Leveraging Large Language Models for Arabic Short Answer Grading and Feedback Generation

Emad Nabil[1,*], Mostafa Mohamed Saeed[2], Rana Reda[3], Safiullah Faizullah[4], and Wael Hassan Gomaa[5]

[1]Faculty of Computer and Information Systems, Islamic University of Madinah, Madinah, Saudi Arabia
[2]Computational Approaches to Modeling Language (CAMeL) Lab, New York University, Abu Dhabi, United Arab Emirates
[3]Digital Egypt for Investment Co., Cairo, Egypt
[4]Faculty of Computer and Information Systems, Islamic University of Madinah, Madinah, Saudi Arabia
[5]Faculty of Computers and Artificial Intelligence, Beni-Suef University, Beni-Suef, Egypt
Email: e.nabil@fci-cu.edu.eg (E.N.); mms10094@nyu.edu (M.M.S.); rana.reda@defi.com.eg (R.R.); safi@iu.edu.sa (S.F.); wael.goma@gmail.com (W.H.G.)
*Corresponding author
Manuscript received May 7, 2025; revised June 10, 2025; accepted July 25, 2025; published December 16, 2025

***Abstract*—This paper explores the potential of Large Language Models (LLMs) in automating the grading and feedback generation of short-answer responses in Arabic. Arabic poses unique challenges due to its linguistic complexity and the relative scarcity of well-developed Natural Language Processing (NLP) resources compared to languages such as English and Chinese. The study evaluates both proprietary models (GPT-4) and open-source models (Llama 3-8B, Llama 3-70B, and DeepSeek-V3) using the Environmental Science Corpus—a custom-designed dataset tailored for Arabic short-answer assessment. Two core tasks are addressed: grading and feedback generation. In the grading task, DeepSeek-V3 achieved the best performance, with a Quadratic Weighted Kappa (QWK) score of 0.8273, a Pearson correlation of 86.09%, and a Root Mean Squared Error (RMSE) of 0.76, indicating near-perfect agreement with human evaluators. GPT-4 ranked second, followed by Llama 3-70B, while Llama 3-8B was the lowest-performing model. In feedback generation, DeepSeek-V3 again led the performance with a human evaluation score of 79.61% for generating accurate and constructive feedback. GPT-4 ranked second, followed by Llama 3 models. Statistical analysis using the Wilcoxon test revealed significant performance differences among all models ($p < 0.05$), indicating that each LLM offers unique capabilities in handling Arabic short-answer grading. Overall, the results underscore the effectiveness of LLMs in Arabic educational assessment, highlighting the critical role of prompt engineering in enhancing model performance. The study demonstrates that LLMs can not only grade student responses with high accuracy but also generate meaningful feedback, thereby supporting the development of more effective automated learning tools. Practical recommendations and best practices are presented to help educators and developers optimize the use of LLMs in Arabic-language educational settings, laying the groundwork for future advancements in Arabic NLP.**

***Keywords*—Arabic short answer grading, Large Language Models (LLMs), prompt engineering, GPT-4, Llama-3, DeepSeek-V3**

I. Introduction

Grading is a cornerstone of educational assessment, serving as a fundamental tool through which educators gauge student understanding and provide essential feedback to guide future learning. Despite its pivotal role, traditional grading methods are primarily reliant on manual processes and pose several challenges that compromise their effectiveness, especially as educational systems scale. Manual grading, while offering a personal touch, is inherently labor-intensive, time-consuming, and susceptible to various forms of subjectivity. These issues, including inconsistencies in grading and the difficulty in delivering personalized feedback on a large scale, have long been recognized as significant impediments to the overall educational experience [1].

Traditional approaches, such as rubric-based evaluations and norm-referenced grading, aim to provide a structured and objective framework for assessment. However, these methods often fall short of their intended goals due to the inherent subjectivity involved in human judgment. Instructors, despite their best efforts, may introduce conscious or unconscious biases into the grading process, leading to variability in assessment outcomes not only across different evaluators but also within the assessments of a single educator over time [2, 3]. This variability undermines the fairness and reliability of grading, raising concerns about the validity of the assessments. Moreover, the manual nature of these methods makes it challenging to provide timely feedback, particularly in large classes, where the volume of work to be graded can be overwhelming [4].

The limitations of traditional grading methods have prompted a growing interest in leveraging technological advancements to enhance the efficiency, consistency, and scalability of educational assessments. In this context, Large Language Models (LLMs) have emerged as a transformative force, reshaping the landscape of grading and feedback delivery. LLMs, powered by advancements in Natural Language Processing (NLP) and machine learning, have demonstrated remarkable capabilities in understanding, analyzing, and generating complex textual data [5, 6]. Their ability to process natural language with near-human proficiency positions them as ideal tools for automating various aspects of educational assessment, particularly in grading short answers and providing personalized feedback [7].

However, LLMs, whether open-source or closed-source, face significant limitations related to computational resources, time, and financial costs. These models, particularly those with billions of parameters, require substantial computational power for both fine-tuning and inference. Running inferences on these models often necessitates specialized hardware, such as Graphics Processing Units (GPUs) or Tensor Processing Units (TPUs), which can be costly and energy-intensive, leading to increased operational expenses.

Fine-tuning these models, especially on domain-specific

tasks, is even more resource-demanding, requiring large datasets, extended training times, and continuous access to high-performance computing infrastructure. Additionally, the financial burden associated with maintaining and scaling these models can be prohibitive, making it challenging for smaller organizations or researchers to fully leverage the potential of LLMs. Even with open-source models, the costs related to computational resources and time can outweigh the benefits, limiting their accessibility and usability in various applications.

LLMs have two promising applications in educational assessment. They significantly improve grading efficiency first. LLMs can quickly evaluate large amounts of student work by automating the review process, producing reliable results that reduce the subjectivity and unpredictability of manual grading [8]. In addition to saving time, this automation frees up teachers to focus on more complex teaching responsibilities like developing curricula and providing individualized student support. Second, LLMs are uniquely capable of providing customized comments based on each student's specific answers. Since this tailored feedback addresses each learner's particular needs and challenges, it is essential for fostering a more effective learning environment and enhancing the learner's overall academic growth [9].

The integration of LLMs into the grading process represents a paradigm shift that extends beyond mere efficiency improvements. It marks a move towards a more adaptive and expansive educational feedback mechanism capable of evolving with the needs of both students and educators. Unlike traditional methods, which often struggle to scale without compromising quality, LLMs offer a scalable solution that maintains high standards of accuracy and consistency in assessment [10]. Additionally, the ability of LLMs to provide context-aware feedback further enriches the educational experience, helping students to understand their mistakes better and learn from them in a meaningful way [11].

The concept of prompt engineering is particularly relevant in the context of automatic scoring systems, where the challenge lies in the models' ability to accurately interpret and evaluate student responses. By utilizing prompt engineering, researchers can guide LLMs to focus on specific aspects of a student's answer, improving the model's ability to capture nuances in reasoning and content-specific knowledge [12]. This method not only reduces the dependency on extensive training datasets but also enhances the flexibility of LLMs in adapting to various educational tasks.

Large language models, such as those from the Generative Pre-trained Transformer (GPT) family, have demonstrated considerable potential in this domain. However, the effectiveness of these models varies depending on the specific task and how the prompts are engineered. For instance, the integration of Chain-of-Thought (CoT) prompting methods, where the model is guided through a sequence of intermediary reasoning steps, has been shown to significantly enhance the accuracy of automatic scoring systems in science education [13]. This approach aligns more closely with human scoring outcomes, as it allows the model to mimic the thought process of a human grader.

While LLMs have been applied to automated assessment in English, to the best of our knowledge, no previous work has explored their use for the domain-specific task of Arabic Short Answer Grading (ASAG). This highlights a significant gap in the literature, particularly given the linguistic complexity of the Arabic language.

