Evaluation of Open-ended Question's Answers using Large Language Model (LLM): A Case Study of a Language Learning Center in University

Faisal Wilmar*
Faculty of Computer Science
Universitas Indonesia
Depok, Indonesia
faisalwilmar@gmail.com

Panca Oktavia Hadi Putra
Faculty of Computer Science
Universitas Indonesia
Depok, Indonesia
hadiputra@cs.ui.ac.id

Abstract

Massive open online courses (MOOCs) are a transformative tool in education. Their benefits are increasingly significant along with the development of artificial intelligence (AI) in the form of Large Language Models (LLMs). However, the use of LLMs in education raises several ethical issues, such as accuracy and fairness, especially when LLMs are used for assessment or evaluation. This is because of the risk of bias, inaccuracy, and inconsistency in the AI output. These risks can be reduced with the development of LLM capabilities and the use of prompt engineering techniques, namely, a way of communicating with LLM models. This study aims to evaluate the influence of factors affecting the accuracy of evaluation results from open-ended questions in an English MOOC at a university's language learning center. The results of this evaluation are used to determine recommendations for the consistent and accurate use of LLMs in evaluating open-ended questions in English courses using the MOOC platform. This quantitative quasi-experimental research involved 580 participants divided into three proficiency groups who answered open-ended questions. Participants' answers were evaluated by one human rater and three LLMs using three prompt engineering techniques to generate an evaluation result. The assessment results were analyzed using a three-way analysis of variance (ANOVA) to examine the factors influencing the LLM output. The mean absolute difference (MAE) between the human and LLM assessment results was used to assess the accuracy of LLM. The evaluation result was then calculated using the quadratic weighted kappa to generate inter-rater reliability to assess the raters' consistency. The results showed that the participant group, LLM model used, and prompt engineering technique used influenced the assessment results. The ability level of the evaluated participants had the greatest impact on the results. A combination of LLM and prompt engineering technique, ChatGPT-4.1 with Chain-of-Thought, provided the best results, but it should not be used for high-stakes assessments. LLM can be used for initial assessment or as a complement to human assessment results without replacing actual human raters for high-stakes assessments.

Keywords: MOOC, generative ai, artificial intelligence, large language model, prompt engineering technique, open-ended question, English course

* Corresponding Author

Introduction

Massive Open Online Courses (MOOC) have emerged as transformative platforms in higher education, offering unprecedented access to quality educational resources regardless of geographical location, socioeconomic status, or prior educational attainment (Kay et al., 2013). These platforms, often provided by prestigious institutions, present an up-to-date, flexible, and cost-effective approach to learning (Hodgkinson-Williams, 2014). In the context of language education, MOOCs enable learners to access courses from renowned institutions globally, democratizing language learning opportunities.

The integration of MOOCs into university curricula presents promising opportunities because of their scalability, personalization capabilities, and technological affordances (Shi et al., 2020; Yu et al., 2017). Unlike traditional classroom settings constrained by class size and instructor availability, MOOCs can leverage technology to track student progress, identify strengths and weaknesses, and provide guidance through interactive multimedia content, including videos, interactive exercises, and language-specific activities (Tang & Qian, 2022). This multimedia approach accommodates diverse learning styles and preferences, enhancing student engagement.

However, integrating MOOCs into university curricula remains a significant challenge: the limitation in assessing learning progress and providing meaningful feedback. The self-paced nature and high student-to-instructor ratios of MOOCs render conventional assessment methods impractical (Haefner et al., 2021). This assessment challenge is particularly critical in language education, where timely and constructive feedback is essential for skill development.

Recent advances in artificial intelligence (AI), particularly large language models (LLM), offer potential solutions to this gap in assessment. These models, which are trained on vast textual datasets using advanced deep learning algorithms, demonstrate remarkable capabilities in understanding and generating human-like responses (OpenAI Team, 2024). AI technology can automate the assessment of student assignments, examinations, and essays, saving valuable time and resources while providing constructive feedback to enhance literacy skills (Jalil et al., 2023; Kasneci et al., 2023).

This study addresses the research gap in utilizing LLMs for MOOC-based language education assessment grading and feedback within university settings. While previous research has explored MOOCs in language learning and AI applications in education separately, limited investigation exists on the systematic integration of LLMs for addressing assessment challenges specific to MOOCs in language learning. A critical limitation of the existing research is the lack of or inadequate implementation of scoring rubrics in LLM-based assessment. Previous studies have either operated without structured rubrics or employed purely numeric scoring scales (e.g., 0-100), which lack explicit criteria explaining how and why specific scores should be assigned. The absence of clear evaluative guidelines may result in inconsistent and unreliable AI-generated assessments, as the model lacks structured reference points for judgment.

