JSCDM

JOURNAL OF SOFT COMPUTING AND DATA MINING
e-ISSN: 2716-621X

Vol. 7 No. 2 (2026) 58-74
https://publisher.uthm.edu.my/ojs/index.php/jscdm

Closed-Source vs. Open-Weights Language Models: Accuracy, Stability, and Deployment Trade-Offs in Educational Assessment (GPT-4o vs. LLaMA 3.1)

Handaru Jati[1]*, Yuniar Indrihapsari[1], Susilo Romadhon[1], Pradana Setialana[1], Danang Wijaya[2], Satya Adhiyaksa Ardy[3]

[1] *Department of Electronics and Informatics Engineering Education, Faculty of Engineering, Information Technology Studi Program, Universitas Negeri Yogyakarta, Karangmalang Number 1, Sleman, 55281, INDONESIA*
[2] *Department of Computer Science and Information Engineering, National Central University, Taoyuan, TAIWAN*
[3] *Department of Electronic and Computer Engineering, National Taiwan University of Science and Technology, Taipei, TAIWAN*

*Corresponding Author: handaru@uny.ac.id
DOI: https://doi.org/10.30880/jscdm.2026.07.02.005

Article Info

Received: 4 January 2026
Accepted: 20 April 2026
Available online: 30 June 2026

Keywords

educational technology, GPT-4o, large language models (LLMs), LLaMA 3.1, model comparison, multiple choice questions, short answer evaluation

Abstract

Large language models (LLMs) are increasingly used to support educational assessment, yet institutional adoption requires balancing performance with governance constraints related to auditability, data control, and deployability. We compare a closed-source model (GPT-4o) and an open-weights model (LLaMA 3.1) on two expert validated assessment formats across five academic domains (Science, History, Literature, Technology, and Social Studies). Experiment 1 evaluates multiple choice question (MCQ) option selection using 148 items, five prompting strategies (Raw, Brief Instruction, Long Instruction, Chain-of-Thought, and Question-Answer Prompt Generation), and five independent repetitions per condition (7,400 total model calls) to estimate both accuracy and run-to-run stability under a strict A to E output constraint. Experiment 2 evaluates 147 one sentence short answer items under a fixed instruction prompt. In the multiple choice questions (MCQs), the accuracy and stability of the results for GPT-4o were higher compared to those of LLaMA 3.1, and the incorporation of the additional prompt structure was associated with higher stability compared to the average improvement in correctness. In the short answer questions, the performance of GPT-4o was slightly higher compared to that of LLaMA 3.1 for ROUGE-L and METEOR, and these results mostly corresponded to the results obtained for the semantic similarity questions and the evaluation rubric. The results of the study are important for understanding the accuracy-stability trade-off for the two language generators and for supporting the evaluation approach that considers the correctness of the results for MCQs, their stability for multiple runs, and the validation of the results for short answer questions using the evaluation rubric and the results for the semantic similarity questions.

1. Introduction

Academic work in higher education is starting to change due to large language models (LLMs). Chatbot-based systems are already used by teachers and students for practice, formative feedback, and study assistance [1], [2]. Recent studies also describe current LLMs as foundation models, that is, general purpose systems whose capabilities can transfer across tasks and domains [3],[4]. This may be one reason why these systems have been adopted so quickly in assessment contexts. As these tools become more common in routine campus use, an important practical question emerges: which types of models are suitable for assessment contexts where accountability and oversight cannot be compromised [1],[5].

However, in multiple-choice tests, there is another issue that is not simply a matter of calculating average accuracy. In these cases, it is equally important that given inputs produce the same output when the material is used repeatedly. Multiple-choice questions (MCQs) continue to be popular in many tests because it is possible to validate answers against a given key, comparing different student cohorts [6]. However, in practical usage, it is not usually the case that MCQ creation is done in one go. Instead, it is common to reuse the same questions, albeit with some changes in wording, in different contexts. Given that it is a matter of generating text, there is a possibility that even when the question is unchanged, different answer choices could be given due to small changes in wording, as well as the inherent randomness of language generation. If different outputs are given in these cases, it might be necessary to validate, leading to a lack of confidence in model usage, as has already been suggested in other contexts [5], [6].

Short answer questions represent a different kind of challenge, but the problem is similar to a certain extent. Many tests rely on one-sentence or short constructed responses due to their ability to measure certain aspects of understanding that cannot be captured via multiple-choice questions alone [1]. However, these responses defy easy classification and automatic verification. Although the metrics have some inherent value, the single reference evaluation methodologies have inherent and widely reported shortcomings. A response can be perfectly faithful in terms of interpretation while at the same time scoring poorly if the wording of the response significantly deviates from the reference. In this way, the right interpretation is penalized if lexical overlap is limited [5]. Current evaluation methodologies suggest the use of a combination of lexical overlap metrics, semantic similarity metrics, and rubric-based evaluation when the interest is in evaluating the conceptual alignment of the responses and not the similarity of the responses in terms of wording [5].

Recent survey and review studies have identified the potential benefits and challenges associated with the incorporation of large language models in the classroom. However, there is a significant and critical empirical issue that remains unaddressed. This is because institutions need data that can inform practice and not simply contribute to the discourse. Specifically, the existing literature has yet to examine the relationship between the accuracy of multiple-choice questions and the stability of the output from run-to-run in the context of constrained and machine-checkable output. Such comparisons are especially limited when closed-source and open-weight models are evaluated under the same controlled experimental protocol [1], [5], [6]. A similar limitation is evident in the evaluation of short answer responses. Many studies continue to rely primarily on lexical similarity metrics, despite the well-recognized sensitivity of single reference evaluation to paraphrasing and variation in wording [5]. Taken together, these limitations make it difficult for institutions to interpret published findings as practical guidance for deployment rather than as isolated benchmarking results.

This study bridges this gap by comparing GPT-4o to LLaMA 3.1 in five academic domains: Science, History, Literature, Technology, and Social Studies. The study employs a comparison of the models based on two expert-validated item banks containing 148 multiple-choice questions and 147 one-sentence short-answer questions. The study has five prompting techniques, and each of the experiments was conducted five times to obtain an accurate and stable result. The study focused on a single MCQ option in Experiment 1. The models have to provide exactly one answer letter from A to E. The study also employed Experiment 2, where one-sentence short answers were used to evaluate the models. The study used a fixed instructional prompt for the short answers. The study used lexical overlap metrics to obtain reproducible results. The study was conducted to evaluate the models based on BERT Score on the entire dataset. The study also used rubric-based evaluations conducted by two human raters on a stratified subset of the results. The study was conducted to find out if the results showed a difference based on word similarity or based on a difference in meaning.

The current study is informed by three specific research questions. RQ1 investigates whether there are discernible differences in the accuracy of multiple-choice questions (MCQs) for GPT-4o and LLaMA 3.1 using various prompting strategies and academic domains. RQ2 involves an examination of the consistency of the output generated for both models after multiple runs under the same conditions, including an evaluation of whether the form of the prompt influences such consistency. RQ3 involves an evaluation of the quality of short answer questions that are composed of only one sentence, including whether lexical overlap indicators are consistent with both semantic similarity and human evaluation based on the rubric. The rest of the paper involves an overview of the relevant literature, the study design, the results from both studies, and the implications and limitations of the study, including suggestions for future research.