Our work goes beyond simply adapting English-based approaches by directly addressing the unique challenges posed by Arabic, such as its rich morphology and inherent ambiguity. In addition, we introduce a novel feedback generation component, a task that has not yet been explored in Arabic educational contexts. To evaluate the quality of the generated feedback, we incorporate human judgment through expert scoring.

The proposed pipeline demonstrates the effectiveness of LLMs in ASAG. This contribution represents a meaningful step toward advancing automated educational assessment in Arabic and clearly differentiates our work from prior studies focused on English.

Despite the promise of LLMs in automatic scoring, challenges remain, particularly in ensuring that these models can fully grasp the depth of content-specific knowledge and the rationale behind students' answers. The selection of appropriate LLM models and the tuning of hyperparameters are critical factors that influence the accuracy and reliability of automatic scoring [14, 15]. As research continues to explore the potential of different LLM variants, it becomes evident that prompt engineering is not just an ancillary technique but a foundational aspect of leveraging AI in educational assessments.

The effectiveness of prompt engineering, coupled with the strategic application of LLMs, particularly in scenarios involving zero-shot and few-shot learning, has the potential to revolutionize automatic scoring. As studies delve deeper into the optimal configurations for these models, including the impact of versions and hyperparameters [16], the field moves closer to realizing the full potential of AI in education. This exploration is not just about improving scoring accuracy but also about understanding how LLMs can be harnessed to provide explainable, transparent, and fair assessments, thereby enhancing the overall learning experience for students.

Implementing LLMs in educational assessment presents challenges. One major concern is that if these models are not properly managed, they could reinforce biases present in their training data, leading to biased grading results [17]. This highlights the need for ongoing human oversight and the development of strategies to counteract such biases to ensure the validity and reliability of LLM-based evaluations. Additionally, it is crucial to carefully consider the ethical implications of using AI-driven systems in education, particularly regarding data protection and the transparency of the decision-making processes these models employ [1].

In conclusion, the incorporation of LLMs into the grading process offers a promising solution to the limitations of traditional assessment methods. By automating the evaluation of short answers and providing personalized feedback, LLMs have the potential to enhance the accuracy, consistency, and scalability of educational assessments. This technological advancement not only addresses the existing challenges of manual grading but also opens new possibilities for creating more adaptable and effective educational

feedback mechanisms. As educational institutions increasingly explore the integration of LLMs into their assessment practices, it is crucial to continue evaluating the impact of these models on both the quality of education and the fairness of the grading process.

A. Research Questions

The research questions of the paper are listed below:

- Can LLMs effectively automate the grading of Arabic short answers?
- Can LLMs generate meaningful feedback for incomplete or incorrect Arabic short answers?
- How critical is prompt engineering in optimizing LLM performance for Arabic ASAG?
- What specific prompt engineering techniques prove most effective for Arabic ASAG applications?
- Among GPT-4, DeepSeek-V3, and Llama-3, which model delivers the highest grading accuracy for Arabic short answers?
- Among GPT-4, DeepSeek-V3, and Llama-3, which model generates the most educationally valuable feedback for Arabic responses?

B. Structure of the Paper

The remainder of this paper is structured as follows: Section II reviews related work on the use of LLMs in education. Section III outlines the methodology adopted for utilizing LLMs in educational assessment, including best practices for prompt creation. Section IV presents the experimental results, evaluating which LLM performs best in grading and feedback generation. Section V discusses the findings, and finally, Section VI concludes the research.

II. Related Work

Table 1 explains the related work in utilizing the LLMs for automatic grading. Lee *et al.* [17] explored the use of GPT-3.5 and GPT-4 with CoT prompting to improve the automatic scoring of student-written responses in science assessments. The study employed a dataset of 1,650 student responses across six assessment tasks, using various prompt engineering strategies that combined zero-shot learning with CoT and scoring rubrics. The results showed that few-shot learning significantly outperformed zero-shot learning with CoT prompting combined with item stem and scoring rubrics, leading to notable improvements in scoring accuracy. GPT-4, particularly with the single-call greedy sampling strategy, demonstrated superior performance over GPT-3.5, highlighting the potential of LLMs to enhance both the accuracy and the transparency of automatic scoring systems in educational contexts.

Carpenter *et al.* [18] investigated the use of LLMs like GPT-3.5, GPT-4 [19], Llama 2 [20], and FLAN-T5 [21] to automatically assess student self-explanations in undergraduate computer science courses. The study leverages the EXPLAINIT classroom response system, collecting and evaluating 356 responses from 36 students during a pilot study. Different prompting strategies, including rubric-based and exemplar-based approaches, were tested to optimize model performance. The findings reveal that fine-tuned FLAN-T5 models achieved the highest accuracy (82.4%) and weighted F1 score (0.798) when using rubric information and exemplar responses, while GPT-4 with few-shot learning attained the highest macro F1 score (0.664) using ten labeled student responses. The results demonstrate the potential of LLMs, particularly FLAN-T5 and GPT-4, for effectively assessing student explanations in educational settings.

Lin *et al.* [11] investigated the use of LLMs, particularly BERT, with Named Entity Recognition (NER), to offer real-time explanatory feedback to human tutors. The study focuses on classifying tutor responses within an online lesson on giving effective praise, categorizing responses into effort-based and outcome-based praise. Utilizing a dataset of 129 annotated tutor responses, the BERT model demonstrated strong performance in identifying effort-based praise with an accuracy of 73.1% and an F1 score of 0.811, though it was less effective with outcome-based praise, achieving an F1 score of 0.350. The findings suggest that LLMs have the potential to provide automated, real-time feedback to tutors, but further enhancements, especially in data augmentation and handling various types of praise, are necessary to improve overall model performance.

Kortemeyer [22] investigated the application of GPT-4 for ASAG on two standard benchmark datasets, SciEntsBank and Beetle, which include student responses to general science and basic electronics questions. The study compares GPT-4's performance, both with and without reference answers, against specialized ASAG models using metrics like precision, recall, and F1 scores. The findings reveal that GPT-4's performance on the SciEntsBank dataset was higher, with an F1 score of 0.744 in the 2-way task, and overall, its performance was comparable to hand-engineered models but lagged behind specialized LLMs. Notably, GPT-4 performed better without reference answers in the Beetle dataset, achieving an F1 score of 0.651, indicating the model's potential as an out-of-the-box tool for ASAG tasks.

Schneider *et al.* [23] investigated the use of LLMs, specifically ChatGPT (GPT-3.5), to support the automatic grading of short textual answers. The study evaluated the effectiveness of LLMs in grading exam responses from two courses—one at the master's level in data science and another at the bachelor's level in information systems—by comparing the LLM's assessments with those of human educators. While LLMs provided a complementary perspective, their alignment with human grading was weak, with a tendency to give more generic and lenient evaluations, especially for students with poor language skills. The study highlighted that LLMs are not yet ready for independent auto-grading, as they are highly sensitive to minor changes in responses and require human oversight to ensure grading accuracy.

Mansour *et al.* [24] investigated the effectiveness of GPT-4 and GPT-3.5 in grading short-answer questions across Science and History subjects for K-12 students. Using a novel dataset of 1,710 student responses from the Carousel quizzing platform, the study compares the models' performance with human raters. The best-performing model, GPT-4 with few-shot prompting, achieved a Cohen's kappa score of 0.70, which is close to the human-level agreement of 0.75. The study highlights GPT-4's potential for use in low-stakes formative assessments, though it also notes the need for further fine-tuning and exploration of LLMs' capabilities in educational settings.

Xiao *et al.* [25] presented a dual-process framework for

Automated Essay Scoring (AES), integrating fast and slow thinking modules using LLMs like GPT-4 and Llama3 [26]. The study evaluates these models on both public (ASAP dataset) and private (CSEE dataset) essay datasets, with a focus on enhancing human-AI collaboration in grading. While LLMs did not surpass traditional AES methods in scoring accuracy, they excelled in generating high-quality explanations that improved the efficiency and performance of human graders. The framework, particularly with the fine-tuned Llama3-8B model, demonstrated the potential to enhance grading accuracy and offer robust feedback, making it a promising tool for educational applications.