This research contributes to methodological innovation by implementing categorical scoring criteria that provide LLMs with explicit, structured guidelines for assessment. This study focuses on English language courses within the context of a university language learning center and examines how different rubric designs influence LLM assessment accuracy. The purposes of this research are to identify factors affecting the accuracy and reliability of LLM-based assessments, determine which LLM models and prompt engineering techniques are best suited for educational assessment grading, and propose practical frameworks for integrating LLM technology into existing MOOC assessment systems. By investigating both the opportunities and challenges of this integration through empirical evaluation, this study aims to inform responsible and effective utilization of AI technology in MOOC platforms to enhance language education quality and scalability.

Literature Review

Related Work

Questions can be classified into two types: selected-response (SR) and constructed-response (CR). SR questions provide predetermined answers from which students select one or more options, typically

appearing as true/false items, matching statements, or multiple-choice questions. In contrast, CR questions, which are commonly referred to as open-ended questions, comprise two subtypes: restricted-response and extended-response questions. The restricted-response questions focus on limited aspects and require students to provide brief and precise answers. Extended-response questions, such as essays, problem-based examinations, and scenario-based tasks, require examinees to engage in analysis, evaluation, creativity, and higher-order cognitive abilities (Isaacs et al., 2013). The assessment of open-ended questions presents unique challenges due to their subjective nature and the complexity of evaluating diverse student responses, particularly in language learning contexts where linguistic nuance and creativity must be considered.

Recent advances in AI have introduced potential solutions to these assessment challenges through LLMs, which can generate human-like text, answer questions, and accomplish various language-related tasks (OpenAI Team, 2024). The sophisticated architecture of these models allows them to process and generate contextually appropriate responses across diverse domains, making them potentially valuable tools for educational applications. However, LLMs have several limitations that warrant consideration. Model outputs carry risks of bias, inaccuracy, and inconsistency, which can occasionally produce misleading results (Kasneci et al., 2023; Yan et al., 2024). These limitations are particularly significant in the context of educational assessment, where reliability and fairness are paramount concerns.

Prompt engineering techniques have emerged as critical methodologies for maximizing the effectiveness of LLMs in educational applications. Prompt engineering techniques are systematic approaches for developing and deploying LLMs to direct or control model responses through specific prompt design. This methodology involves crafting effective instructions to guide models in generating the desired outputs. Prompt engineering has become critical in the development of text-based AI, as well-designed prompts can facilitate relevant, logical, and contextually appropriate responses from models. Prompt engineering encompasses several key strategies, such as providing clear instructions and using examples. Through these approaches, users can direct models to generate more specific responses aligned with requirements, especially in scenarios where direct interpretation might yield less accurate or irrelevant answers (Brown et al., 2020; Srivastava et al., 2023). As a rapidly evolving field of study, the standardized prompting framework terminology has not yet been established (Liu et al., 2023). Nevertheless, current research identifies several prompt engineering techniques, including Zero-shot, Few-shot, and Chain-of-Thought prompting (Touvron et al., 2023; Wei et al., 2024).

Zero-shot prompting is the most straightforward technique in which the prompts used to interact with models contain neither examples nor demonstrations of reasoning processes. Zero-shot prompts directly request models to perform tasks without additional instructions (Wei et al., 2022). This approach relies entirely on the pre-existing knowledge and capabilities of the models to interpret and respond to novel tasks. Few-shot prompting extends this approach by incorporating one or more examples of reasoning processes within the prompt. For more complex tasks, few-shot prompts can be enhanced by increasing the number of examples or reasoning demonstrations. Research on in-context learning suggests several recommendations for employing few-shot prompts, including the use of labels to assist LLMs in generating responses. Additionally, the prompt composition format influences how LLMs learn from the provided examples (Min et al., 2022). By providing contextual patterns through examples, such as demonstrating word usage in sentences, it will help the model understand and apply new concepts appropriately to generate a similar construct. Chain-of-Thought (CoT) prompting represents the most sophisticated approach, enabling complex reasoning capabilities through intermediate reasoning steps. This method provides examples of reasoning processes within prompts used for LLMs. In the context of AI and NLP, complex reasoning refers to the models' ability to perform reasoning involving multiple logical steps. This complex reasoning capability is activated by encouraging models to generate intermediate reasoning steps rather than merely presenting final answers directly. In this technique, models receive thought examples that guide them through several logical thinking steps to reach conclusions (Wei et al., 2024). The provision of these intermediate reasoning steps demonstrates patterns that models can follow when solving similar problems, thereby producing the expected responses through explicit logical progression.

