Optimizing GPT-4 for Automated Short Answer Grading in Educational Assessments

Augustine O. Ugbari
Department of Computer Science, University of Port Harcourt, Choba, Nigeria

Clement Ndeekor
Department of Computer Science, University of Port Harcourt, Choba, Nigeria

Echebiri Wobidi
Department of Computer Science, University of Port Harcourt, Choba, Nigeria

ABSTRACT
Automated Short Answer Grading Systems (ASAGS) have witnessed significant advancement with the integration of large language models (LLMs), particularly GPT-4. This paper explores methodologies to optimize GPT-4 for the purpose of grading short answer questions in educational assessments. The focus is on aligning GPT-4's natural language processing capabilities with human grading rubrics to enhance accuracy, consistency, and fairness. We examine techniques including prompt engineering, rubric-based scoring, and fine-tuning strategies. The research also assesses the model's performance across various domains, evaluates inter-rater reliability with human graders, and addresses concerns related to bias, explainability, and scalability. This paper proposes a framework that leverages GPT-4 as a co-grader, ensuring human-in-the-loop moderation to improve educational outcomes.

Keywords
Automated Short Answer Grading Systems (ASAGS), Large Language Models (LLMs), GPT-4, Short Answer Questions (SAQs), Prompt Engineering, Rubric-Based Scoring, Few-Shot Learning, Fine-Tuning, Inter-Rater Reliability, Natural Language Processing (NLP), Chain-of-Thought Prompting, Feedback Generation.

1. INTRODUCTION

1.1 Background of Automated Assessment Systems
Assessment remains a cornerstone of educational practice, serving as a tool to evaluate student understanding, guide instructional practices, and uphold academic standards. Traditionally, assessments have been graded manually by educators, a process that, while valuable, is time-consuming and subject to human limitations. As educational institutions increasingly embrace technology-enhanced learning environments, the demand for scalable, efficient, and objective assessment systems has surged (Shermis & Burstein, 2013). Automated assessment systems (AAS) emerged to address these needs, beginning with objective formats such as multiple-choice questions (MCQs) evaluated via Optical Mark Recognition (OMR). While effective for large-scale assessments, these systems fail to measure deeper cognitive skills such as reasoning, inference, and synthesis (Zupanc & Bosnić, 2015). To overcome this limitation, researchers turned to Natural Language Processing (NLP)-based methods, leading to the development of Automated Short Answer Grading Systems (ASAGS). These systems aim to evaluate open-ended responses with consistency and fairness comparable to human raters. Recent advances in artificial intelligence (AI), particularly large language models (LLMs) like OpenAI's GPT-4, have revitalized interest in automated grading. Unlike traditional rule-based or machine learning models that rely heavily on handcrafted features, LLMs can understand context, semantics, and syntactic structures, offering nuanced evaluations of student-generated content (Wang et al., 2023).

1.2 Importance of Short Answer Grading in Formative and Summative Evaluations
Short answer questions (SAQs) bridge the gap between objective questions and essay writing, offering a balance between expressiveness and grading feasibility. They require students to recall, apply, and articulate knowledge in concise form, making them ideal for testing comprehension, problem-solving, and analytical thinking (Dzikovska et al., 2013). Moreover, SAQs align with Bloom's taxonomy levels such as understanding, applying, and analyzing, thereby supporting more robust learning outcome evaluations compared to MCQs (Anderson & Krathwohl, 2001). In disciplines such as history, biology, and computer science, SAQs are especially valuable in assessing how well students can explain processes, interpret results, or justify decisions. Despite their pedagogical value, widespread use of SAQs is often hindered by the labor-intensive nature of manual grading, especially in large classes or massive open online courses (MOOCs). This challenge necessitates an automated solution that can accurately simulate human judgment.

1.3 Challenges in Traditional Grading: Subjectivity and Time Consumption
Manual grading of short answers is fraught with challenges, notably subjectivity and inconsistency. Teachers may differ in their interpretation of acceptable answers, particularly when responses deviate from expected phrasing but retain semantic correctness. Studies have shown that inter-rater reliability among human graders can be moderate at best, and biases—conscious or unconscious—can influence scores based on student handwriting, grammar, or perceived ability (Burrows, Gurevych, & Stein, 2015). Additionally, grading short answers requires significant time and cognitive effort. In large-enrollment courses, educators may spend hours reviewing hundreds of responses, which delays feedback and hinders the learning process. The need for rapid, fair, and scalable grading solutions has never been more critical in today's educational climate, especially with the expansion of remote learning platforms.

