning models performed on ASAG tasks from a multilingual perspective. In doing so, we addressed which prompt elements were more likely to improve LLMs' performance (RQ1), how LLMs peformed in comparison to machine and deep learning models (RQ2), and how these models' performances varied across different questions (RQ3) in the context of ASAG for English and Brazilian Portuguese. Table 7 summarizes the answers to these RQs, which are discussed next.

Our results in response to RQ1 (Section 4.2) first showed "few-shot" was important for the Texas dataset, while "Time-to-Think" is relevant for PT_ASAG. This confirms the relevant role of providing examples and giving the model space to reason, a finding that aligns with previous research in prompt engineering [28, 35]. Second, the analysis reveals a language-dependent or/and level of questions variation in the effectiveness of prompt components. For the English dataset, "few-shot" and "Context" were the most impactful, while for the Portuguese dataset, "think" and "rubric" were the leading contributors. This adds empirical results to the literature on prompt engineering for ASAG, which is still inconclusive in designing efficient prompts. For instance, some previous studies demonstrated the relevance of "few-shot" [24], whereas others showed that the "few-shot" prompt did not improve the final result of GPT [8]. Therefore, this study contributes by demonstrating the relevance of different prompt elements from the context of both English and Brazilian Portuguese answers.

Our results in response to RQ2 (Section 4.2) demonstrated that, regardless of the language, traditional machine learning models substantially outperformed LLMs. Although we explored a wide range of prompt combinations, all the traditional machine learning models considered yielded lower error rates compared to GPT-4o regarding both MAE and RSME for both English and Brazilian Portuguese answers. Given the increasing interest in LLMs, much has been discussed regarding their applicability to various tasks, especially within the educational context [58]. However, previous research did not extensively evaluate these models in light of the various ways they might be prompted compared to traditional machine learning models [9, 35, 39, 40, 51, 55]. Therefore, this study contributes empirical evidence demonstrating that, even with prompt engineering, state-of-the-art LLMs, represented by GPT-4o in this study, are unlikely to perform as well as traditional machine learning models in ASAG.

Our results for RQ3 (Section 4.3) expand the previous analyses with insights concerning RQ2. Overall, the findings suggest that, in grading answers for new questions, the models' performances were comparable to those they yielded in grading questions available in the training process. For instance, the best performance in terms of MAE was 0.59 for models with answers to all questions, whereas models trained with a limited sample of questions achieved a MAE value of 0.64 when grading answers to new questions. Additionally,

Table 4: Performance of GPT models for ASAG

		Dataset Split 1		Dataset Split 2	
Dataset	Prompt Components	MAE	RMSE	MAE	RMSE
Texas	few-shot	1.02	1.48	0.69	1.01
	context + role + few-shot	1.04	1.50	0.70	1.04
	role + few-shot	1.06	1.53	0.71	1.06
	few-shot + justification	1.06	1.50	0.72	1.05
	think + few-shot + rubric	1.06	1.50	0.72	1.05
PT_ASAG	think + rubric + justification	1.23	1.72	0.51	0.94
	context + think + rubric + justification	1.32	1.76	0.56	1.02
	role + think + rubric	1.35	1.80	0.59	1.06
	role + think + justification	1.35	1.80	0.59	1.06
	context + role + think + rubric + justification	1.37	1.85	0.59	1.06

Table 5: Results for the analyzed algorithms in terms of MAE and RMSE for Dataset split 1.

		Texas Dataset		PT_ASAG Dataset	
Features/Prompt	Model	MAE	RMSE	MAE	RMSE
TFIDF	SVM	0.60	1.15	0.43	0.81
TFIDF	RF	0.63	1.16	0.37	0.73
TFIDF	DT	0.71	1.24	0.47	0.84
TFIDF	XGB	0.64	1.18	0.41	0.77
BERT	SVM	0.71	1.22	0.49	0.89
BERT	RF	0.79	1.30	0.67	1.09
BERT	DT	0.79	1.30	0.67	1.09
BERT	XGB	0.71	1.22	0.47	0.86
FastText	BILSTM	0.62	1.13	0.58	1.00
Glove	BILSTM	0.60	1.07	0.49	0.89
BERT	BERT	0.59	1.02	0.34	0.67
few-shot	GPT4o	1.02	1.48	1.66	2.13
context + role + few-shot	GPT4o	1.04	1.50	1.65	2.12
role + few-shot	GPT4o	1.06	1.53	1.55	2.01
think + rubric + justification	GPT4o	1.55	1.96	1.23	1.67
think + step by step + rubric + justification	GPT4o	1.65	2.13	1.28	1.76
step by step + rubric	GPT4o	1.51	2.00	1.28	1.75