Recent empirical studies have demonstrated varying degrees of success in employing LLMs for educational assessment across different contexts and methodologies. Henkel et al. (2024) evaluated the

performance of LLM in marking short-answer questions in K-12 science and history education using 1700 student responses from the Carousel platform. Their study employed few-shot prompting with simple rubrics and achieved Cohen's Kappa value of 0.7, approaching the inter-rater reliability between human evaluators (κ=0.75), thereby demonstrating the viability of LLM for low-stakes formative assessment (Henkel et al., 2024). Similarly, Tate et al. (2024) investigated holistic essay scoring using minimally prompted LLMs, comparing GPT-3.5 and GPT-4 performance against certified human raters across three essay samples involving 1786 students. Their findings revealed AI-AI consistency rates of 60% for GPT-3.5 and 80% for GPT-4, compared to 96% human-human agreement and 92-95% human-AI agreement, respectively, suggesting that while LLMs show promise for low-stakes assessment, they remain insufficient for higher-stakes evaluation without further refinement (Tate et al., 2024).

Investigations into specific prompting techniques have yielded important insights regarding optimal LLM deployment strategies. Mendonça (2024) examined ChatGPT-4 Vision and ChatGPT-4 Turbo performance on Brazil's National Undergraduate Computer Science Exam (ENADE 2021) using zero-shot chain-of-thought prompting with multi-step reassessment processes. Despite the sophisticated prompting approach, the model achieved only 54.2% accuracy, with subsequent expert panel analysis identifying ambiguous statement construction as a significant contributor to performance degradation, consistent with findings by Srivastava et al. (2023) regarding bias introduction through ambiguous prompts (Mendonça, 2024; Srivastava et al., 2023). Lee et al. (2024) conducted a comparative analysis of zero-shot, few-shot, and chain-of-thought prompting techniques for middle school science assessment in the United States, evaluating 1,650 student responses using a three-level rubric (Beginning, Developing, Proficient). Their results demonstrated that few-shot prompting outperformed zero-shot approaches, while chain-of-thought techniques yielded significant improvements only when accompanied by explicit scoring rubrics, achieving accuracy increases of up to 13.44% when rubrics and problem context were provided (Lee et al., 2024).

Further research has explored LLM capabilities across cognitive complexity levels and assessment types. Murali et al. (2024) assessed five popular chatbots (OpenAI ChatGPT, Bing Chat, Google's Bard, Baize, and OpenAssistant's LLM) on computer science assignments categorized by Bloom's Taxonomy levels, employing zero-shot and instruction prompting techniques. Their findings indicated strong chatbot performance on lower cognitive levels (remembering and understanding) with significant performance deterioration on higher-order thinking tasks (applying and analyzing) (Murali et al., 2024). Most recently, Stasuik (2025) evaluated LLM performance in essay assessment for university English courses by comparing self-hosted models (Llama 3.1 and 3.2) with OpenAI-hosted models (GPT-4o-mini, o3-mini, and o1) using various prompting techniques, including chain-of-thought, tree-of-thought, and few-shot approaches. The study employed numeric scoring (1-100) and evaluated results using quadratic weighted kappa (QWK) and mean square error (MSE), with OpenAI's o1 model achieving the highest accuracy (MSE=10.19) and o3-mini with few-shot prompting attaining the highest inter-rater reliability ($\kappa = 0{,}6075$) (Stasuik, 2025).

Analysis of the existing literature reveals several consistent patterns and persistent gaps in LLM-based assessment research. Zero-shot and chain-of-thought prompting techniques emerge as the most frequently investigated approaches across diverse educational contexts. However, a critical limitation pervades previous research: studies examining open-ended question assessment, whether for evaluation or response generation, typically omit explicit answer criteria or employ purely numeric rubrics that generate numerous possible outcomes without structured evaluative guidance. While some investigations have incorporated scoring rubrics, these predominantly utilize numeric formats that lack the categorical distinctions necessary for consistent LLM interpretation. Furthermore, previous studies have employed various inter-rater reliability metrics, including Cohen's kappa, quadratic weighted kappa, and Fleiss's kappa, with selection depending on the number of raters involved in the assessment process.