In the study of Song *et al.* [27] addressing the challenges of Automated Essay Scoring (AES) and Automated Essay Revising (AER), researchers explored the use of open-source LLMs to enhance these tasks' efficiency and cost-effectiveness. The study utilized a dataset of 600 manually scored essays from 2870 Chinese primary school students, employing models like Baichuan 13B, InternLM-7B, and ChatGLM2-6B. These models were selected for their low cost, data security, and adaptability. The researchers applied zero-shot, few-shot, and continuous prompt tuning methods to assess the AES capabilities of the LLMs, comparing their performance to traditional statistical learning models and deep-learning baselines. Results indicated that while the zero-shot LLMs' performance lagged behind baseline models, few-shot and continuous prompt tuning methods significantly improved their effectiveness, with the best model achieving a Quadratic Weighted Kappa (QWK) of 0.531. For AER, qualitative and quantitative analyses demonstrated that the LLMs effectively enhanced essay quality while maintaining high similarity to original texts, thus proving useful for educational applications. However, the study noted the limitations of sample size and context length on model performance, and the challenges of achieving high consistency with manual scoring.

ASAG [28] aims to reduce teachers' workloads by using computational methods to evaluate student responses. This study evaluates ChatGPT models based on GPT-3.5 and GPT-4 for scoring short-answer questions in Finnish from ten undergraduate courses. GPT-4 achieved a QWK score of 0.6+ in 44% of one-shot settings, outperforming GPT-3.5 at 21%. While GPT-4 shows promise, further research is needed before it can be considered reliable. Models were instructed to follow a specific output format, but compliance varied. Despite format issues, all outputs contained valid predicted grades. GPT-3.5 occasionally added explanations to its grades, especially with binary grading scales. In the zero-shot setting, GPT-3.5 predicted 18 out of 200 answers as 'failed', while in the one-shot setting, it predicted 53 as 'failed'. Quantitatively, models tended to assign higher scores than human evaluators. One-shot GPT-4 performed best, achieving higher QWK scores than GPT-3.5. For binary grading scales, one-shot GPT-4 consistently achieved high accuracy. The TAA survival curve suggests that providing one example per grade allows educators to expect over 60% of scores to be within one point of their actual score 95% of the time.

Chamieh *et al.* [29] investigated the efficacy of LLMs, specifically GPT and Llama, for automated scoring of short-answer responses, focusing on zero-shot, few-shot, and fine-tuning settings. Models tested include GPT-3.5, GPT-4, Llama-7b, Llama-13b, Llama-70b, BERT, and SVM, using datasets such as ASAP, MindReading, and Powergrading. The evaluation was based on QWK scores. The results showed that LLMs perform poorly in zero-shot and few-shot settings, making them impractical for real-world applications without fine-tuning. Fine-tuning improved performance but was computationally expensive and sometimes led to overfitting. The best results for each approach were: for zero-shot, GPT-4 with 0.86 on Powergrading; for few-shot, GPT-4 with 0.87 on Powergrading; and for fine-tuning, GPT-3.5 with 0.83 on Powergrading. While GPT-4 showed promise on simpler datasets like Powergrading, it struggled with more complex tasks in ASAP and MindReading, indicating the limitations of current LLMs for automated scoring.

Xie *et al.* [30] introduced a multi-agent grading system called "Grade-Like-a-Human," which redefines the automated grading process by breaking it down into three stages: rubric generation, grading, and post-grading review. This approach involves refining grading rubrics based on student answers, applying optimized rubrics for grading, and conducting post-grading reviews to ensure accuracy and fairness. The system was tested using the newly collected OS dataset, derived from an undergraduate operating systems course, and the widely used Mohler dataset. The results showed significant improvements in grading accuracy, particularly for complex questions, with the best performance observed using batching and one-shot prompt strategies. The post-grading review process further enhanced the system's reliability by identifying and correcting anomalies in grading.

Aggarwal *et al.* [31] introduced the Engineering Short Answer Feedback (EngSAF) dataset and explore the use of LLMs to generate both grades and content-focused, elaborated feedback for student answers in the context of ASAG. The study curated a dataset of approximately 5.8k student responses across various engineering domains and utilized a Label-Aware Synthetic Feedback Generation (LASFG) strategy to enhance the educational value of automated grading. The system was tested and successfully deployed in a real-world end-semester exam at the Indian Institute of Technology Bombay (IITB), demonstrating high accuracy and quality in both grading and feedback, highlighting its potential for broader use in educational institutions.

Jiang and Bosch [32] explored the use of GPT-4 for ASAG by examining how different prompt configurations, such as including key elements of correct answers, scoring examples, and the order of tasks like score generation and rationale, impact the model's grading performance. Using the Automated Student Assessment Prize Short Answer Scoring (ASAP-SAS) dataset, which includes student responses to 10 questions across subjects like English, science, and biology, the study found that GPT-4 achieved an average QWK of 0.677. The results indicate that adding scoring examples generally improved performance, especially in science and biology, while the effectiveness of rationale generation varied depending on the evaluation metric, revealing important trade-offs in prompt design for ASAG tasks.

Morris *et al.* [33] presented the development of LLMs for automatically scoring constructed response items in mathematics. The approach involved extensive preprocessing,

including balancing class labels and tailoring input modifications for each item, followed by fine-tuning pre-trained models, particularly DeBERTa. Using a dataset from the National Assessment of Educational Progress (NAEP) 2023 Automatic Math Scoring Challenge, the study applied data augmentation and filtering techniques to address data imbalances. The results showed that the models achieved human-like agreement with human raters, with a less than 0.05 difference in QWK scores for nine out of ten items, demonstrating the potential of LLMs to efficiently and accurately score math assessments at scale.

After conducting a comprehensive search of published research, we found no existing studies utilizing DeepSeek models for Arabic ASAG at the time of writing. To our knowledge, this represents the first investigation of DeepSeek's application in ASAG tasks.

Table 1. Comparison of related work

Ref.	Task	Dataset	Model	Evaluation Metric	Language
[17]	Automatic scoring	Dataset of 1,650 student responses across six assessment tasks	GPT-3.5, GPT-4 with Chain-of-Thought (CoT) prompting	Accuracy	English
[18]	Assessing student self-explanations in undergraduate computer science	EXPLAINIT classroom response system, 356 responses from 36 students	GPT-3.5, GPT-4, Llama 2, FLAN-T5	FLAN-T5: 82.4% accuracy, weighted F1 score 0.798; GPT-4: macro F1 score 0.664	English
[11]	Providing explanatory feedback to human tutors	Dataset of 129 annotated tutor responses	BERT with Named Entity Recognition (NER)	Effort-based praise: accuracy 73.1%, F1 score 0.811; Outcome-based praise: F1 score 0.350	English
[22]	Automated Short Answer Grading	SciEntsBank and Beetle datasets	GPT-4	SciEntsBank: F1 score 0.744 in 2-way task; Beetle: F1 score 0.651	English
[23]	Automated Short Answer Grading	Responses from two courses: master's level data science, bachelor's level information systems	ChatGPT (GPT-3.5)	-	English
[24]	Automated Short Answer Grading	Carousel platform dataset, 1,710 student responses	GPT-4, GPT-3.5	GPT-4 with few-shot prompting: Cohen's kappa score 0.70	English
[25]	Essay Scoring	ASAP dataset, CSEE dataset	GPT-4, Llama3	Fine-tuned Llama3-8B model: QWK	English
[27]	Automated essay scoring and revising (AES, AER)	Dataset of 600 essays from 2870 Chinese primary school students	Baichuan 13B, InternLM-7B, ChatGLM2-6B	Best model QWK: 0.531 (few-shot, continuous prompt tuning)	Chinese
[28]	Automated Short Answer Grading	10 undergraduate courses (responses in Finnish)	GPT-3.5, GPT-4	GPT-4 QWK score >0.6 in 44% of one-shot	Finnish
[29]	Automated Short Answer Grading	ASAP, MindReading, Powergrading datasets	GPT-3.5, GPT-4, Llama-7b, Llama-13b, Llama-70b, BERT, SVM	Best QWK: GPT-4 in Powergrading (0.86 in zero-shot, 0.87 in few-shot)	English
[30]	Automated assessment with multi-agent system	OS dataset (undergraduate operating systems course), Mohler dataset	-	-	English
[31]	Automatic Short Answer Grading with feedback	Engineering Short Answer Feedback (EngSAF) dataset, 5.8k student responses	LLMs with Label-Aware Synthetic Feedback Generation (LASFG) strategy	High accuracy and quality in grading and feedback in real-world exam setting	English
[32]	Automatic Short Answer Grading	ASAP-SAS dataset (10 questions across subjects)	GPT-4	Average QWK: 0.677; scoring examples improved performance, varied effectiveness	English
[33]	Automatic Short Answer Grading	NAEP 2023 Automatic Math Scoring Challenge dataset	DeBERTa, data augmentation, fine-tuning	Human-like agreement with human raters; <0.05 difference in QWK scores for 9/10 items	English