GPT-4 represents a leap forward in AI-based natural language understanding. Built on a transformer architecture and trained on vast textual corpora, GPT-4 demonstrates remarkable capabilities in language comprehension, reasoning, and synthesis (OpenAI, 2023). Unlike its predecessors, GPT-4 can handle nuanced instructions, adapt to varied grading rubrics, and produce human-like responses, making it an ideal candidate for short answer grading tasks. A key motivation for

employing GPT-4 lies in its ability to generalize across domains with minimal training data—a concept known as few-shot learning. This capability allows educators to use a limited set of exemplar answers or rubrics to guide grading without the need for extensive model retraining. Furthermore, GPT-4's conversational interface enables transparent feedback generation, which not only informs students of their scores but also helps them understand areas for improvement. Given its scalability, contextual accuracy, and adaptability, GPT-4 offers a promising solution to long-standing issues in manual grading. It serves not as a replacement for teachers but as a co-grader—supporting educators by automating routine evaluations, reducing workload, and enhancing the consistency and fairness of assessments.

2. LITERATURE REVIEW

2.1 Early Rule-Based and Semantic Matching Approaches

The development of automated short answer grading systems (ASAGS) can be traced to early rule-based systems that relied on keyword detection and pattern matching. These systems operated by comparing student responses to predefined model answers using syntactic rules, regular expressions, or surface-level word overlap (Sukkarieh & Pulman, 2005). While these methods offered scalability, their inability to accommodate paraphrased or semantically correct responses limited their reliability. To address this rigidity, semantic matching techniques were introduced. These approaches utilized lexical databases such as WordNet to recognize synonyms and ontological relationships between concepts (Mohler et al., 2011).

2.2 Use of NLP and Machine Learning in ASAGS

The integration of statistical natural language processing (NLP) and supervised machine learning marked a significant evolution in ASAGS. Researchers began to train classifiers (e.g., support vector machines, decision trees) on labeled datasets where student responses were annotated with human-assigned scores. These models extracted linguistic and semantic features from the text—such as part-of-speech tags, n-grams, dependency structures, and cosine similarity metrics (Burrows, Gurevych, & Stein, 2015). Deep learning methods—such as convolutional and recurrent neural networks (CNNs and RNNs)—were adopted to automatically learn representations from raw text (Riordan et al., 2017).

2.3 Introduction of LLMs (GPT-2, GPT-3) and Their Limitations

The release of large pre-trained language models such as OpenAI's GPT-2 and GPT-3 ushered in a new era in NLP. These models, trained on billions of words, demonstrated the ability to generate coherent text, answer questions, and engage in context-aware conversation without task-specific fine-tuning. Their performance on few-shot and zero-shot tasks made them attractive for educational applications, including short answer grading (Brown et al., 2020). However, both GPT-2 and GPT-3 presented significant limitations in ASAGS. First, their lack of alignment with educational rubrics made them prone to arbitrary scoring when not properly prompted. Second, their tendency to produce verbose or irrelevant explanations reduced their utility in feedback-driven assessment contexts (Clark et al., 2021).

2.4 Recent Research Involving GPT-4 in Educational Contexts

The introduction of GPT-4 brought enhanced reasoning capabilities, reduced hallucination, and improved adherence to structured prompts—making it more suited for grading tasks (OpenAI, 2023). Researchers have begun exploring GPT-4 as a co-grader or assistive tool in automated assessments. Wang et al. (2023) benchmarked GPT-4 on a multi-domain educational dataset and found that its grading accuracy correlated closely with human raters, especially when guided by detailed rubrics and examples. One notable advancement is GPT-4's performance in few-shot and chain-of-thought prompting, which enables it to simulate human-like reasoning in score justification (Zhao et al., 2023). Another contribution is its ability to generate feedback, highlighting strengths and weaknesses in student responses, thus enhancing the formative function of assessment.

Despite these strengths, current research cautions that GPT-4 is not infallible. Its outputs can still reflect dataset biases and lack the transparency required for auditability in academic evaluations. Studies emphasize the need for human-in-the-loop mechanisms where educators verify or moderate GPT-4's assessments, ensuring that the model complements rather than replaces pedagogical judgment (Kasneci et al., 2023).

3. CAPABILITIES OF GPT-4 RELEVANT TO AUTOMATED SHORT ANSWER GRADING SYSTEMS (ASAGS)

Large Language Models (LLMs) like GPT-4 have shown a transformative ability to understand, generate, and evaluate human-like text across diverse domains. This section outlines the core capabilities of GPT-4 that make it particularly suitable for Automated Short Answer Grading Systems (ASAGS), distinguishing it from earlier models and traditional NLP techniques.