J. of Soft Computing and Data Mining Vol. 7 No. 2 (2026) p. 58-74

2. Related Work

2.1 LLM in Higher Education Assessment and Learning Workflows

Large Language Models (LLMs) are no longer restricted to experimental or pilot implementations in a classroom setting. For instance, in a higher education context, LLMs are increasingly integrated into day-to-day scholarly activity, including tutoring, generation of feedback drafts, as well as supporting various types of assessment activity [1], [2]. The existing body of review studies suggests that these technologies could help in reducing time taken to provide feedback as well as in supporting the scaling of practice activity. However, these studies also point to a variety of salient issues, including the possibility of overreliance on these technologies, variability in model effectiveness across subject domains, as well as difficulties in accountability in contexts where institutions heavily rely on services whose inner workings are not transparent [1]. Because many educational applications involve assessment resources such as questions, model answers, scoring rubrics, and evaluation rationales evaluation cannot focus solely on overall performance metrics. It must also consider whether model behavior remains consistent, transparent, and manageable when these systems are deployed under realistic institutional constraints [1], [7].

Recent research on assessment has moved beyond one time measures of correctness and has instead examined how LLMs perform in scoring related workflows, where model outputs may influence real instructional decisions [8]. Work on MCQ authoring raises a related point. Model responses are more likely to remain usable when the task is clearly specified, and the expected output is tightly constrained. This is particularly important in settings where instructors need consistency with the answer key and a response in the form of one machine checkable option letter [9].

2.2 Closed-Source vs Open-Weights LLMs: Governance Trade-off

Recent comparative studies have revealed that the adoption of a large language model in institutional settings is not just limited to the selection of the model with the highest benchmark performance. Governance factors are also likely to impact the decision-making processes [7]. Closed-source models are often available through API-based services, while the weights and the overall development processes of these models are not disclosed. In the absence of these factors, the overall auditing of the models may become more challenging and may not be easily replicable or verifiable, despite the high performance of the models [7], [10]. These are the reasons why the focus is given to models that can be locally deployed, thereby providing the institution with more flexibility and the ability to review, replicate, and verify the overall evaluation processes.

Large language models are no longer limited to the pilot or experimental phases in instructional settings. In the case of higher education, these models are increasingly being integrated into the overall scholarly processes, including the provision of tutoring, the creation of feedback drafts, and the overall support of various assessment processes [1], [2]. The overall review of the literature has revealed that these models have the potential for providing faster feedback and scaling the overall practice processes. At the same time, the literature has identified several salient factors associated with the overall adoption of these models in institutional settings, including the possibility of over-reliance on the models, the overall performance of the models in various disciplines, and the overall accountability of the models in the context of institutional settings [1]. Since many educational processes are likely to include the overall assessment of various resources, including questions, model answers, scoring rubrics, and evaluation rationales, the overall evaluation of these models cannot be limited to the overall performance of the models. Rather, the overall evaluation must focus on the overall assessment of the consistency of the models in institutional settings [1], [7].

2.3 Reliability Under Repeated Use in MCQ Centered Workflows

Multiple choice questions are commonly used, and their results can be objectively measured and compared among different groups [6]. However, their practicality brings with it a critical aspect that has often been missed: the need to ensure that the results obtained are identical when the questions are used for different purposes, like drafting or checking the questions or their distractors. This has been seen to not always hold true, especially given that the process of AI generation is stochastic and that small changes to the prompt can produce significant effects. This has made teachers verify the results manually to align them with the ones produced by the AI tool, which has made the process unreliable [6], [7]. This is because the accuracy of the AI tool in grading depends on the absolute ability to select distinct and correct options without any form of error.

Previous studies on the evaluation process shed more light on the reasons behind the discrepancies that may occur when reported scores or rankings differ across different testing processes. From the empirical evidence provided, it is evident that benchmark results vary because of data contamination and the effects of specific protocol modifications [5], [11]. These discrepancies create more problems when examined under closed-source systems because the architecture of the systems is not clearly understood. Fortunately, with the HELM and MT

Bench standardized reporting process, more transparency has been achieved by providing strict guidelines on accuracy and robustness, which helps readers understand discrepancies that may occur when interpreting data from different sources [12], [13]. However, most benchmarking processes that exist today mainly focus on providing coverage of tasks with performance results obtained from a single run process. This has made them inadequate in providing evidence on the stability of systems during repeated processes under the strict conditions that exist when used to solve problems like grading multiple choice questions [5], [6].

2.4 Constructed Response/Short Answer Assessment and Scoring Evidence

From In the normal classroom situation, educators are able to assess significantly more than just the results of a series of multiple choice questions. In fact, educators can assess the results of written responses that include concise explanations as well as detailed articulations of reasoning. As such, a wide range of courses have continued to include a variety of constructed response items that include short answer items. Unlike the results of a traditional multiple choice question, these items offer a unique window into a student's level of knowledge as well as their capacity to articulate concepts. While the body of research on the assessment of the results of a constructed response has a long tradition of scholarship, recent studies are increasingly considering the potential that modern language models can offer to this field of assessment, while the final decision rests with the instructor [14], [15]. From this perspective, the assessment of a short answer question offers a valuable complement to objective assessment that closely mirrors the written assessment that is normally used.

In the evaluation of the short answer questions, a unique set of methodological issues arises. Unlike the multiple choice questions, the answers do not possess a binary label for verification. Though the application of metrics proves useful to a certain extent, the evaluation of the results becomes a challenge while relying on a single reference answer. A student may articulate the correct concept with precision, yet the evaluation of the answer proves to be low due to the various phrasings of the answer. Extensive research on the evaluation of the language models has consistently demonstrated that the metrics applied for evaluation need to be congruent with the evaluation goals [5]. Therefore, the recent research suggests the application of a hybrid evaluation model that incorporates the conventional application of lexical analysis with the evaluation of semantic similarity [14], [16].

2.5 Synthesis and Gap Linking to RQ1–RQ3

Existing research literature suggests that large language models can be highly instrumental in facilitating the process of assessments in the realm of higher learning institutions. At the same time, the literature also points to some critical aspects of governance that become important when institutions opt to use closedsource models [1], [7]. First and foremost, the critical aspects pertain to data control, auditability of the system, and reproducibility of output. It is important to mention here that despite the increase in research literature on the subject, there is a significant scarcity of literature that is focused on the accuracy of multiple choice questions with severe output constraints and repeated iterations in a single framework with stringent testing. This is especially important in the context of assessments in an academic setting since teachers often require the same questions to be repeated at different stages of the academic process [5], [6].

Another gap exists with regards to the assessment of short answer questions. Various studies have demonstrated the use of lexical metrics due to the ease of replicability; however, short answer questions with single reference sources are easily subject to paraphrasing. Therefore, lexical similarity could be considered an attribute of educational quality without semantic validation and/or rubric based assessment [5], [14].

These differences are in line with the research questions posed in this study. Specifically, Research Question 1 (RQ1) aims to examine differences in multiple choice question (MCQ) accuracy among models using different prompting strategies and academic domains. Research Question 2 (RQ2) aims to explore differences in repeated runs under similar conditions and determine if prompt structure affects this consistency. Research Question 3 (RQ3) aims to extend this comparison to one sentence short answer questions by examining answer quality and determining if lexical overlap signal corresponds to semantic similarity and human rubric judgments [5], [6], [14].

3. Method

3.1 Overview

Two complementary experiments were conducted to compare GPT-4o and LLaMA 3.1 on educational assessment tasks with different output formats and evaluation criteria. Experiment 1 focused on multiple-choice option selection, mirroring a common need in many evaluation processes. Experiment 2 focused on one-sentence short-answer question answering, mirroring a common need in many evaluation processes. The academic domains used in these experiments were identical, with five being used in total: Science, History, Literature, Technology, and Social Studies. The outcomes of these experiments were analyzed at the item level, comparing pairs of items. This was done as the inputs to both models were identical. The study was conducted as a study of controlled

comparison with deployment-oriented relevance. The two items were used to draw inferences at the item-paired level under fixed decoding. The study was limited to five academic domains, as well as two datasets, in order to ensure reproducibility of the study. Even though this increases the internal validity of the study, it limits its external validity.