III. Materials and Methods

A. Dataset

The Environmental Science Dataset (ESD) [34] is a dataset designed for automated grading of Arabic short answers. It comprises 61 questions from Egypt's Environmental Science curriculum, each with 10 student responses, totaling 610 answers. Two annotators graded responses on a 0 to 5 scale, providing an average score per answer. The dataset includes four question types: definitions, explanations, consequences, and reasons. Structured in three XML files with a detailed schema, it facilitates research and educational applications in Arabic language processing. High- and low-graded sample answers, translated into English, are provided for reference.

Table 2 shows a high-grade student's answer, while Table 3 shows a low-grade answer. Both samples are translated into English for non-Arabic readers.

Table 2. A high-grade sample of the Environmental Science Dataset (ESD)

ESD Sample	Question	Model Answer	Student Answer	Grade
A dataset original sample	بعض الحيوانات الصحراوية كاليرابيع لا تقرب الماء طيلة حياتها.	لأنه قد يفيد فى خفض نسبة النتح و البخر بإزالة أجزاء من المجموع الخضرى.	الرعى المنظم يساعد فى خفض نسبة النتح و البخر و إزالة أجزاء من المجموع الخضرى	5/5
Translation of the sample into English	Some desert animals, such as jerboas, do not go near water throughout their lives.	Because it may help reduce the rate of transpiration and evaporation by removing parts of the vegetative cover.	Organized grazing helps reduce the rate of transpiration and evaporation and removes parts of the vegetative cover.	5/5

Table 3. A low-grade sample of the Environmental Science Dataset (ESD)

ESD Sample	Question	Model Answer	Student Answer	Grade
A dataset original sample	يتعذر على الإنسان الغوص فى المياه العميقة بدون جهاز غطس.	لأن ضغط عمود الماء يتزايد بمعدل واحد ضغط جوى لكل عشرة أمتار تحت سطح الماء بالإضافة إلى الضغط الجوى على سطح البحر، فإذا أراد الإنسان أن يغوص إلى عمق 100 متر فإنه سيتحمل ضغط قدره 11 ضغط جوى و يتعذر ذلك بدون جهاز الغطس المخصص لذلك.	لأن جهاز الغطس يخفف من تأثير ضغط المياه فى الأعماق.	1/5
Translation of the sample into English	It is impossible for a person to dive into deep water without a diving device.	Because the pressure of the water column increases at a rate of one atmosphere for every ten meters below the surface, in addition to the atmospheric pressure at sea level. So, if a person wants to dive to a depth of 100 meters, they would have to withstand a pressure of 11 atmospheres, which is impossible without the specialized diving apparatus.	Because the diving apparatus mitigates the effect of water pressure at great depths.	1/5

B. Methodology

As shown in Fig. 1, this study employs a structured approach to evaluate the performance of LLMs in terms of the automatic grading task and feedback generation task for student responses in comparison to the model answer. The methodology is divided into two primary modules: the Grading Evaluation Module and the Feedback Evaluation Module, both of which are supported by a centralized Data Module.

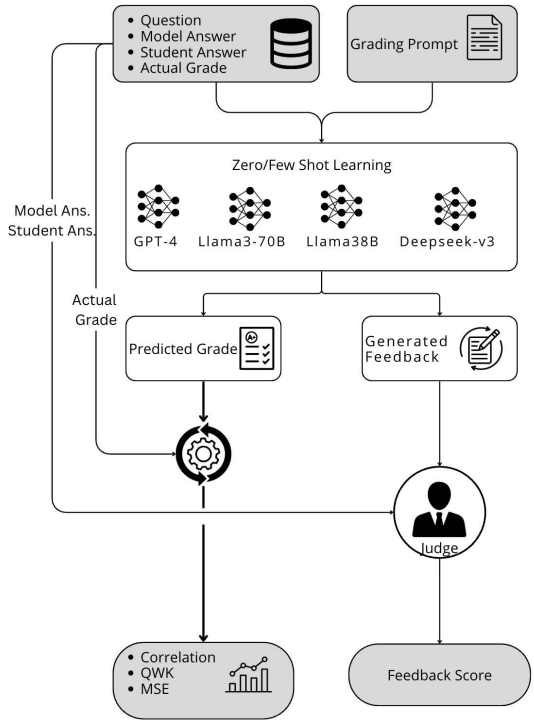

Fig. 1. Proposed methodology of utilizing LLMs in ASAG and feedback generation.

1) Data module

The Data Module is built upon a carefully curated dataset, the ESD, which encompasses 610 meticulously paired student answers alongside their corresponding model answers. This dataset plays a pivotal role as the underlying resource for both the grading and feedback generation processes. The ESD is specifically tailored to include responses related to subjects within the domains of physiology and earth sciences, ensuring that the content is relevant and challenging. By leveraging this dataset, the grading system is able to evaluate student performance against high-quality model answers, while the feedback mechanisms are designed to offer targeted insights that guide student learning in these scientific fields.

2) Prompt engineering

The Prompt Engineering Module serves as the foundational step in the evaluation and feedback processes, setting the stage for the effective use of LLMs in educational assessment. This module is meticulously designed to harness the full potential of LLMs by employing advanced prompt engineering techniques, ensuring that the models operate with precision, contextual awareness, and adaptability.

Prompt engineering begins with the careful selection of techniques such as zero-shot and few-shot learning, where the models are either provided with minimal or no prior examples to guide their responses. This approach is particularly powerful in enabling the LLMs to generalize from limited data, making it possible to assess a wide range of student answers with high accuracy.

At the heart of this module lies the construction of prompts that are not only tailored to specific educational tasks but also optimized for the strengths and capabilities of LLMs. The prompts are crafted in the language most relevant to the educational context, ensuring that instructions are clear and contextually appropriate. The effectiveness of prompt engineering in this context is crucial, as it directly influences the accuracy of the models in generating grades and feedback. Below is the list of the tactics used in the research.

Figs. 2 and 3 illustrate the final prompts used for few-shot and zero-shot learning, respectively, along with the strategies employed in their construction. The final prompts adopted in this study were developed through multiple iterative trials. Each iteration incorporated some or all of the six prompt engineering tactics outlined below. The optimal prompts were selected based on performance metrics, such as correlation with human scores, with higher-performing prompts retained and lower-performing ones discarded. The findings suggest that prompts incorporating all six engineering tactics yielded the best results. Each tactic addresses a specific aspect of the input query, collectively enhancing the relevance and completeness of the model's response. The selection of the best-performing prompt was guided by empirical evaluation, ensuring that the chosen prompts maximized alignment with desired outputs. The six prompt engineering tactics used in this process are detailed below.