the difference between GPT-4o and traditional models remained in this setting. Thereby, these insights expand our understanding of how different kinds of machine learning approaches perform in ASAG, revealing that traditional models' advantages compared to LLMs remains, even in a context where LLMs were expected to excel due to their generalization potential powered by their training procedure.

Based on that context, this paper has two main implications to LA research and practice. *First*, the differences in relevant components for different datasets indicate the need to tailor prompt engineering strategies based on the target domain (e.g., language and questions level) in the ASAG contexts. Moreover, components like "step by step" had a minimal influence across both datasets, suggesting that their contribution to performance improvement might be limited. These findings could support educators and researchers when developing prompts to maximize the performance of LLMs in ASAG tasks [16]. *Second*, the results from comparing LLMs and traditional models highlight the need for a critical analysis between performance and practicality in deploying ASAG. On the one hand, traditional models clearly outperformed LLMs, which supports choosing such methods in cases wherein predictive performance is the priority. On the other hand, traditional methods demand data, expertise, time and other resources to be developed, whereas LLMs are *ready to use*. For instance, one might build upon our insights to define their prompts and automatically grade short answers. Accordingly, opting for LLMs might be suitable in the absence of resources for training machine and deep learning models, or in semi-automated contexts. Thus, these findings could support educators and developers in defining the best AI model to empower their ASAG systems, depending on their priorities and resources available.

Table 6: Results for the analyzed algorithms in terms of MAE and RMSE for Dataset split 2.

		Texas Dataset		PT_ASAG Dataset	
Features/Prompt	Model	MAE	RMSE	MAE	RMSE
TFIDF	SVM	0.65	1.15	0.69	1.12
TFIDF	RF	0.73	1.21	0.67	1.09
TFIDF	DT	0.88	1.32	0.63	1.02
TFIDF	XGB	0.76	1.24	0.66	1.08
BERT	SVM	0.71	1.19	0.74	1.16
BERT	RF	0.96	1.44	0.79	1.21
BERT	DT	0.96	1.44	0.79	1.21
BERT	XGB	0.69	1.18	0.72	1.14
FastText	BILSTM	0.78	1.26	0.99	1.38
Glove	BILSTM	0.87	1.31	0.73	1.10
BERT	BERT	0.64	1.11	0.58	0.96
few-shot	GPT4o	0.69	1.01	0.93	1.42
context + role + few-shot	GPT4o	0.70	1.04	0.90	1.39
role + few-shot	GPT4o	0.71	1.06	0.80	1.27
think + rubric + justification	GPT4o	1.37	1.71	0.51	0.94
think + step by step + rubric + justification	GPT4o	1.28	1.57	0.56	1.02
step by step + rubric	GPT4o	1.22	1.52	0.59	1.06

Table 7: Summary of the answers to this study's research questions (RQs).

RQ	Anwer
Which elements of prompt engineering can improve the performance of LLMs in the context of ASAG for Brazilian Portuguese and English?	Prompt engineering elements such as "few-shot" examples and "Time-to-Think" play significant roles in improving the performance of LLMs, though their effectiveness varies depending on the dataset and language, where the combination of "few-shot" and "Context" was found to be the most impactful in English, while "think" and "rubric" were the leading contributors in Brazilian Portuguese.
To what extent can LLM models outperform traditional machine learning models in the context of ASAG?	LLMs underperform compared to traditional machine learning models even after optimizing prompt engineering strategies, with traditional models consistently yielding lower error rates in terms of both MAE and RMSE for both English and Brazilian Portuguese answers.
How does the classification performance of traditional machine learning models and LLMs vary across different questions?	The performance of traditional machine learning models and LLMs remains consistent across new and previously trained questions, although traditional models continue to have an edge.