3.2 Models and Settings

This study examined two models, GPT-4o and LLaMA 3.1. The former is a closed source system queried through the OpenAI API, while the latter is an open weights model queried through the Hugging Face Inference API. To keep inference comparable across systems, the decoding settings were held constant at temperature = 0.5, top p = 1.0, and max tokens = 200. Each run was issued as a separate request with the same inputs and parameters. Every response was logged verbatim for automated scoring and subsequent analyses.

3.3 Datasets

To accommodate the two assessment modes, two datasets were created. For Experiment 1, 148 expert-reviewed multiple-choice questions were created in five domains: Science, History, Literature, Technology, and Social Studies. For Experiment 2, 147 short answer question-answer pairs were created in the same five domains. For both datasets, expert reviewers were consulted to fine-tune the questions and remove unnecessary ambiguity. Model outputs were compared against expert-defined targets, which were defined as the keyed answer option in the case of the MCQs and the reference answer in the case of the short answer questions. The chosen dataset sizes were deemed adequate to perform an item-paired assessment while being manageable in terms of repeated runs measurement. For Experiment 1, each of the MCQs was replicated five times across five prompting strategies and two models, resulting in 7,400 runs. This is deemed adequate to ensure stable estimation of both accuracy and stability across runs under identical circumstances. At the same time, the datasets used in the experiment were representative of the five domains rather than exhaustive. Thus, it is not appropriate to generalize the findings to other curricula, grade levels, or writing styles.

3.4 Experiment 1: MCQ Option Selection

3.4.1 Study Design

A controlled quantitative comparison was conducted between a closed source model (GPT-4o) and an open weights model (LLaMA 3.1) on educational multiple choice questions (MCQs). MCQ option selection was treated as the target task because assessment workflows require a discrete and machine checkable response rather than open ended generation. Both models were evaluated under five prompting strategies, and each condition was repeated five times. This made it possible to estimate two main outcomes, namely MCQ accuracy and stability across repeated runs.

3.4.2 Data Collection

A total of 148 multiple choice questions were compiled across five academic domains: Science, History, Literature, Technology, and Social Studies. Each item was reviewed by domain experts to ensure clear wording, plausible distractors, and one unambiguous correct answer. In every trial, both models were given the same question stem together with the same five options from A to E. Outputs were scored by checking whether the selected letter matched the expert answer key.

3.4.3 Prompting Strategies (Operational Definitions)

Five prompting strategies were evaluated using the constant question stem and A-E options. For all prompting strategies, the response format was constrained to a single machine-checkable option letter. For the Raw condition, the item is used as is. Brief Instruction (BI) adds a brief instruction that usually consists of one or two sentences asking the model to select the best option and return the letter. Long Instruction (LI) uses an even longer instruction that explicitly guides the model to read the question carefully, follow the single-letter constraint, and make a simple decision such as selecting the best-supported option without providing an explanation. For the Chain of Thought (CoT) prompting strategy, the instruction explicitly asks the model to reason before making a response. However, the final response is still evaluated based on the last letter in the range of A to E. QAPG uses an even more structured prompt that asks the model to perform intermediate checks such as restating the question, identifying the key concept, before making the final option selection. This is again evaluated based on the last letter.

3.4.4 Primary Outcomes: Accuracy and Stability

The two main outcomes in this experiment were MCQ accuracy and how stable the results remained across repeated runs. Accuracy was scored as 1 when the model selected the expert keyed option and as 0 otherwise. To quantify stability, each condition was repeated five times. For each item, the modal answer share was calculated as the proportion of runs in which the most frequently selected option was repeated. For item i, this quantity was defined as:

$$\boldsymbol{S_i} = \frac{max(n_A, n_B, n_C, n_D, n_E)}{R}, \boldsymbol{R = 5} \quad (1)$$

where n_A, n_B, n_C, n_D, and $n_E$represent the counts of options A to E across the Rrepeated runs for item i. With five repetitions, $S_i$ranges from 0.20 to 1.00, and higher values indicate greater consistency in option selection across repeated runs. Stability was reported as the mean ± SD of $S_i$across the 148 items for each model and prompting strategy.

3.4.5 Exploratory Budget and Implementation Details

Across all conditions, the evaluation involved 7,400 model runs, calculated from 148 items, 5 prompting strategies, 5 independent runs, and 2 models. Outputs were scored automatically using a strict post processing rule in which only the final option letter from A to E was retained. When a response contained additional text, the last valid letter was extracted and the remaining text was disregarded.

The generation parameters were kept constant across the two models, with temperature set to 0.5, top p to 1.0, and max tokens to 200. GPT-4o was queried through the OpenAI API, whereas LLaMA 3.1 was queried through the Hugging Face Inference API. Each run was issued as an independent request with identical inputs and parameters so that the results could be reproduced and aggregated consistently across runs.

3.4.6 Statistical Analysis

All analyses were conducted at the item level because each MCQ produced paired outcomes from GPT-4o and LLaMA 3.1 under identical inputs. Nonparametric tests were used because accuracy is binary and stability is bounded.

- Cross model comparisons (within each prompting strategy): Two sided paired Wilcoxon signed rank tests were applied to item level outcomes. For accuracy, each condition produced 148 paired binary scores. For stability, each condition produced 148 paired stability values.
- Prompting strategy effects (within each model): The Friedman test was used to evaluate item level outcomes across the five prompting strategies. When the Friedman test indicated an overall effect, we used Nemenyi post hoc comparisons and applied Holm correction within each model.
- Domain level variation (within each model): Kruskal-Wallis tests were used across the five domains. When the Kruskal-Wallis test indicated an overall effect, we conducted planned Mann-Whitney U follow ups and applied the Holm correction within each model.

The significance threshold was set at α = 0.05. We report Holm-adjusted p-values for families of follow up comparisons and report unadjusted p-values only for tests that did not belong to a multiple comparison family.

3.5 Experiment 2: One Sentence Short Answer Question Answering

3.5.1 Study Design

The second experiment dealt with short answer question answering, with responses restricted to a single sentence. The protocol was designed to resemble classroom short response assessment, where students are expected to answer briefly while still including the main factual content. The same five academic domains from Experiment 1 were kept so that the two tasks could still be compared. The design was also item paired, which meant that both models answered the same prompts under the same decoding configuration.

3.5.2 Data Collection

The short answer materials comprised 147 question and answer pairs taken from the same five domains. Each item consisted of an Indonesian question and a one sentence reference answer prepared by domain experts. The reference answer was written to capture the minimum information needed for a correct response. Subject matter experts also reviewed the full item set in order to reduce ambiguity and to confirm that the reference answers reflected the intended target content rather than a narrowly specific wording.

J. of Soft Computing and Data Mining Vol. 7 No. 2 (2026) p. 58-74

3.5.3 Prompting Strategies

The prompt was kept the same for all items. Each question was followed by the instruction, "Answer briefly in one sentence." This helped keep response length consistent and reduced variation caused by changes in prompt wording. For each item, one response was collected from each model because the purpose of this experiment was to examine average short answer quality rather than stability across repeated runs.