1) Role Specification: By defining the AI's role as an "excellent short-answer grader," the prompt encourages the AI to adopt a persona of expertise. This tactic leverages the AI's ability to contextualize its output based

on the role it is assigned, leading to more accurate and relevant grading.

2) Contextual Examples: The inclusion of grading examples for each possible score (from 0 to 5) provides the AI with clear reference points. These examples ensure that the AI consistently applies the grading criteria, reducing the likelihood of bias or inconsistency.
3) Detailed Feedback Instruction: The AI is directed to generate feedback in Arabic, which ensures that the output is culturally and contextually appropriate. This tactic also enhances the relevance of the feedback, making it more useful for students.
4) JSON Output Requirement: Specifying JSON as the output format ensures that the AI's responses are structured and easy to process. This is particularly important for integrating the AI's output into larger systems or workflows that may require automated data handling.
5) Focus on Accuracy: The grading examples demonstrate a range of student answer qualities, helping the AI discern subtle differences and assign appropriate grades. This focus on detailed examples improves the accuracy and reliability of the grading process.
6) Chain of Thought (CoT): The prompt incorporates elements of Chain of Thought reasoning by guiding the AI through a logical, step-by-step evaluation process:
   - Comparison: The AI compares each student's answer with the model answer.
   - Evaluation: Based on the comparison, the AI determines the degree of correctness and assigns an appropriate grade.
   - Feedback Generation: The AI then generates feedback explaining the reasoning behind the assigned grade.

Fig. 2. The Arabic few-shot learning prompt used for scoring short answers, along with the tactics employed in its creation.

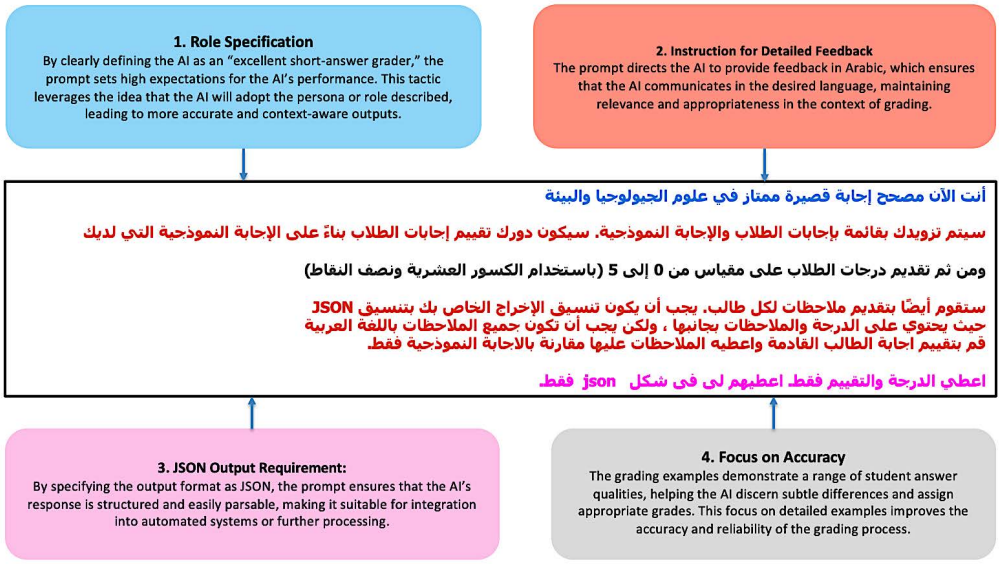

Fig. 3. The Arabic zero-shot learning prompt used for scoring short answers, along with the tactics employed in its creation.

3) Grading task module

The Grading Module is an integral component of the evaluation system, meticulously designed to leverage the capabilities of LLMs in the assessment of student responses. This module begins by constructing a few-shot prompt, as illustrated in Fig. 2, and a zero-shot prompt, as explained in Fig. 3.

The two prompts are carefully tailored to guide the LLMs in evaluating the quality and accuracy of student answers. These prompts are crafted based on a nuanced understanding of the subjects at hand, ensuring that the instructions provided to the models are both precise and contextually relevant.

The few-shot learning setting includes 11 examples, one for each possible grade available in the dataset (0, 0.5, 1, …, 5), enabling the model to learn how to map student responses to the appropriate grade more effectively. In contrast, zero-shot learning involves the model being given no examples.

In the initial phase, both the student's answer and the corresponding model answer from the ESD are input into these custom-designed prompts. This process involves a series of carefully selected techniques that enable the LLMs to interpret the answers in relation to the expected outcomes. The prompts are then fed into various LLMs, including state-of-the-art models such as GPT-4, Llama-3, with configurations of 70 billion (70B) and 8 billion (8B) parameters, and DeepSeek-V3. These models process the prompts, applying their extensive training on diverse datasets to generate two critical outputs: a predicted grade and detailed feedback.

The predicted grade represents the LLM's evaluation of how closely the student's response aligns with the model answer. This assessment is based on the content, structure, and relevance of the answer, as interpreted by the LLM. Once the predicted grade is generated, it is systematically compared with the ground truth grades provided in the ESD. To quantify the accuracy of the LLMs in replicating human grading, Pearson correlation metrics are employed. This statistical measure assesses the strength and direction of the linear relationship between the predicted grades and the actual grades, providing a robust evaluation of the LLMs' grading capabilities.

Simultaneously, the generated feedback, which is designed to offer constructive insights into the student's performance, is passed to the Feedback Module for further evaluation. This step is crucial, as it determines the effectiveness of the LLMs in providing meaningful, actionable feedback that can guide students in improving their future responses. The Feedback Module assesses the clarity, relevance, and utility of the generated feedback, ensuring that it meets the high standards required for educational contexts.

Through this comprehensive methodology, the Grading Module not only facilitates an in-depth evaluation of student responses but also provides valuable insights into the performance of different LLMs and the effectiveness of various prompting strategies. By comparing the predicted grades with the actual grades and evaluating the quality of the generated feedback, this module makes a significant contribution to advancing the use of LLMs in educational assessment, paving the way for more accurate, scalable, and efficient grading systems.

4) Feedback task module

The Feedback Evaluation Module plays a vital role in ensuring the accuracy and effectiveness of the feedback produced by the Grading Module. To preserve objectivity, the feedback is assessed by a human evaluator, as explained in Fig. 1, who serves as an independent and impartial judge. This dual-layered evaluation approach, encompassing both grading and feedback assessment, not only verifies the accuracy of the grades assigned by the LLM but also ensures that the feedback is clear, constructive, and aligned with educational goals.

IV. Experiments and Results

While conducting the experiments for GPT4 and DeepSeek_V3, we used their cloud API services. For Llama3_8B and Llama3_70B, we downloaded the models and conducted the experiment on a workstation with a Processor: Xeon W-3235 12 cores, 3.3 GHz, 128GB RAM, and an RTX 6000 24GB GPU.

A. Grading Task

Table 4 summarizes the experiments for the grading task. We used four well-known LLMs, three of which are open source, and one is closed source. The LLMs are GPT-4, Llama3-8b, Llama3-70b, and DeepSeek-V3. We employed two different prompts: the first using zero-shot learning and the second using few-shot learning, resulting in a total of eight model combinations.

To determine which prompt language is better, each prompt is tested once in Arabic and once in its translation into English. The Arabic prompt performed better, using the performance metrics, than the translated prompt, so we excluded the latter from our study.

We used three standard performance metrics in the problem domain, namely, the quadratic Weighted Kappa (QWK) [35], the Person Correlation, and the Root Mean Squared Error (RMSE).

Each of the three metrics is computed by directly comparing the gold grades, taken from the dataset, with the predicted grades for each student response.