6 Limitation and Future Directions

We acknowledge the following limitations of the study. *First*, while various prompt components were evaluated, the precise composition of specific components 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 craft more detailed and specific prompt components. This approach is expected to provide a deeper understanding of the impact of each component's articulation on performance. Second, while we employed two datasets, our experimentation was restricted to 30% of the dataset due to cost constraints. However, it is noteworthy that multiple prior studies in the field of learning analytics and NL 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 may 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.

Acknowledgments

This study was partially supported by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (310888/2021-2). Also, we would like to thank OpenAI research Program for offering the credits required for this experiment.

References

[1] Sridevi Bonthu, S Rama Sree, and MHM Krishna Prasad. 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. Springer, 61–78.

[2] S Brindha, K Prabha, and Sandeep Sukumaran. 2016. A survey on classification techniques for text mining. In *2016 3rd International Conference on Advanced Computing and Communication Systems (ICACCS)*, Vol. 1. IEEE, 1–5.

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL]

[4] Steven Burrows, Iryna Gurevych, and Benno Stein. 2015. The eras and trends of automatic short answer grading. *International journal of artificial intelligence in education* 25 (2015), 60–117.

[5] Leon Camus and Anna Filighera. 2020. 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*. Springer, 43–48.

[6] Leon Camus and Anna Filighera. 2020. Investigating Transformers for Automatic Short Answer Grading. In *Artificial Intelligence in Education*, Ig Ibert Bittencourt, Mutlu Cukurova, Kasia Muldner, Rose Luckin, and Eva Millán (Eds.). Springer International Publishing, Cham, 43–48.

[7] Dan Carpenter, Wookhee Min, Seung Lee, Gamze Ozogul, Xiaoying Zheng, and James Lester. 2024. Assessing Student Explanations with Large Language Models Using Fine-Tuning and Few-Shot Learning. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, Ekaterina Kochmar, Marie Bexte, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Anaïs Tack, Victoria Yaneva, and Zheng Yuan (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 403–413. https://aclanthology.org/2024.bea-1.33

[8] Dan Carpenter, Wookhee Min, Seung Lee, Gamze Ozogul, Xiaoying Zheng, and James Lester. 2024. Assessing student explanations with large language models using fine-tuning and few-shot learning. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*. 403–413.

[9] Imran Chamieh, Torsten Zesch, and Klaus Giebermann. 2024. LLMs in Short Answer Scoring: Limitations and Promise of Zero-Shot and Few-Shot Approaches. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*. 309–315.

[10] Li-Hsin Chang and Filip Ginter. 2024. Automatic Short Answer Grading for Finnish with ChatGPT. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 23173–23181.

[11] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology* 15, 3 (2024), 1–45.

[12] Benjamin Clavié and Kobi Gal. 2019. Edubert: Pretrained deep language models for learning analytics. *arXiv preprint arXiv:1912.00690* (2019).

[13] Aubrey Condor, Max Litster, and Zachary A. Pardos. 2021. Automatic short answer grading with SBERT on out-of-sample questions. In *Educational Data Mining*. https://api.semanticscholar.org/CorpusID:236880619

[14] Emiliano del Gobbo, Alfonso Guarino, Barbara Cafarelli, and Luca Grilli. 2023. GradeAid: a framework for automatic short answers grading in educational contexts—design, implementation and evaluation. *Knowledge and Information Systems* 65, 10 (2023), 4295––4334. https://doi.org/10.1007/s10115-023-01892-9

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[16] Bronwyn Eager and Ryan Brunton. 2023. Prompting higher education towards AI-augmented teaching and learning practice. *Journal of University Teaching & Learning Practice* 20, 5 (2023), 02.

[17] Rafael Ferreira, Elyda Freitas, Luciano Cabral, Filipe Dawn, Luiz Rodrigues, Mladen Rakovic, Jackson Raniel, and Dragan Gasevic. 2024. Words of Wisdom: A Journey through the Realm of NLP for Learning Analytics-A Systematic Literature Review. *Journal of Learning Analytics* (2024), 1–24.

[18] Lucas Galhardi, Rodrigo C Thom de Souza, and Jacques Brancher. 2020. Automatic grading of portuguese short answers using a machine learning approach. In *Anais Estendidos do XVI Simpósio Brasileiro de Sistemas de Informação*. SBC, 109–124.