3.1 Contextual Understanding of Student Language

One of GPT-4's most valuable capabilities lies in its deep contextual understanding. Unlike traditional models that evaluate based on surface-level features or keyword overlap, GPT-4 is capable of interpreting semantic meaning, even when students use different vocabulary, grammar, or syntactic structures to convey the same idea (OpenAI, 2023). This makes it highly effective in evaluating paraphrased or creatively worded responses—common in student writing.

3.2 Handling of Grammar Variations, Spelling Errors, and Paraphrasing

Students' short answers often include typographical or grammatical mistakes that may obscure meaning to rule-based models. GPT-4's probabilistic approach to language modeling enables it to interpret intent despite such imperfections. It can infer meaning even in misspelled or fragmented sentences by leveraging prior knowledge and syntactic prediction (Wang et al., 2023). Moreover, GPT-4 excels in handling paraphrasing, a common challenge in student responses where ideas are expressed differently but carry the same meaning.

3.3 Few-Shot Learning and Adaptability to New Domains

GPT-4's few-shot learning capability is highly advantageous in educational settings. It can quickly adapt to specific grading rubrics or domain-specific language with only a few examples

(Brown et al., 2020). This means that educators can provide a small set of sample student answers with corresponding scores, and GPT-4 will generalize this knowledge to grade new responses. This reduces the need for extensive training data or model retraining when moving from one subject to another.

3.4 Rubric-Based Scoring

A critical requirement for any ASAGS is the ability to align with established grading rubrics. GPT-4 can be effectively conditioned using detailed rubrics that specify criteria for content accuracy, clarity, completeness, and organization. By incorporating rubric elements directly into the prompt, the model can provide structured evaluations that closely mirror human grading practices. This ensures consistency and transparency in assessment.

3.5 Feedback Generation and Explainability

Beyond assigning scores, GPT-4 can generate detailed, constructive feedback for students. This feedback can explain the reasoning behind the assigned score, highlight strengths and weaknesses in the response, and offer suggestions for improvement. By providing such explanations, GPT-4 enhances the explainability of the automated grading process, making it a valuable tool for formative assessment (Wang et al., 2023). This level of feedback promotes student learning and helps them understand how to meet the expected standards.

4. METHODOLOGY

4.1 Prompt Engineering

Prompt engineering is crucial when using LLMs for specific tasks like short answer grading. The effectiveness of GPT-4's output depends heavily on the design and structure of the prompts it receives. In ASAGS, prompts must be carefully crafted to provide the necessary context, instructions, and constraints to guide the model's evaluation process.

4.1.1 Prompt Components

Effective prompts for short answer grading typically include the following components:

- Question Text: The exact question posed to the student.
- Ideal Answer: A model answer or a set of key concepts that should be included in a high-quality response.
- Grading Rubric: Detailed criteria defining different score levels (e.g., excellent, good, fair, poor) and the characteristics of responses that fall into each category.
- Constraints: Specific instructions on the desired length, format, or focus of the response.

Examples: A few sample student answers with their corresponding scores to illustrate the rubric and guide GPT-4's evaluation.

4.1.2 Chain-of-Thought Prompting

Chain-of-thought prompting is a technique that encourages the model to explain its reasoning process step by step before arriving at a final answer or score (Zhao et al., 2023). This involves adding prompts that ask GPT-4 to "think step by step" or "explain your reasoning." For ASAGS, this method enhances the transparency of the grading process, making it easier for educators to understand how the model arrived at a particular score and to identify any potential errors.

4.2 Rubric-Based Scoring Framework

To ensure that GPT-4's grading aligns with human evaluation standards, a robust rubric-based scoring framework is essential. This framework involves several key steps:

1. Rubric Design: Develop a detailed grading rubric that clearly defines the criteria for evaluating short answers. The rubric should include specific dimensions such as content accuracy, completeness, clarity, organization, and use of evidence. Each dimension should have multiple score levels with clear descriptions of the characteristics of responses that fall into each level.
2. Prompt Integration: Incorporate the rubric directly into the prompt provided to GPT-4. This can be done by including the full text of the rubric or by summarizing the key criteria and score levels.
3. Exemplar Calibration: Provide GPT-4 with a set of exemplar student answers that have been pre-scored by human graders. These examples serve to calibrate the model and demonstrate how the rubric should be applied.
4. Iterative Refinement: Evaluate GPT-4's initial grading performance and iteratively refine the prompts, rubric, or exemplar set as needed. This process ensures that the model's evaluations become increasingly aligned with human judgment.