DeepSeek-V3 (Few-shot) emerged as the most effective model for ASAG, achieving a QWK score of 0.8273, a Pearson correlation score of 86.09%, and an RMSE of 0.7602. Coming in second place, GPT4 (Few-shot) achieved a QWK score of 0.816, a Pearson correlation score of 86%, and an RMSE of 0.76. According to Table 5, DeepSeek-V3 (Few-shot) and GPT4 (Few-shot) achieved almost perfect agreement with the ground truth, while other models achieved an agreement ranging from moderate to substantial. The QWK score shows that LLMs have great potential in ASAG, especially Deep_Seek and GPT-4.

As shown in Table 4, the DeepSeek_V3, ChatGPT4, Llama3_8B, and Llama3_70 are compared with the baseline models of the Environmental Science Corpus results published in [34]. As explained in Table 4, DeepSeek_V3, GPT4, and Llama3_70B performed better than the baselines in correlation, and GPT4 and DeepSeek_V3 performed better than the baselines in RMSE. DeepSeek_V3 and GPT4 surpassed all baselines in correlation and RMSE. Results show that LLMs are promising in ASAG as they exceed the baselines. Table 4 does not include QWK because the QWK

baselines are reported in the original paper of the Environmental Science Corpus [34].

Table 4. Grading task experiment results

	Models	QWK	Correlation	RMSE
Zero-Shot Models	GPT4 (Zero Shot)	0.6938	84%	1.127
	Llama3 8B (Zero Shot)	0.4881	65%	1.8819
	Llama3 70B (Zero Shot)	0.5142	68%	2.0157
	DeepSeek-V3 (Zero Shot)	0.6101	79.49%	1.4164
Few-Shot Models	GPT4 (Few Shot)	0.816	86%	**0.76**
	Llama3 8B (Few Shot)	0.4674	57%	3.7119
	Llama3 70B (Few Shot)	0.7183	75%	1.3168
	DeepSeek-V3 (Few Shot)	**0.8273**	**86.09%**	**0.76**
Baseline Models	Char_Based_Bi_Cluster_11 [34]	-	72%	1.11
	Char_Based_Tri_ Cluster_11 [34]	-	71%	1.07
	Char_Based_Quad_ Cluster_11 [34]	-	70%	1.11
	Word_Based_Bi_Cluster_11 [34]	-	43%	1.16
	Word_Based_Tri_Cluster_1 1 [34]	-	39%	1.18
	Word_Based_Quad_Cluster 11 [34]	-	38&	1.20

Table 5. Interpretation of Kappa [35]

Kappa	Interpretation
<0	Less than chance agreement
0.01–0.20	Slight agreement
0.21–0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–1.00	Almost perfect agreement

B. Feedback Task

Table 6 shows the experiments conducted to assess the effectiveness of the underlined LLMs in generating feedback on students' answers. The feedback provided by the LLM includes explanations for the grades assigned to each response. To evaluate the quality of this feedback, a human expert is used to review it. Each piece of feedback was rated on a scale from 0 to 5. The expert followed Algorithm 1 to evaluate the generated feedback for each LLM.

The feedback assessment component is designed to evaluate the model's ability to generate detailed, constructive, and contextually relevant feedback based on grading outcomes. The primary focus is on the quality, clarity, and relevance of the feedback, particularly how effectively the model communicates insights and suggestions that help students understand their grades and improve their performance. This additional evaluation phase aims to rigorously test the model's text generation capabilities in a practical, educational context.

As shown in Table 6, the DeepSeek-V3 model in the few-shot setting outperformed all other LLMs in the feedback generation task. It achieved the highest feedback score of 79.61%, surpassing the performance of all other models across various prompt configurations.

Table 6. Feedback human scoring results

Model	Prompt Type	Feedback Score
GPT-4	Zero-shot	66.85%
GPT-4	Few-shot	67.77%
Llama-3 70b	Zero-shot	66.49%
Llama-3 70b	Few-shot	62.82%
DeepSeek-V3	Zero-shot	75.57%
DeepSeek-V3	**Few-shot**	**79.61%**

Algorithm 1. Scoring of the generated feedback for an LLM

```
FeebackScore(LLM):
Max_Score = 5
Min_Score=0
Scores=0

For each generated feedback F by LLM for a Question Q
Scores += HumanEvaluator(Student Answer, Model answer,
F, Max_Score, Min_Score)

LLM_Score = (average(Scores)/Max_Score)*100
Return LLM_Score
```

Furthermore, the experiments consistently demonstrated that few-shot learning led to improved performance. This trend suggests that providing example-based prompts helps guide the model, enhancing its ability to produce higher-quality responses.

Table 7 presents representative samples of the generated feedback along with the corresponding human evaluation scores. Each entry includes the model answer, the student response, the feedback generated by the model, and the score assigned by a human evaluator. To ensure accessibility for non-Arabic readers, each sample is accompanied by an English translation.

To evaluate the statistical significance of performance differences among the eight LLMs: GPT-4 Zero-Shot, GPT-4 Few-Shot, Llama 8B Zero-Shot, Llama 8B Few-Shot, Llama 70B Zero-Shot, Llama 7B Few-Shot, DeepSeek-V3 Zero-Shot, and DeepSeek-V3 Few-Shot, we conducted a paired Wilcoxon test. The results, presented in Table 8, show statistically significant differences between all model pairs ($p < 0.05$). These findings suggest that each model produces meaningfully distinct outputs, indicating a lack of redundancy among the evaluated LLMs.

Table 7. Samples of the generated feedback with their scoring using a human evaluator

LLM	Model Answer	Student Answer	Feedback	Human Score
Llama 3 70b-Few Shot	حيث تستخدم بعض الكائنات الحية مثل الطحالب الفضلات كغذاء لها. Some organisms, such as algae, use waste as food.	لأنه مثلا فى النظام البيئى البحرى نجد أن: الفضلات العضوية التى تخرجها الأسماك تتغذى عليها الطحالب التى تتغذى عليها الأسماك و هكذا لا تبقى هذه الفضلات فى ماء البحر فيظل الماء محتفظا بخواصه، غاز ثانى أكسيد الكربون الناتج من تنفس الكائنات البحرية تستخدمه النباتات البحرية فى عملية البناء الضوئى و ينتج غاز الأكسجين اللازم لعملية التنفس وهكذا تظل نسبة الغازين ثابتة فى الماء. For example, in the marine ecosystem, we find that: organic waste excreted by fish is fed to the algae that feed on the fish. Thus, this waste does not remain in the seawater, and the water retains its properties. Carbon dioxide produced by the respiration of marine organisms is used by marine plants in photosynthesis, producing oxygen necessary for	الإجابة جيدة لكنها لم تذكر دور الفضلات فى النظام البيئى البحرى بشكل كامل، كما أنها لم تذكر علاقة الفضلات بالغازين فى الماء. The answer is good, but it does not fully address the role of waste in the marine ecosystem, nor does it mention the relationship between waste and the two gases in the water.	3

		respiration. Thus, the ratio of these two gases in the water remains constant.		
GPT 4-Few Shot	اهى العلاقة بين فترة الإضاءة و الظلام التى يتعرض لها النبات خلال 24ساعة. It is the relationship between the period of light and darkness a plant experiences over a 24-hour period.	العلاقة بين فترة الإضاءة التى يحصل عليها النبات و فترة الإظلام التى يتعرض لها بعد ذلك بالتعاقب كل 24 ساعة It is the relationship between the period of light a plant receives and the period of darkness it experiences after that, alternating every 24 hours.	إجابة جيدة، ولكن يفتقد الجزء الذي يتحدث عن أن النبات يحصل على الإضاءة. Good answer, but missing the part about the plant receiving light.	4
DeepSeek – V3-Few Shot	لأن البيئات المائية متصلة بينما البيئات الأرضية منفصلة متباعدة و تختلف عليها الظروف الفيزيائية و الكيميائية و البيولوجية كالمناخ و التربة. Because aquatic environments are connected, while terrestrial environments are separate and distant, and their physical, chemical, and biological conditions, such as climate and soil, vary.	نظرا لاتصال مياه البحار و المحيطات ببعضها البعض لذلك فهى ثابتة نسبيا عن البيئات الأرضية التى تتفاوت فى ظروفها الفيزيائية و الكيميائية و البيولوجية نظرا لانفصالها فى شكل قارات و جزر متباعدة. Because the waters of seas and oceans are connected to each other, they are relatively stable compared to terrestrial environments, which vary in their physical, chemical, and biological conditions due to their separation into separate continents and islands.	إجابة الطالب ممتازة وتغطي معظم النقاط الرئيسية في الإجابة النموذجية، لكنها لم تذكر صراحة أن مياه البحار والمحيطات متصلة ببعضها البعض مما يجعلها ثابتة نسبيًا. The student's answer is excellent and covers most of the key points in the model answer, but it does not explicitly state that the waters of seas and oceans are connected to each other, making them relatively stable.	5