3.5.4 Primary Outcomes: BLEU, ROUGE-L, and METEOR

Short answer quality was assessed at the item level by means of three automatic metrics, all of which were computed with respect to a reference answer for each item: BLEU, ROUGE-L, and METEOR. The three metrics vary in how they calculate n-gram overlaps, with BLEU including a brevity penalty. ROUGE-L, in turn, measures the length of the longest subsequence shared between a response and a reference. METEOR is more flexible in its use of stemming and synonyms. The same preprocess was applied to both models, as were the same settings for the metrics. The availability of only one reference answer for each item in this dataset is another limitation. The response may be semantically correct but receive low scores if it is formulated differently from the reference. Lexical similarity metrics, therefore, were seen as replicable evidence of compliance with the reference answer, not as a measure of conceptual correctness.

3.5.5 Semantic and Rubric Validation

Lexical metrics were complemented by two additional validation layers. First, BERT Score F1 was computed for the full set of 147 items using the same scorer model and the same preprocessing steps for both systems. Paired differences between the two models were then tested with a two sided Wilcoxon signed rank test. Second, a stratified subset of 100 items was selected for rubric based human evaluation, with approximate balance across domains. Two raters scored each model response using a three level rubric, where Incorrect = 0, Partial = 1, and Correct = 2. When disagreement occurred, the final label was established through discussion. Inter rater agreement was calculated on the ratings before adjudication using Cohen's κ and quadratic weighted κ.

3.5.6 Implementation Details

The same decoding parameters as Experiment 1 (temperature = 0.5, top-p = 1.0, max_tokens = 200). Each generated answer was stored verbatim before standardized preprocessing was applied for metric computation.

All inferential analyses were conducted at the item level. GPT-4o and LLaMA 3.1 were compared on BLEU, ROUGE L, and METEOR using both a paired t test and a Wilcoxon signed rank test. A significance level of $\alpha = 0.05$ was used. Since these metrics had been identified in advance as the primary outcomes, the planned comparisons are reported with unadjusted p values. Effect sizes are also reported to aid practical interpretation.

3.6 Methodological Limitations and Ethical Considerations

The present study was designed to allow a controlled comparison under fixed decoding settings, and this makes the results easier to reproduce. Even so, several limitations need to be kept in mind when the findings are interpreted. The item banks cover five domains only and represent a selected set of materials rather than the full range of curricula, difficulty levels, or item writing styles that may appear in practice. The comparison also reflects one particular access route and one particular point in time for each system. If a provider updates the model or changes its safety tuning, endpoint configuration, or quantization, both accuracy and stability may shift. In the short answer task, the use of a single reference answer helps keep the scoring procedure consistent. At the same time, it also makes the evaluation more sensitive to paraphrasing.

Assessment use raises questions that go beyond performance alone. For that reason, these experiments are better read as evidence for low stakes use under human supervision than as a basis for fully automated scoring. Model outputs are more appropriately treated as decision support than as final judgments. Institutions also need to protect student data through suitable privacy safeguards and to check performance against local curricula before any operational use takes place, especially when the results may affect grades or student progression.

4. Results

4.1 Experiment 1: MCQ Option Selection

4.1.1 MCQ Accuracy Across Prompting Strategies

The mean MCQ accuracy of GPT-4o and LLaMA 3.1 across the five prompting strategies is shown in Table 1.

Table 1 *Mean MCQ accuracy of GPT-4o and LLaMA 3.1 across prompting strategies (N = 148; R = 5)*

Prompt Style	GPT-4 accuracy (%)	LlaMa 3.1 accuracy (%)	p-value (Wilcoxon)
Raw	86.5	78.4	0.018
Brief Instruction	88.5	79.7	0.009
Long Instruction	87.2	79.1	0.023
Chain-of-Thought	89.2	78.4	0.002
QAPG	83.1	71.6	0.003

Note: Two sided paired Wilcoxon signed rank tests were conducted to compare GPT-4o and LLaMA 3.1 within each prompting strategy. Holm adjusted p values for the five prompt specific comparisons are reported in the table, with α set at 0.05. Accuracy was computed over 148 items using five independent runs for each item under each prompting condition.

GPT-4o outperformed LLaMA 3.1 under every prompting strategy shown in Table 1. Its accuracy ranged from 83.1% in QAPG to 89.2% in CoT, whereas LLaMA 3.1 ranged from 71.6% in QAPG to 79.7% in BI. Two sided paired Wilcoxon signed rank tests confirmed crossmodel differences within each prompting strategy, and all five comparisons remained significant after Holm adjustment (α = 0.05).

4.1.2 Run-to-Run Stability Across Prompting Strategies

Table 2 summarizes run-to-run stability as modal answer share across five repeated runs for each condition.

Table 2 *Run-to-run stability across prompting strategies (modal answer share; R=5; N=148)*

Prompt Style	GPT-4o stability (mean ± SD)	LLaMA 3.1 stability (mean ± SD)	p-value (Wilcoxon)
Raw	0.69 ± 0.06	0.63 ± 0.08	0.028*
Brief Instruction	0.76 ± 0.05	0.69 ± 0.06	0.014*
Long Instruction	0.80 ± 0.04	0.73 ± 0.05	0.006**
Chain-of-Thought	0.85 ± 0.03	0.78 ± 0.04	0.002**
QAPG	0.88 ± 0.02	0.82 ± 0.03	0.001**

Note: Stability denotes modal answer share across R = 5 runs per item (range 0.20–1.00); the table reports two sided paired Wilcoxon tests for GPT-4o vs. LLaMA 3.1 with Holm-adjusted p-values across the five prompt specific comparisons (α = 0.05).

GPT-4o showed higher stability than LLaMA 3.1 under every prompting strategy (Table 2). Paired Wilcoxon signed rank tests detected stability differences within each prompt condition, and Holm correction across the five prompt specific stability comparisons preserved statistical significance for all contrasts (α = 0.05). Both models showed higher stability under more structured prompts (Table 2).

4.1.3 Prompting Strategy Effects within Each Model

We tested whether the prompting strategy altered accuracy within each model using a Friedman test across the five prompting strategies on item level outcomes (N = 148). The test showed no significant strategy effect for GPT-4o, $X^2(4) = 2.095$, p = 0.72, or LLaMA 3.1, $X^2(4) = 2.10$, p = 0.715. Mean ranks were tightly clustered across strategies, and Nemenyi post hoc comparisons did not identify significant pairwise differences within either model at α = .05.

4.1.4 Accuracy by Domain

Figure 1 visualizes accuracy differences across domains, and Table 3 provides domain by prompt accuracy for both models.

J. of Soft Computing and Data Mining Vol. 7 No. 2 (2026) p. 58-74

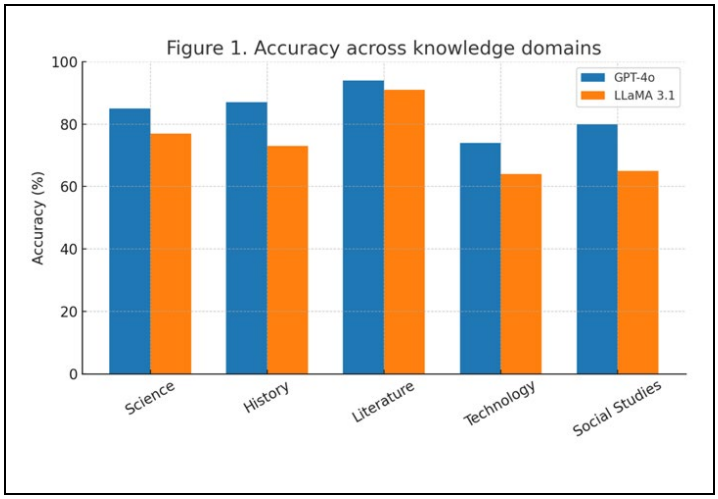

Fig. 1 *Accuracy across knowledge domains*

Table 3 *Domain wise MCQ accuracy by prompting strategy for GPT-4o and LLaMA 3.1 (N = 148 across five domains)*