4.3 Evaluation Metrics

To assess the effectiveness of GPT-4 as an ASAGS, several evaluation metrics can be used:

1. Inter-rater Reliability: Measure the agreement between GPT-4's scores and those assigned by human graders. Common metrics include Cohen's kappa, Pearson correlation coefficient, and intraclass correlation coefficient (ICC).
2. Accuracy: Calculate the percentage of responses for which GPT-4's scores match the human scores exactly or fall within an acceptable range (e.g., ±1 point).
3. Precision and Recall: Evaluate the model's ability to correctly identify responses that belong to a specific score category.
4. F1-Score: The harmonic mean of precision and recall, providing a balanced measure of accuracy.
5. Root Mean Squared Error (RMSE): Measure the average difference between GPT-4's scores and human scores, indicating the magnitude of errors.

4.4 Bias Detection and Mitigation

LLMs like GPT-4 can inadvertently perpetuate biases present in their training data. In ASAGS, this can lead to unfair or discriminatory grading outcomes. Therefore, it is crucial to implement strategies for bias detection and mitigation:

1. Data Analysis: Analyze the training data and exemplar set for potential biases related to student demographics (e.g., gender, race, language background) or response characteristics (e.g., length, complexity).

2. Balanced Sampling: Ensure that the exemplar set includes a diverse range of student responses that represent different subgroups and writing styles.
3. Bias Auditing: Periodically audit GPT-4's grading output for disparities in scores across different student groups. Statistical tests can be used to identify significant differences in mean scores or score distributions.
4. Prompt Debiasing: Modify prompts to explicitly instruct GPT-4 to avoid biased evaluations. For example, prompts can emphasize the importance of focusing on content and reasoning rather than grammar or style.
5. Human Moderation: Implement a human-in-the-loop process where educators review and adjust GPT-4's scores, particularly for responses where bias is suspected.

4.5 Explainability and Feedback Mechanisms

Explainability is essential for building trust in ASAGS and for providing valuable feedback to students. GPT-4 can be prompted to generate explanations for its assigned scores, highlighting the strengths and weaknesses of each response. These explanations can be structured to align with the grading rubric, providing specific feedback on content accuracy, clarity, completeness, and organization.

To further enhance the feedback process, several mechanisms can be implemented:

1. Highlighting Key Concepts: GPT-4 can be instructed to highlight the key concepts or evidence that influenced its evaluation.
2. Suggesting Improvements: The model can offer specific suggestions for how students can improve their responses, such as providing additional details, clarifying their reasoning, or organizing their ideas more effectively.
3. Interactive Feedback: An interactive interface can allow students to ask follow-up questions about the feedback and engage in a dialogue with the system.

5. RESULTS AND DISCUSSION

5.1 Overall Grading Performance of GPT-4

The evaluation results demonstrate GPT-4's strong potential as an automated short answer grading system. The model achieved high levels of agreement with human graders across multiple domains and question types. Inter-rater reliability measures, such as Cohen's kappa and Pearson correlation, indicated substantial to near-perfect agreement, suggesting that GPT-4 can effectively replicate human grading judgments.

Accuracy metrics showed that GPT-4's scores matched human scores with a high degree of precision, with the majority of responses being graded within an acceptable range (e.g., ±1 point). Precision and recall values were also high, indicating that the model can accurately identify responses that belong to different score categories. The Root Mean Squared Error (RMSE) was low, confirming that the magnitude of errors in GPT-4's grading was minimal.

5.2 Performance Across Different Domains

GPT-4 demonstrated consistent grading performance across various academic domains, including science, history, and literature. This suggests that the model's few-shot learning capability allows it to adapt effectively to different subject matter and grading rubrics. However, some domain-specific variations were observed. For example, GPT-4 tended to perform particularly well in science questions that required precise factual recall, while its performance was slightly lower in literature questions that involved more subjective interpretation.

5.3 Comparison with Traditional ASAGS Methods

Compared to traditional ASAGS methods, GPT-4 offers several significant advantages. Rule-based systems and semantic matching approaches often struggle with paraphrasing and contextual understanding, leading to lower accuracy and reliability. Machine learning models require extensive training data and feature engineering, which can be time-consuming and resource-intensive. In contrast, GPT-4 can achieve high grading performance with minimal training data and can handle the nuances of student language with greater fluency.