Table 8. Wilcoxon test between every pair of LLMs

Model		Zero Shot				Few Shot			
		GPT4	Llama3_8B	Llama3_70B	DeepSeek-V3	GPT4	Llama3_8B	Llama3_70B	DeepSeek-V3
Zero Shot	GPT4	—	0.0202	0.000	0.0024	0.000	0.000	0.000	0.000
	Llama3_8B	—	—	0.000	0.0002	0.000	0.000	0.000	0.000
	Llama3_70B	—	—	—	0.000	0.000	0.000	0.000	0.000
	DeepSeek-V3	—	—	—	—	0.000	0.000	0.000	0.000
Few Shot	GPT4	—	—	—	—	—	0.000	0.000	0.0048
	Llama3_8B	—	—	—	—	—	—	0.000	0.000
	Llama3_70B	—	—	—	—	—	—	—	0.000
	DeepSeek-V3	—	—	—	—	—	—	—	—

V. Discussion

The experiments highlight the strong performance of several well-known and widely cited LLMs (LLMs) in the task of ASAG. Among them, DeepSeek-V3 emerged as the top-performing model, achieving a Quadratic Weighted Kappa (QWK) score of 0.8273, a correlation of 86.09%, and a Root Mean Square Error (RMSE) of 0.76 on the designated dataset. These results demonstrate DeepSeek-V3's high accuracy in score prediction, with its QWK indicating an almost perfect agreement with human grading.

GPT-4 ranked second, demonstrating almost-perfect agreement with human scores, with a QWK of 0.816. Although it showed a slightly lower correlation, its RMSE matched that of DeepSeek-V3, further affirming its reliability. Llama-70B placed third, consistently outperforming Llama-8B across all evaluation metrics. These results highlight the strong potential of LLMs for Arabic educational assessment, particularly in the context of ASAG.

Given the superior performance of DeepSeek-V3, GPT-4, and Llama-70B, only these top models were included in the feedback evaluation phase to ensure meaningful comparisons and focus on the most effective LLMs.

According to Table 6, DeepSeekV3-FewShot excels in generating feedback for student responses, with an evaluation score of 79.61%. DeepSeekV3-ZeroShot achieved the second rank with a feedback score of 75.57%, and GPT4-FewShot achieved a 67.77% score. Again, DeepSeek-V3 proved to be the best among other models in grading and feedback generation.

A. Relationship between Training Data and Performance

Our experiments revealed significant performance differences among LLMs. DeepSeek-V3 outperforms GPT-4 and Llama 3 in both grading and feedback tasks. We hypothesize that this discrepancy may stem from variations in the models' training data, particularly their exposure to Arabic linguistic resources.

- **DeepSeek-V3**: Public documentation suggests this model was trained on a diverse multilingual corpus with substantial Arabic content, including educational texts. This aligns with its strong performance in handling morphological complexity and domain-specific terminology in our dataset.
- **GPT-4**: While robust in English, its lower performance than DeepSeek-V3 may reflect limited fine-tuning for Arabic educational contexts, as OpenAI has not disclosed detailed language-specific data ratios.
- **Llama 3**: The lower performance of Llama 3-8B/70B could indicate insufficient Arabic data or a lack of dialectal diversity in its pretraining corpus.

B. Classes of Errors

After analyzing the dataset, the errors in students' answers fall into the following classes. As future work, the types of errors can be included in the generated feedback.

1) **Missing Keywords or Terminology:** Important terms found in the model answer are absent.
2) **Logical Error or Misunderstanding:** The answer may be well-written linguistically, but it reflects a wrong or unrelated concept.
3) **Partial Answer:** Captures some correct elements but is incomplete.
4) **Irrelevant Answer:** The answer is completely off-topic or unrelated to the model answer.
5) **Language Clarity or Ambiguity:** The idea is roughly correct but poorly expressed, leading to unclear meaning.

C. Few-Shot Learning Effect

Few-shot learning proved to be the most effective approach

for grading, offering greater accuracy and adaptability. In this method, the model is provided with a few examples, allowing it to learn the specific patterns and associations between student answers and their corresponding marks.

Grading is a structured task where numerical associations based on defined criteria need to be established, and the few-shot learning approach enables the model to map these connections more accurately. The examples serve as a form of calibration, helping the model refine its understanding of the grading scale and deliver more precise predictions. Without these examples, the model might struggle to interpret the nuances of the grading process, leading to less reliable results.

All prompts were translated into English before being passed to the LLMs, rather than using the original Arabic prompts. However, our findings indicate that the Arabic prompts yielded better results, suggesting that prompts should be in the same language as the students' answers.

This study addresses a notable gap in the literature by exploring the use of LLMs in educational assessment for the Arabic language, a domain that has seen significant development in English but remains underexplored in Arabic due to limited linguistic resources and relatively less research compared to languages like English and Chinese. The work presented can be seen as an initial step toward establishing the foundation for leveraging LLMs in Arabic educational contexts. However, further research is needed to expand the scope of this work. Specifically, future studies should explore the application of LLMs to the evaluation of long-form answers, such as essays, in addition to short responses. There is also a need to validate the approach across a broader range of Arabic datasets to enhance generalizability. Moreover, evaluating newer and more advanced LLMs on the same educational assessment tasks is essential, given the rapid pace of development in this field.

VI. Conclusion and Future Work

In this study, we addressed the challenge of ASAG leveraging LLMs. Our research involved extensive experimentation with various open-source models to assess their effectiveness in both grading and generating feedback. We also investigated the impact of different learning paradigms, particularly zero-shot and few-shot learning, on model performance. Furthermore, we examined how the language used in prompts influenced the results. The findings demonstrate that LLMs perform exceptionally well in ASAG, consistently yielding strong results across experiments while also providing valuable feedback tailored to students' responses. Potential research directions include applying our findings to other datasets, comparing new LLMs with those used in this study, and exploring the fine-tuning of LLMs in other Arabic educational assessments, such as essay assessments, instead of short answers.

Ethics Statement

All data used in this study were obtained from actual student exams, with the explicit consent and authorization of the course instructor responsible for the assessments. At no point during the research was any Personally Identifiable Information (PII) collected or used. Specifically, no student names, photos, or other personal identifiers were accessed or stored. The study fully complies with ethical standards for academic research and upholds the privacy and anonymity of all participants.

We emphasize that our proposal of using LLMs in assessment must prioritize fairness across all students. For instance, the same prompt must be provided to every student, and if a specific LLM is used to evaluate one group, the same model should be used for all other groups. More broadly, any implementation of automatic evaluation must consider transparency and the ethical integrity of the educational process.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Methodology: Emad Nabil, Mostafa Mohamed Saeed, Rana Reda, Safiullah Faizullah, and Wael Gomaa; Software: Mostafa Mohamed Saeed, Rana Reda; Supervision: Emad Nabil, Safiullah Faizullah, and Wael Gomaa; Writing original draft: Mostafa Mohamed Saeed, Rana Reda; Writing review and editing: Emad Nabil, Mostafa Mohamed Saeed, Rana Reda, Safiullah Faizullah, and Wael Gomaa; Administration: Emad Nabil. All authors had approved the final version.