Model	Model	Raw	BI	LI	CoT	QAPG
Science	GPT-4o	85.2	87.6	85.2	86.6	83.9
	LLaMA 3.1	71.6	74.1	75.3	74.1	67.9
History	GPT-4o	86.0	88.3	87.5	88.3	82.5
	LLaMA 3.1	72.5	76.1	74.2	73.8	68.2
Literature	GPT-4o	94.4	94.4	94.4	94.4	88.9
	LLaMA 3.1	91.3	91.3	87.0	82.6	78.2
Technology	GPT-4o	73.3	80.0	80.0	80.0	73.3
	LLaMA 3.1	72.7	72.7	63.6	72.7	63.6
Social Studies	GPT-4o	74.0	79.0	79.0	79.0	73.3
	LLaMA 3.1	69.3	70.1	65.1	68.0	61.5

Note: Values are percent correct at the item level; domain sample sizes were Science (n = 30), History (n = 30), Literature (n = 30), Technology (n = 30), and Social Studies (n = 28) (total N = 148).

Both models achieved their strongest performance in Literature and their weakest performance in Technology and Social Studies (Table 3). GPT-4o maintained higher accuracy than LLaMA 3.1 across all domains and prompting strategies (Table 3).

4.1.5 Domain Level Significance Tests

Table 4 reports domain level significance tests for within-model domain contrasts.

Table 4 *Domain level pairwise significance tests within each model (Mann-Whitney U, p-values)*

Model	Comparison	p-value
GPT-4o	Literature vs Technology	0.0007
	Literature vs Social Studies	0.003
	Science vs Social Studies	0.011
LLaMA 3.1	Literature vs Technology	0.009
	Science vs Social Studies	0.045

Note: We tested the predefined within model domain contrasts using MannWhitney U tests and report Holm-adjusted p-values within each model across the follow up comparisons shown (α = 0.05).

For GPT-4o, Literature significantly outperformed Technology and Social Studies, and Science significantly outperformed Social Studies (Table 4). For LLaMA 3.1, Literature significantly outperformed Technology, and Science significantly outperformed Social Studies (Table 4).

4.1.6 Option Distribution Alignment Analysis

We assessed whether each model's option selection tendencies track the reference option distribution across prompting strategies by computing Spearman correlations between each model's option distribution and the reference distribution. We computed each model distribution by taking the modal option per item across five runs and aggregating modal letters across items within each prompting condition. Table 5 reports alignment scores by prompting strategy.

Table 5 *Option distribution alignment across prompting strategies (Spearman's ρ).*

Model	GPT-4o ρ	LLaMA 3.1 ρ
Raw	0.98	0.90
Brief Instruction	0.92	0.90
Long Instruction	0.91	0.87
Chain-of-Thought	0.89	0.36
QAPG	0.90	0.30

Note: For each prompting strategy, we computed an option selection distribution as the proportion of items assigned to options A to E (a 5 element vector) and computed Spearman's ρ between each model's vector and the answer key option distribution. Because the correlation uses only five categories, we interpret ρ as a descriptive alignment statistic.

GPT-4o showed high alignment across prompting strategies, whereas LLaMA 3.1 showed lower alignment under CoT and QAPG (Table 5). Because the correlation uses only five option categories, we interpret ρ as a descriptive alignment statistic rather than a fine grained fit measure.

4.2 Experiment 2: Short Answer Question Answering

4.2.1 Domain Wise Short Answer Scores

Table 6 reports domain wise short answer performance for GPT-4o and LLaMA 3.1 using BLEU, ROUGE-L, and METEOR. All metrics are computed at the item level against a single reference answer and then averaged within each domain.

Table 6 *Domain wise automatic evaluation for short answers (N = 147)*

Model	Model	Bleu	Rouge-L	METEOR
Science	GPT-4o	0.010255	0.111638	0.203773
	LLaMA 3.1	0.009313	0.103301	0.213592
History	GPT-4o	0.005444	0.084746	0.164942
	LLaMA 3.1	0.005551	0.068702	0.133292
Literature	GPT-4o	0.006301	0.078839	0.139026
	LLaMA 3.1	0.002990	0.064567	0.100848
Technology	GPT-4o	0.004660	0.078315	0.135754
	LLaMA 3.1	0.005877	0.079827	0.132660
Social Studies	GPT-4o	0.003830	0.069955	0.138423
	LLaMA 3.1	0.004131	0.073961	0.130024

Note: Mean scores were calculated from item level metrics using one reference answer for each item (N=147). Higher values indicate a closer match to the reference.The domain sample sizes were 30 for Science, 29 for History, 30 for Literature, 30 for Technology, and 28 for Social Studies, giving a total of 147 items.

Across domains, BLEU scores were low for both models. This is expected in a one sentence, single reference setting, where a correct paraphrase may share only limited n gram overlap with the reference answer. Larger differences between the models appeared in ROUGE L and METEOR for several domains. In History and Literature, GPT-4o scored higher than LLaMA 3.1 on both ROUGE L and METEOR (Table 6). In Science, GPT-4o scored higher on BLEU and ROUGE L, whereas LLaMA 3.1 was slightly higher on METEOR (Table 6). For Technology and Social

Studies, the gap between the two models was smaller, and the results did not follow a single pattern across metrics (Table 6).

4.2.2 Global Score

Table 7 summarizes global mean scores across all 147 items.

Table 7 *Global short answer scores across all items (N = 147)*

Model	BLEU	ROUGE-L	METEOR
GPT-4o	0.006133	0.084899	0.156570
LLaMA 3.1	0.005592	0.078191	0.142308

Note: Global means across all 147 items.

GPT-4o scores higher than LLaMA 3.1 on overall ROUGE-L and METEOR, whereas the gap in BLEU stays relatively narrow. A similar pattern can be seen across the domain level findings. Because only one reference answer was available, n-gram overlap did not separate the models very clearly, while sequence based comparison and synonym aware matching showed the difference more steadily.

4.2.3 Paired Tests and Effect Sizes

To examine whether the two models differed on short answer metrics for the same items, paired comparisons were carried out at the item level (N=147). The paired t test and Wilcoxon signed-rank results for BLEU, ROUGE L, and METEOR are summarized in Table 8.

Table 8 *Paired statistical tests for short-answer metrics (N = 147)*

Metric	Test	Statistic	p-value	Result
BLEU	Paired t-test	t=-0.6708	0.503436	Not Significant
BLEU	Wilcoxon signed-rank	W=5006.0000	0.402411	Not Significant
ROUGE-L	Paired t-test	t=-2.4947	0.013720	Significant
ROUGE-L	Wilcoxon signed-rank	W=3974.0000	0.004612	Significant
METEOR	Paired t-test	t=-1.9917	0.048265	Significant
METEOR	Wilcoxon signed-rank	W=4212.0000	0.017657	Significant

Note: All tests were carried out at the item level (N=147), using identical items for both models. For the Wilcoxon signed-rank test, W denotes the sum of signed ranks.

As Table 8 shows, BLEU does not separate GPT-4o and LLaMA 3.1 under either paired test (both $p > .05$). In contrast, ROUGE-L and METEOR show statistically reliable paired differences under both test families ($p < .05$), indicating that GPT-4o achieves higher reference adherence scores than LLaMA 3.1 on these two metrics when we compare responses item by item.

Statistical significance alone does not convey practical magnitude, so we report effect sizes for the metrics with significant paired differences. Table 9 reports Cohen's d (paired) and Wilcoxon effect size r for ROUGE-L and METEOR, and Table 10 consolidates the key inferential results, effect sizes, and direction of the differences.