[19] Lucas Busatta Galhardi and Jacques Duílio Brancher. 2018. Machine learning approach for automatic short answer grading: A systematic review. In *Advances in Artificial Intelligence-IBERAMIA 2018: 16th Ibero-American Conference on AI, Trujillo, Peru, November 13-16, 2018, Proceedings 16*. Springer, 380–391.

[20] Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Pal: Program-aided language models. In *International Conference on Machine Learning*. PMLR, 10764–10799.

[21] Louie Giray. 2023. Prompt Engineering with ChatGPT: A Guide for Academic Writers. *Annals of Biomedical Engineering* (2023), 1–5.

[22] Luis Gutiérrez and Brian Keith. 2019. A systematic literature review on word embeddings. In *Trends and Applications in Software Engineering: Proceedings of the 7th International Conference on Software Process Improvement (CIMPS 2018) 7*. Springer, 132–141.

[23] Jochen Hartmann, Juliana Huppertz, Christina Schamp, and Mark Heitmann. 2019. Comparing automated text classification methods. *International Journal of Research in Marketing* 36, 1 (2019), 20–38.

[24] Owen Henkel, Libby Hills, Adam Boxer, Bill Roberts, and Zach Levonian. 2024. Can Large Language Models Make the Grade? An Empirical Study Evaluating LLMs Ability To Mark Short Answer Questions in K-12 Education. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale* (Atlanta, GA, USA) *(L@S '24)*. Association for Computing Machinery, New York, NY, USA, 300–304. https://doi.org/10.1145/3657604.3664693

[25] Abdullah İhsanoğlu, Mounes Zaval, and Olcay Taner Yıldız. 2024. Comparison of Machine Learning Algorithms and Large Language Models for Product Categorization. In *2024 32nd Signal Processing and Communications Applications Conference (SIU)*. IEEE, 1–4.

[26] 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*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 14197–14203. https://doi.org/10.18653/v1/2023.findings-emnlp.946

[27] Faiza Khan Khattak, Serena Jeblee, Chloé Pou-Prom, Mohamed Abdalla, Christopher Meaney, and Frank Rudzicz. 2019. A survey of word embeddings for clinical text. *Journal of Biomedical Informatics* 100 (2019), 100057.

[28] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems* 35 (2022), 22199–22213.

[29] Junyoung Lee and Elizabeth Koh. 2023. Teamwork dimensions classification using bert. In *International conference on artificial intelligence in education*. Springer, 254–259.

[30] Jiazheng Li, Lin Gui, Yuxiang Zhou, David West, Cesare Aloisi, and Yulan He. 2023. Distilling ChatGPT for Explainable Automated Student Answer Assessment. *arXiv preprint arXiv:2305.12962* (2023).

[31] Zongmin Li, Qi Zhang, Yuhong Wang, and Shihang Wang. 2020. Social media rumor refuter feature analysis and crowd identification based on XGBoost and NLP. *Applied Sciences* 10, 14 (2020), 4711.

[32] Joseph P Magliano and Arthur C Graesser. 2012. Computer-based assessment of student-constructed responses. *Behavior Research Methods* 44 (2012), 608–621.

[33] Christopher Manning. 1999. *Foundations of statistical natural language processing*. The MIT Press.

[34] Elijah Mayfield and Alan W Black. 2020. Should you fine-tune BERT for automated essay scoring?. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*. 151–162.

[35] Rafael Mello, Luiz Rodrigues, Luciano Cabral, Filipe Pereira, Cleon Pereira Júnior, Dragan Gasevic, and Geber Ramalho. 2024. Prompt Engineering for Automatic Short Answer Grading in Brazilian Portuguese. In *Anais do XXXV Simpósio Brasileiro de Informática na Educação* (Rio de Janeiro/RJ). SBC, Porto Alegre, RS, Brasil, 1730–1743. https://doi.org/10.5753/sbie.2024.242424

[36] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2017. Advances in pre-training distributed word representations. *arXiv preprint arXiv:1712.09405* (2017).

[37] Michael Mohler, Razvan Bunescu, and Rada Mihalcea. 2011. Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In *HLT '11: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*. Association for Computational Linguistics, 752—-762. https://dl.acm.org/doi/10.5555/2002472.2002568

[38] Michael Mohler and Rada Mihalcea. 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)*. 567–575.

[39] Ben Naismith, Phoebe Mulcaire, and Jill Burstein. 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)*. 394–403.