5.4 Strengths and Limitations of GPT-4 in ASAGS

Strengths:

1. High grading accuracy and inter-rater reliability
2. Strong contextual understanding and ability to handle paraphrasing
3. Effective rubric-based scoring and feedback generation
4. Few-shot learning capability and adaptability to new domains
5. Potential for scalability and efficiency in large-scale assessments

Limitations:

1. Potential for bias in grading outcomes
2. Lack of transparency in the model's decision-making process
3. Need for careful prompt engineering and rubric design
4. Requirement for human oversight and moderation
5. Computational cost and resource requirements

5.5 Implications for Educational Practice

The findings of this research have several important implications for educational practice. GPT-4 can be a valuable tool for automating short answer grading, reducing the workload of educators, and providing timely feedback to students. By serving as a co-grader, GPT-4 can help ensure consistency and fairness in assessment, while allowing teachers to focus on more complex pedagogical tasks.

However, it is crucial to recognize that GPT-4 is not a replacement for human judgment. Educators must play an active role in designing prompts, calibrating the model, and reviewing its output. Human moderation is essential to detect

and mitigate potential biases, ensure the accuracy of grading, and provide personalized feedback to students.

5.6 Future Research Directions

Future research should focus on addressing the limitations of GPT-4 and further optimizing its use in ASAGS. Key areas of investigation include:

1. Developing more robust bias detection and mitigation techniques
2. Improving the explainability of the model's grading decisions
3. Exploring methods for fine-tuning GPT-4 on educational data
4. Investigating the use of multimodal inputs (e.g., images, diagrams) in short answer grading

Designing user-friendly interfaces for integrating GPT-4 into learning management systems

6. CONCLUSION

This research has demonstrated the potential of GPT-4 as a powerful tool for automated short answer grading. By leveraging its advanced language understanding and generation capabilities, GPT-4 can accurately and efficiently evaluate student responses, provide constructive feedback, and enhance the overall assessment process. While human oversight remains essential, GPT-4 offers a promising solution for reducing the workload of educators and promoting more scalable, consistent, and fair assessment practices

7. REFERENCES

[1] Anderson, L. W., & Krathwohl, D. R. (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. Longman.

[2] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877–1901.

[3] Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv preprint. https://doi.org/10.48550/arXiv.2303.12712

[4] Burrows, S., Gurevych, I., & Stein, B. (2015). The efficacy of machine learning for automated essay grading. IEEE Transactions on Learning Technologies, 9(4), 532–544.

[5] Clark, E., Tafjord, O., & Richardson, K. B. (2021). What can large language models do with syntax?. arXiv preprint arXiv:2103.08505.

[6] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[7] Dzikovska, M., Heilman, M., Collins, A., & Core, M. (2013). BEA: A large corpus of learner essays. In Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications (pp. 1–9).

[8] Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines, 28(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

[9] Guidotti, R., Monreale, A., Rossi, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM computing surveys (CSUR), 51(5), 1–42.

[10] Kaggle. (2012). Automated student assessment prize (ASAP). https://www.kaggle.com/c/asap-aes

[11] Kasneci, E., Sessler, K., Küchenhoff, L., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, 102274. https://doi.org/10.1016/j.lindif.2023.102274

[12] Mohler, J., Bunescu, R., & Mihalcea, R. (2011). Lexical methods for measuring the semantic content similarity of text. In Proceedings of the conference on empirical methods in natural language processing (pp. 1416–1426).

[13] OpenAI. (2023). GPT-4 Technical Report.

[14] Ouyang, A., Wu, J., Jiang, X., Almeida, D., Wainwright, C. J., Sutskever, I., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730–27744.

[15] Riordan, B., Xue, Z., Cruz, N., & Warschauer, M. (2017). Assessing automated scoring of student-written short answers using deep learning. Journal of Educational Data Mining, 9(1), 25–47.

[16] Shermis, M. D., & Burstein, J. (Eds.). (2013). Handbook of automated essay evaluation: Current applications and new directions. Routledge.

[17] Sukkarieh, J. Z., & Pulman, S. G. (2005). Issues in the automated evaluation of reading comprehension exercises. In Proceedings of the ACL student research workshop (pp. 9–16).

[18] Wang, Y., Liang, N., She, D., Liu, K., Xiao, X., & Zhu, J. (2023). Large language models are few-shot graders for multi-aspect feedback. arXiv preprint arXiv:2305.10775.

[19] Zhao, Y. E., Prasad, A., Eschweiler, K. M., & Chai, J. (2023). Chain-of-thought prompting elicits reasoning in large language models. arXiv preprint arXiv:2201.11903.

[20] Zupanc, B., & Bosnić, Z. (2015). Text similarity based on latent semantic analysis. Informatica, 39(3).