Table 9 *Effect sizes for significant short-answer metrics (N = 147)*

Metric	Cohen's d (paired)	r (Wilcoxon)
ROUGE-L	0.206	0.269
METEOR	0.164	0.226

Note: We computed Cohen's d (paired) on itemlevel paired differences, $d = {\Delta}/{SD(\Delta)}$, where $\Delta = \text{score}_{\text{GPT4o}} - \text{score}_{\text{LLaMa3.1}}$. The Wilcoxon effect size $r$was calculated from the standardized Wilcoxon statistic using the normal approximation, $r = {Z}/{\sqrt{N}}$, for a two sided test. Positive values of r indicate highe r scores for GPT-4o.

Table 10 *Consolidated statistical comparison and effect sizes (N = 147)*

Metric	Test	Statistic	p-value	Effect Size Type	Value	Direction
ROUGE-L	Wilcoxon	W=3974	0.004612	r (Wilcoxon)	0.269	GPT-4o>LLaMA 3.1
ROUGE-L	Paired t-test	t (146)=-2.4947	0.013720	d (paired)	0.206	GPT-4o>LLaMA 3.1
METEOR	Wilcoxon	W=4212	0.017657	r (Wilcoxon)	0.226	GPT-4o>LLaMA 3.1
METEOR	Paired t-test	t (146)=-1.9917	0.048265	d (paired)	0.164	GPT-4o>LLaMA 3.1
BLEU	Wilcoxon	W=5006	0.402411	-	-	not reported (n.s.)
BLEU	Paired t-test	t (146)=-0.6708	0.503436	-	-	not reported (n.s.)

Note: The table reports W for the Wilcoxon signed rank test and t(df) for the paired t-test. Effect sizes are reported only for metrics that showed statistically significant paired differences.

Tables 9 and 10 show that the effect sizes are generally small to moderate, with the largest separation appearing in ROUGE L. This indicates that GPT-4o holds a fairly consistent advantage across items, although the average size of the difference is still modest. Such a pattern is in line with the expected sensitivity limits of single reference short answer scoring.

4.2.4 Semantic and Rubric Validation

Because lexical overlap metrics may understate answers that preserve the intended meaning through paraphrase in single reference settings, an additional validation step focused on meaning was included. Table 11 reports item level BERT Score F1 as mean ± SD for all short answer items (N=147), together with the paired Wilcoxon signed rank test for the difference between the two models.

Table 11 *BERTScore F1 across short answer items (mean ± SD; N = 147)*

Model	BertScore F1 (mean)	SD
GPT-4o	0.866	0.08
LLaMA 3.1	0.843	0.08

Note: BERTScore F1 was computed at the item level (N=147) using the same scorer model and the same preprocessing steps for both systems. The cross model difference was evaluated with a two sided paired Wilcoxon signed rank test (p=0.047).

As Table 11 shows, GPT-4o obtained a higher mean BERTScore F1 than LLaMA 3.1, and the paired test indicates that the difference is statistically reliable at α=0.05. This result provides convergent, meaning oriented support for the ROUGE-L and METEOR gaps observed in Table 8.

We then checked whether the metric differences align with human judgments on a stratified subset. Table 12 summarizes rubric outcomes on the 100 item subset (Correct/Partial/Incorrect and mean score), and Table 13 reports inter rater agreement on the pre adjudication ratings using Cohen's κ (unweighted) and quadratic weighted κ.

Table 12 *Rubric results on 100 item subset (Correct/Partial/Incorrect+mean score)*

Model	Correct	Partial	Incorrect	Mean score (0-2)
GPT-4o	66%	18%	16%	1.50
LLaMA 3.1	58%	20%	22%	1.36

Note: Percentages sum to 100% within each model. Two raters independently scored each response using a 3 level rubric (Incorrect=0, Partial=1, Correct=2). When raters disagreed, they resolved the label by discussion to produce an adjudicated final score, and we computed the mean score from these final labels. The subset was selected using stratified sampling across domains (≈20 items per domain) to preserve topic coverage.

J. of Soft Computing and Data Mining Vol. 7 No. 2 (2026) p. 58-74

Table 13 *Inter rater agreement for 3 level rubric scores (0 to 2) on the 100 item subset*

Metric	Value
Cohen's kappa (unweighted)	0.58
Weighted kappa (quadratic)	0.76

Note: Agreement is computed on the pre adjudication independent ratings for N = 200 responses (100 items × 2 models) using the ordinal rubric (0/1/2). Quadratic weights are used for weighted κ to reflect the ordered distance between adjacent rubric levels.

As Tables 12–13 show, GPT-4o attains a higher mean rubric score and a higher share of fully correct responses than LLaMA 3.1, while the agreement statistics indicate moderate to strong reliability between raters. Overall, these results suggest that the observed advantages in the automatic metrics are not explained by surface overlap alone, but are likely to reflect differences in meaning.

5. Discussion

5.1 Experiment 1: Interpretation and Practical Implications

This research attempts to fill a gap in the practical application of large language models (LLMs) for education. Many studies have carried out comparative analyses that focus on the accuracy of a single shot or the average performance of the benchmark. However, from a practical point of view, the application of an LLM in an education setting often requires the repeated usage of the same item for tasks such as quiz creation, formative task verification, or the modification of assessment materials. Therefore, it is not sufficient for an LLM to give the correct answer for a single shot or a single application. Instead, the same item must give the same answer to the machine every time it is used [1], [6] [1], [6].

For this reason, the present study used a deployment oriented protocol that combines MCQ correctness with stability across repeated runs under a strict A to E output constraint. The same protocol was applied across prompting strategies and academic domains in order to inform institutional choices between closed source performance and open weights control [7], [10] [7], [10].

Across prompting strategies, GPT-4o achieved higher MCQ accuracy than LLaMA 3.1 (Table 1). This pattern appeared under all five prompt styles. The result suggests that, in this dataset, model choice influenced correctness more strongly than prompt style. It also indicates that prompt design has limits in workflows that require a single option letter to be parsed reliably. A change in prompt may make the format of the response more uniform and may reduce variation, but the difference between the two models in keyed option selection does not disappear.

Another pattern was related to stability, which was identified as an important operational factor. In all prompting conditions, GPT-4o showed greater stability than LLaMA 3.1 in repeated runs (Table 2). In other words, the closed-source model showed more consistency in selecting the same option when the same item was repeated. This is important in an instructional context because if teachers are using the same items repeatedly, the stability of the output reduces the need to resolve conflicts between those items and decreases the likelihood that automated scoring or screening will produce different results for the same items.

In this dataset, prompt structure appeared to matter more for stability than for correctness. As prompts became more structured, both models produced more repeatable option choices (Table 2). By contrast, changes in accuracy across prompt styles were small when compared with the overall gap between the two models (Table 1). This finding has a practical implication for deployment. Prompt templates do not appear to function as a reliable way to increase accuracy, but they do provide a workable way to reduce variance. Assessment teams may therefore use prompt standardization to limit drift, especially when an open weights model is deployed and predictable downstream behaviour is required.

The domain level results also indicate where deployment risk may be more concentrated. Both models performed best in Literature and least well in Technology and Social Studies (Table 3). The contrasts within each model suggest that these differences are more likely to reflect domain demand than a problem caused by one particular prompt style (Table 4). This means that institutions should not assume uniform performance across subject areas. Instead, performance needs to be checked against local curricula, item review may need to be tightened in weaker domains, and item wording and difficulty should be aligned carefully with the intended learning objectives before model outputs are used in routine MCQ workflows.