[40] Huy A Nguyen, Hayden Stec, Xinying Hou, Sarah Di, and Bruce M McLaren. 2023. Evaluating chatgpt's decimal skills and feedback generation in a digital learning game. In *European Conference on Technology Enhanced Learning*. Springer, 278–293.

[41] Hilário Oliveira, Rafael Ferreira Mello, Péricles Miranda, Bruno Alexandre, Thiago Cordeiro, Ig Ibert Bittencourt, and Seiji Isotani. 2023. Classificaçao ou Regressao? Avaliando Coesao Textual em Redaçoes no contexto do ENEM. In *Anais do XXXIV Simpósio Brasileiro de Informática na Educação*. SBC, 1226–1237.

[42] Shweta Patil and Krishnakant P Adhiya. 2022. Automated evaluation of short answers: A systematic review. *Intelligent Data Communication Technologies and Internet of Things: Proceedings of ICICI 2021* (2022), 953–963.

[43] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[44] Marko Putnikovic and Jelena Jovanovic. 2023. Embeddings for Automatic Short Answer Grading: A Scoping Review. *IEEE Transactions on Learning Technologies* (2023).

[45] Dadi Ramesh and Suresh Kumar Sanampudi. 2022. An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review* 55, 3 (2022), 2495–2527.

[46] Chaudhary Jashubhai Rameshbhai and Joy Paulose. 2019. Opinion mining on newspaper headlines using SVM and NLP. *International journal of electrical and computer engineering (IJECE)* 9, 3 (2019), 2152–2163.

[47] Luiz Rodrigues, Filipe Dwan Pereira, Luciano Cabral, Dragan Gašević, Geber Ramalho, and Rafael Ferreira Mello. 2024. Assessing the quality of automatic-generated short answers using GPT-4. *Computers and Education: Artificial Intelligence* 7 (2024), 100248.

[48] Archana Sahu and Plaban Kumar Bhowmick. 2020. Feature Engineering and Ensemble-Based Approach for Improving Automatic Short-Answer Grading Performance. *IEEE Transactions on Learning Technologies* 13, 1 (2020), 77–90. https://doi.org/10.1109/TLT.2019.2897997

[49] Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management* 24, 5 (1988), 513–523.

[50] S Selva Birunda and R Kanniga Devi. 2021. A review on word embedding techniques for text classification. *Innovative Data Communication Technologies and Application: Proceedings of ICIDCA 2020* (2021), 267–281.

[51] Chamuditha Senanayake and Dinesh Asanka. 2024. Rubric Based Automated Short Answer Scoring using Large Language Models (LLMs). In *2024 International Research Conference on Smart Computing and Systems Engineering (SCSE)*, Vol. 7. IEEE, 1–6.

[52] Sima Siami-Namini, Neda Tavakoli, and Akbar Siami Namin. 2019. The performance of LSTM and BiLSTM in forecasting time series. In *2019 IEEE International conference on big data (Big Data)*. IEEE, 3285–3292.

[53] Chul Sung, Tejas Indulal Dhamecha, and Nirmal Mukhi. 2019. Improving Short Answer Grading Using Transformer-Based Pre-training. In *Artificial Intelligence in Education*, Seiji Isotani, Eva Millán, Amy Ogan, Peter Hastings, Bruce McLaren, and Rose Luckin (Eds.). Springer International Publishing, Cham, 469–481.

[54] Neslihan Süzen, Alexander N Gorban, Jeremy Levesley, and Evgeny M Mirkes. 2020. Automatic short answer grading and feedback using text mining methods. *Procedia computer science* 169 (2020), 726–743.

[55] Felipe Urrutia and Roberto Araya. 2024. Who's the Best Detective? Large Language Models vs. Traditional Machine Learning in Detecting Incoherent Fourth Grade Math Answers. *Journal of Educational Computing Research* 61, 8 (2024), 187–218.

[56] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.

[57] 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* (2023).

[58] Lixiang Yan, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martinez-Maldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin, and Dragan Gašević. [n.d.]. Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology* n/a, n/a ([n. d.]). https://doi.org/10.1111/bjet.13370

[59] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).

[60] Araz Zirar. 2023. Exploring the impact of language models, such as ChatGPT, on student learning and assessment. *Review of Education* 11, 3 (2023).