Funding

This work was funded by the Deanship of Scientific Research, Islamic University of Madinah, Saudi Arabia.

References

[1] J. Rudolph, S. Tan, and S. Tan, "ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?" *J. Appl. Learn. Teach.*, vol. 6, no. 1, Jan. 2023. doi: 10.37074/jalt.2023.6.1.9

[2] L. Yan *et al.*, "Practical and ethical challenges of large language models in education: A systematic scoping review," *Br. J. Educ. Technol.*, vol. 55, no. 1, pp. 90–112, Jan. 2024. doi: 10.1111/bjet.1337

[3] P. Boyd and S. Bloxham, *Developing Effective Assessment in Higher Education: A Practical Guide*, 2007.

[4] E. Kasneci *et al.*, "ChatGPT for good? On opportunities and challenges of large language models for education," *Learn. Individ. Differ.*, vol. 103, 102274, Jan. 2023. doi: 10.1016/j.lindif.2023.102274

[5] Q. Chen, X. Wang, and Q. Zhao, "Appearance discrimination in grading? Evidence from migrant schools in China," *Econ. Lett.*, vol. 181, pp. 116–119, Aug. 2019. doi: 10.1016/j.econlet.2019.04.024

[6] Z. Jiang *et al.*, "Exploring the role of artificial intelligence in facilitating assessment of writing performance in second language learning," *Languages*, vol. 8, no. 4, 247, Dec. 2023. doi: 10.3390/languages8040247

[7] R. Gao, N. Thomas, and A. Srinivasa, "Work in progress: Large language model based automatic grading study," in *Proc. Front. Educ. Conf. (FIE)*, Oct. 2023, pp. 1–4. doi: 10.1109/FIE58773.2023.10343006

[8] J. Han *et al.*, "FABRIC: Automated scoring and feedback generation for essays," arXiv preprint, arXiv: 2310.05191, Oct. 8, 2023

[9] A. Jonsson and G. Svingby, "The use of scoring rubrics: Reliability, validity and educational consequences," *Educ. Res. Rev.*, vol. 2, no. 2, pp. 130–144, Jan. 2007. doi: 10.1016/j.edurev.2007.05.002

[10] K. Taghipour and H. T. Ng, "A neural approach to automated essay scoring," in *Proc. EMNLP*, Nov. 2016, pp. 1882–1891, doi: 10.18653/v1/D16-119

[11] J. Lin *et al.*, "Using large language models to provide explanatory feedback to human tutors," arXiv preprint, arXiv: 2306.15498, Jun. 27, 2023.

[12] X. Wu *et al.*, "Matching exemplar as Next Sentence Prediction (MeNSP): Zero-shot prompt learning for automatic scoring in science education," in *Proc. Int. Conf. Adv. Data Mining Appl.*, 2023, vol. 1831, pp. 401–413. doi: 10.1007/978-3-031-36272-9_33

[13] J. Wei *et al*., "Chain-of-thought prompting elicits reasoning in large language models," in *Proc. the 36th International Conference on Neural Information Processing Systems*, 2022, pp. 24824–24837. doi: 10.5555/3600270.3602070

[14] Z. Liu *et al.*, "Context matters: A strategy to pre-train language model for science education," in *Proc. Int. Conf. Adv. Data Mining Appl.*, 2023, vol. 1831, pp. 666–674. doi: 10.1007/978-3-031-36336-8_103

[15] E. Latif and X. Zhai, "Fine-tuning ChatGPT for automatic scoring," *Computers and Education: Artificial Intelligence*, vol. 2024, no. 6, 100210, 2024.

[16] G.-G. Lee, "Gemini pro defeated by GPT-4V: Evidence from education," arXiv preprint, arXiv:2401.08660, 2023.

[17] G.-G. Lee *et al.*, "Applying large language models and chain-of-thought for automatic scoring," *Comput. Educ. Artif. Intell*., vol. 6, 100213, Jun. 2024. doi: 10.1016/j.caeai.2024.100213

[18] D. Carpenter *et al*., "Assessing student explanations with large language models using fine-tuning and few-shot learning," in *Proc. the 19th Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics*, 2024.

[19] OpenAI *et al.*, "Gpt-4 technical report," arXiv preprint, arXiv:2303.08774, 2023.

[20] H. Touvron *et al*., "Llama 2: Open foundation and fine-tuned chat models," arXiv preprint, arXiv: 2307.09288, Jul. 19, 2023.

[21] H. W. Chung *et al.*, "Scaling instruction-finetuned language models," arXiv preprint, arXiv: 2210.11416, Dec. 6, 2022.

[22] G. Kortemeyer, "Performance of the pre-trained large language model GPT-4 on automated short answer grading," arXiv preprint, arXiv: 2309.09338, Sep. 17, 2023.

[23] J. Schneider *et al*., "Towards LLM-based autograding for short textual answers," in *Proc. 16th Int. Conf. Comput. Supported Educ*., 2024, pp. 280–288. doi: 10.5220/0012552200003693

[24] W. Mansour *et al.*, "Can large language models automatically score proficiency of written essays?" arXiv preprint, arXiv: 2403.06149, Apr. 15, 2024.

[25] C. Xiao *et al*., "Human-AI collaborative essay scoring: A dual-process framework with LLMs," arXiv preprint, arXiv: 2401.06431, Jun. 14, 2024.

[26] A. Dubey *et al*., "The Llama 3 herd of models," arXiv preprint, arXiv: 2407.21783, Aug. 15, 2024.

[27] Y. Song *et al*., "Automated essay scoring and revising based on open-source large language models," *IEEE Trans. Learn. Technol.*, vol. 17, pp. 1920–1930, 2024. doi: 10.1109/TLT.2024.3396873

[28] L.-H. Chang and F. Ginter, "Automatic short answer grading for Finnish with ChatGPT," in *Proc. AAAI Conf. Artif. Intell*., vol. 38, no. 21, pp. 3173–3181, Mar. 2024, doi: 10.1609/aaai.v38i21.30363

[29] W. Chamieh *et al.*, "LLMs in short answer scoring: Limitations and promise of zero-shot and few-shot approaches," in *Proc. 19th Workshop Innovative Use NLP Educ. Appl. (BEA 2024)*, Jun. 2024, pp. 309–315.

[30] W. Xie *et al*., "Grade like a human: Rethinking automated assessment with large language models," arXiv preprint, arXiv: 2405.19694, May 30, 2024.

[31] D. Aggarwal, P. Bhattacharyya, and B. Raman, "'I understand why I got this grade': Automatic short answer grading with feedback," arXiv preprint, arXiv: 2407.12818, Jun. 30, 2024.

[32] L. Jiang and N. Bosch, "Short answer scoring with GPT-4," in *Proc. 11th ACM Conf. Learn. @ Scale*, Jul. 2024, pp. 438–442. doi: 10.1145/3657604.3664685

[33] W. Morris *et al*., "Automated scoring of constructed response items in math assessment using large language models," *Int. J. Artif. Intell. Educ.*, Jul. 2024, doi: 10.1007/s40593-024-00418-w

[34] W. H. Gomma and A. A. Fahmy, "Automatic scoring for answers to Arabic test questions," *Comput. Speech Lang.*, vol. 28, no. 4, pp. 833–857, Jul. 2014. doi: 10.1016/j.csl.2013.10.005

[35] A. Doewes, N. A. Kurdhi, and A. Saxena, "Evaluating quadratic weighted Kappa as the standard performance metric for automated essay scoring," in *Proc. 16th Int. Conf. Educational Data Mining (EDM)*, Bengaluru, India, Jul. 2023, pp. 103–113. https://doi.org/10.5281/zenodo.8115784

Copyright © 2025 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (CC BY 4.0).