A related pattern appears in the alignment of option distributions with the answer key. GPT-4o remained closely aligned across prompts, whereas LLaMA 3.1 shifted more when the prompts became more structured (Table 5). In small scale use, such variation may seem limited. Across large item banks or repeated use across courses, however, this kind of drift may lead to systematic changes in which options are selected more often, even when outputs remain restricted to A to E. From a deployment perspective, this supports the use of prompt

standardization and disciplined decoding settings as basic governance controls, particularly when open weights models are used.

5.2 Experiment 2: Short Answer Question Answering Metric Alignment, Findings, and Implications

A note is needed on the relation between the task and the metrics. BLEU, ROUGE L, and METEOR do not directly measure conceptual understanding, instructional usefulness, or rubric level correctness. In this study, they were used as reproducible indicators of reference adherence in a controlled setting. Each item had one expert reference answer, and each model produced one sentence. This design supports transparent comparison at the item level, but it can assign too little credit to valid paraphrases and too much credit to surface overlap when factual coverage is incomplete [5] [5]. For this reason, lexical metrics were treated as screening indicators and were interpreted together with a semantic similarity check, namely BERTScore on the full set, and rubric based human ratings on a stratified subset (Tables 11–13).

Within that scope, the short answer results show a modest but fairly consistent advantage for GPT-4o on metrics that are more sensitive to sequence alignment and synonym tolerant matching. At the overall level, GPT-4o scored higher than LLaMA 3.1 on ROUGE L and METEOR (Table 7), and the paired tests indicated that both differences were statistically reliable (Table 8). BLEU did not separate the two models (Table 8). This is in line with the limited sensitivity of n gram precision in short answers with a single reference, where correct responses may differ in wording while preserving the intended meaning.

The supplementary validation layers provide clarity on the interpretation of the aforementioned. BERTScore is indeed higher for GPT-4o for the entire dataset (Table 11). For the rubric based dataset, the mean score and proportion of completely correct answers were indeed higher for GPT-4o, although inter rater agreement remained in the moderate to strong range (Tables 12–13). This undercuts the persuasiveness of the surface level account. The pattern of results suggests that the differences in ROUGE-L and METEOR might indeed be more plausibly explained in relation to the content rather than the wording. Nevertheless, the result must be cautiously interpreted. While automatic metrics can play a part in the evaluation process, human judgment is necessary in educational quality.

With respect to instruction, the second experiment reaches a tentative conclusion. Both models were able to demonstrate their ability to produce single sentence responses that are useful in practice; however, GPT4o was more aligned with references in both automatic and human evaluation contexts. In low stakes formative assessments, these measures may still be useful in facilitating quick review and initial screening. In more critical situations, rubric based evaluation should continue to be used. It may also be necessary to design evaluation procedures that are less dependent upon a single reference answer. It may be useful to incorporate more references or other types of keys or structured response types to help insure that paraphrasings are more equitably evaluated.

With respect to Research Question 3 (RQ3), the second experiment reveals that the lexical measures, especially ROUGE-L and METEOR, together with the semantic and human validation measures, tend towards a general overarching pattern. At the same time, the variability between models is limited. As such, it is still feasible to compare the quality of short answer responses, although the reliability of such an approach is improved if interpreted with the help of secondary evidence related to meaning.

5.3 Limitation and Future Work

There are also practical issues that deserve closer attention in future work. Institutions may respond to the present findings in several ways. Prompt standardization, tighter decoding constraints, retrieval augmentation for domain grounding, and local adaptation under governance controls are some possible examples. What remains unclear is how much each of these steps would actually reduce the performance and reliability differences observed in this study. This is a critical question for future research. Deployment guidance would be more useful in practice if evidence was provided on this point.

The findings need to be interpreted with caution. Part of the basis for the caution stems from the scope of the materials used for the research. Two expert validated item banks were used for the analysis. These banks contained 148 multiple choice questions and 147 short answer questions across five academic domains. Though this allows for paired comparisons of the items, it does not necessarily translate to other curriculum types or levels. Therefore, an expanded set of items would be useful for future research. This would also enable the exploration of the ability to cluster the items by level of difficulty and cognitive demands. Additionally, the ability to incorporate other subject areas would be useful to see if the findings regarding accuracy and reliability continue to hold.

A second limitation concerns the systems themselves. The comparison provides a snapshot of each model at a given point in time based on a given access path to the model. GPT-4o is accessed through the Open AI API, while LLaMA 3.1 is accessed through an inference endpoint. These conditions are not static over time. The accuracy and

stability of the model may change over time based on changes to a model provider's accuracy or stability when they make changes to a model or to the serving configuration, quantization, or preprocessing of the model. Replication is necessary to verify the findings. The same procedure should be repeated with future releases of the model, and future studies should report clearly the version identifiers used to access the model.

Similarly, the constraints of the MCQ protocol apply to the interpretation of results. For instance, the participants were only allowed to choose an option letter from A to E. This is because the protocol is deterministic in terms of assessment. While it is in line with the machine-checkable assessment paradigm, there is still more to explore in terms of understanding the behavior of the models. For example, it is important to consider whether the rationale behind the chosen option is sound, what type of uncertainty is present, and what the instructional value of the result is. As such, it would be interesting to consider extending the protocol to include more than just the selection of options. This would include evaluating the rationale using a rubric or evaluating the confidence statement. This would provide valuable insights into the balance between simplicity in operation and instructional value.

There is an analogous limitation in the short answer protocol. For each question, the answer is evaluated against a single expert answer. This ensures that there is consistency in the assessment. However, it is apparent that the range of acceptable paraphrasing is restricted. As such, it is conceivable that correct answers would be penalized if the wording is divergent. As such, it would be interesting to consider extending the protocol in the future. This would include using multiple answer keys or even a checklist.

There are various methods that can be used to ensure reliability. For instance, in the present study, the modal answer share was used. This is because it is relevant in scenarios involving repeated use. It is also useful in illustrating the propensity of the model to produce the same answer repeatedly. However, it is apparent that it does not illustrate all changes. As such, it would be interesting to consider extending the protocol in the future. This would include incorporating item entropy, agreement coefficients, or even calibration plots that illustrate confidence against correctness. This would be particularly interesting if it were used in an open-weight system. There are a number of additional practical questions that future research should address more concretely. There are the possibilities for mitigation strategies, for example, including the application of rapid standardization, tighter decoding constraints, retrieval augmentation for domain grounding, or local adaptation with governance controls. It would be interesting to see how much each of these influences the performance and reliability gaps.

6. Conclusion

Two forms of assessment were investigated in this current study. One of these forms involved multiple choice option selection, where it was necessary for the response to be machine evaluable. In another form, one sentence short answer questions were employed, making this comparison more similar to brief written work, as is commonly seen in a classroom context. GPT-4o acted as the closed-source model, and LLaMA 3.1 acted as the open-weights model. For both tasks, an item paired approach within five different domains was employed, enabling not only a comparison in terms of correctness but also in terms of reliability when using reused material.

The outcome of the multiple choice question (MCQ) test revealed clear results, indicating GPT-4o's superiority in terms of precision over LLaMA 3.1 using any of the prompting strategies applied. GPT-4o also showed more consistency in its tests, a factor that carries considerable significance in day to day teaching processes. It is not uncommon for instructors to revisit questions during the creation of tests, quizzes, screening questions, and even when revising distracters. If a model gives different answers to similar questions even after a second evaluation, further verification is needed, thus undermining the overall utility of such a system. Structured prompts were seen to play a positive role in enhancing stability in both systems. However, they were not successful in eliminating the overall difference in precision between these two systems when subjected to a stringent response system using A to E. It is clear, therefore, that structured prompts were more successful in enhancing consistency than in altering precision outcomes.

In this case, in the short answer task, it was observed that the performance gap between the two models remained small but distinguishable. GPT-4o received a higher ROUGE-L and METEOR score, while BLEU failed to distinguish between the two in terms of a single reference. Such an observation is in line with expectation, as it is known that these automatic evaluation systems are not a true measure of actual understanding and may even penalize accurate factual responses simply because of a difference in articulation. As such, these systems were not relied upon in isolation; they were simply a preliminary measure of reference similarity. To gain more meaningful insights, various evaluation techniques were also employed by the researchers. For instance, both BERTScore evaluations on the full data set and human evaluation on a designated subset also confirmed the earlier findings; overall, it may be concluded that the performance gap is indeed a true measure of semantic correctness and not simply a measure of superficial similarity.

These findings directly answer the three central research questions posed by the current study. In terms of the first research question, GPT-4o was found to consistently perform better than LLaMA 3.1 in terms of accuracy in multiple choice questions regardless of prompting strategy and academic domains considered. In terms of the

second research question, the findings suggested that prompt engineering mainly improved response stability as opposed to improving accuracy under stringent automated evaluation. In terms of the third research question, the findings highlighted the limitations of solely depending on lexical measures to evaluate response quality. In particular, while semantic measures seemed to point to GPT-4o as the preferred model, BLEU scores were generally inconclusive. However, additional semantic evaluation measures strongly corroborated the findings. Overall, the entire framework of evaluating the performance of GPT-4o against its competitor shows a generally moderate but persistent performance advantage of GPT-4o over LLaMA 3.1.

Several practical implications arise from the above findings. Firstly, for any institution or system that seeks to incorporate large language models (LLMs) into multiple choice question (MCQ) systems, repeatability and correctness should be a concern, especially for cases where the content is reused. Secondly, for any system or developer working with open weight models, it would be advisable to view prompt standardization and decoding disciplines as a fundamental form of system governance since the structured prompts were seen to reduce variance even if they did not necessarily boost accuracy. Thirdly, for the evaluation of short answer questions, it would be unwise to rely entirely on lexical matching for correctness.

The results need to be interpreted with the limitations of the research. The analysis employed two expert validated item banks, consisting of 148 multiple choice questions and 147 short answer questions, from five academic domains. Additionally, the models were evaluated through a particular access route at a particular time. Though this allows for a comparative evaluation, it alone is not sufficient to support generalization. Future research could involve increasing the item set, utilizing different curriculum levels of difficulty, or repeating the entire procedure with subsequent versions of the models and serving configurations. This research could help to shed light on the phenomenon of temporal drift and whether the general pattern repeats.

Acknowledgement

The authors express their sincere appreciation for the assistance provided by Universitas Negeri Yogyakarta, National Central University, and National Taiwan University of Science and Technology throughout the course of this research. The collaboration among these institutions, together with the academic resources and encouragement received, contributed substantially to the completion of the study.

OpenAI's ChatGPT and Grammarly were used only for language editing, including grammar, readability, and style. They were not used to generate data, perform analyses, or produce the results of the study. All revisions were reviewed and approved by the authors, who take full responsibility for the final submitted manuscript.

Conflict of Interest

Authors declare that there is no conflict of interests regarding the publication of the paper.

Author Contribution

Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data Curation, Writing-Original Draft, Writing Review & Editing, Supervision, Project administration, Funding acquisition: *Handaru Jati;* ***Conceptualization, Methodology, Formal analysis, Investigation, Resources, Writing - Original Draft, Writing - Review & Editing, Project administration:*** *Yuniar Indrihapsari;* ***Conceptualization, Software, Validation, Investigation, Data Curation, Writing - Original Draft, Visualization:*** *Pradana Setialana, Susilo Romadhon;* ***Methodology, Software, Validation, Formal analysis, Data Curation, Writing - Original Draft, Writing - Review & Editing:*** *Danang Wijaya;* ***Validation, Resources, Writing - Original Draft, Writing - Review & Editing, Visualization:*** *Satya Adhiyaksa Ardy.*

References

[1] Enkelejda Kasneci *et al.* (2023) ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences*, 103, 102-274. https://doi.org/10.1016/j.lindif.2023.102274

[2] Christian Stöhr, Amy Wanyu Ou & Hans Malmström (2024) Perceptions and usage of AI chatbots among students in higher education across genders, academic levels and fields of study. *Computers and Education: Artificial Intelligence*, 7, 100-259. https://doi.org/10.1016/j.caeai.2024.100259

[3] Chunyuan Li *et al.* (2024) Multimodal foundation models: From specialists to general-purpose assistants. *Foundations and Trends in Computer Graphics and Vision*, 16(1-2), 1-214. https://doi.org/10.1016/j.caeai.2024.100259

[4] Michael Moor *et al.* (2023) Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956), 259-265. https://doi.org/10.1038/s41586-023-05881-4

[5] Yupeng Chang *et al.* (2024) A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3), 1-45. https://doi.org/10.1145/3641289

[6] Philip Newton & Maira Xiromeriti (2024) ChatGPT performance on multiple choice question examinations in higher education. A pragmatic scoping review. *Assessment & Evaluation in Higher Education*, 49(6), 781-798. https://doi.org/10.1080/02602938.2023.2299059

[7] Emanuele La Malfa *et al.* (2024) Language-models-as-a-service: Overview of a new paradigm and its challenges. *Journal of Artificial Intelligence Research*, 80, 1497-1523. https://doi.org/10.1613/jair.1.15865

[8] Yishen Song, Qianta Zhu, Huaibo Wang & Qinhua Zheng (2024) Automated essay scoring and revising based on open-source large language models. *IEEE Transactions on Learning Technologies*, 17, 1880-1890. https://doi.org/10.1109/TLT.2024.3396873

[9] Giorgio Biancini, Alessio Ferrato & Carla Limongelli (2024) Multiple-choice question generation using large language models: Methodology and educator insights, in *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, 2024, 584-590, https://doi.org/10.1145/3631700.3665233

[10] J. Jakob Mökander, J. Jonas Schuett, Hannah Rose Kirk & Luciano Floridi (2024) Auditing large language models: a three-layered approach. *AI and Ethics*, 4(4), 1085-1115. https://doi.org/https://doi.org/10.2139/ssrn.4361607

[11] Simone Balloccu, Patrícia Schmidtová, Mateusz Lango & Ondřej Dušek (2024) Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs, in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, 67-93, https://doi.org/10.18653/v1/2024.eacl-long.5

[12] Rishi Bommasani, Percy Liang & Tony Lee (2023) Holistic evaluation of language models. *Annals of the New York Academy of Sciences*, 1525(1), 140-146. https://doi.org/10.1111/nyas.15007

[13] Lianmin Zheng *et al.* (2023) Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36, 46595-46623. https://doi.org/10.52202/075280-2020

[14] Wesley Morris, Langdon Holmes, Joon Suh Choi & Scott Crossley (2025) Automated scoring of constructed response items in math assessment using large language models. *International journal of artificial intelligence in education*, 35(2), 559-586. https://doi.org/10.1007/s40593-024-00418-w

[15] Cagatay Neftali Tulu, Ozge Ozkaya & Umut Orhan (2021) Automatic short answer grading with semspace sense vectors and malstm. *IEEE Access*, 9, 19270-19280. https://doi.org/10.1109/access.2021.3054346

[16] Mark D Shermis & Joshua Wilson (2024) *The Routledge international handbook of automated essay evaluation*. Taylor & Francis.https://doi.org/10.4324/9781003397618-2