Building upon these foundations, the present study addresses identified methodological gaps by implementing zero-shot, few-shot, and chain-of-thought prompting techniques augmented with explicit answer criteria designed to facilitate LLM evaluation consistency. This study employs the quadratic weighted kappa as the primary evaluation metric for calculating inter-rater reliability, a method selected for its capacity to account for disagreement magnitude by assigning greater weight to larger

discrepancies compared to minor differences. The evaluation result generated through the research process will be analyzed with consideration of agreement by chance, thereby providing a more nuanced assessment of LLM reliability in educational contexts. This study aims to enhance the consistency and reliability of LLM assessment while maintaining the flexibility necessary for evaluating complex language learning outcomes by incorporating structured categorical scoring criteria rather than purely numeric scales.

Hypotheses

Based on the evaluation of previous studies, this study explores the potential use of LLMs in evaluating student responses across diverse student groups. The experiment involves research participants divided into three proficiency levels (Basic Elementary, Intermediate, and Advanced Proficient), with each participant assigned to only one proficiency level. Each proficiency level receives different questions of the same type, collectively referred to as Student Groups. This design enables between-subject comparisons of mean differences across independent participant groups (Howell, 2007). Each participant's responses will be evaluated using three different prompt engineering techniques, each of which will be applied to three LLM models. This configuration allows within-subject testing of effects across different conditions, which is statistically more powerful as it controls individual variability (Howell, 2007). Hypotheses for this research are presented in Table 1.

Table 1. Research Hypotheses

No	Hypotheses
H1	Significant differences exist in mean absolute difference scores between students across student groups
H2	Significant differences exist in mean absolute difference scores among the three prompt engineering techniques
H3	Significant differences exist in mean absolute difference scores among the three LLM models utilized
H4	The effect of prompt engineering technique depends on the utilized LLM model
H5	The effect of prompt engineering technique depends on student group
H6	Performance differences between LLM models depends on student group
H7	The interaction between prompt engineering technique and LLM model varies depending on student group

These hypotheses are tested against the null hypothesis (H_0) stating "no differences/interactions exist between factors" and the alternative hypothesis (H_α) stating "differences/interactions exist between factors".

H1. Significant differences exist in mean absolute difference scores between students across student groups

The first dimension of the study categorizes students into Basic, Intermediate, and Proficient ability levels, serving as the sole between-subjects factor, as individual essays cannot simultaneously represent multiple proficiency levels. This categorization derives from the theoretical relationship between linguistic noise and model uncertainty in large language model operations. LLMs fundamentally operate via next-token prediction by minimizing perplexity, and Research shows that these models exhibit varying perplexity levels depending on the quality of the input text (Zhong et al., 2025). Essays from Basic students contain non-standard syntax, orthographic errors, and fragmented logical structures, which constitute out-of-distribution data for models trained primarily on high-quality, curated text. High perplexity input destabilizes attention mechanisms and increases output variance, leading to scoring instability in automated essay scoring contexts. Conversely, Proficient essays

adhering to standard grammatical structures produce lower perplexity and higher scoring confidence (Liang et al., 2023). This study, therefore, hypothesizes significant differences in mean absolute difference scores across the three ability groups.

H2. Significant differences exist in mean absolute difference scores among the three prompt engineering techniques

Prompt engineering has evolved into a structured discipline where reliability depends on both engineering methods and evaluation metrics (Sarim et al., 2025). The three techniques examined—zero-shot, few-shot, and chain-of-thought—fundamentally alter the cognitive pathways models use to reach decisions. Zero-shot prompting assesses models' emergent capability to apply rubrics without examples (Sivarajkumar et al., 2024). Few-shot technique leverages in-context learning, where provided examples serve as anchors in vector space, effectively fine-tuning the model's immediate state to align with specific scoring distributions (Cheng et al., 2025). Chain-of-thought prompting externalizes decision-making processes, making it particularly effective for complex reasoning tasks such as essay grading (Wu et al., 2025). Because these techniques represent distinct approaches, this study hypothesizes significant differences in mean absolute difference scores among the three prompt engineering methods.

H3. Significant differences exist in mean absolute difference scores among the three LLM models utilized

Comparative analyses reveal that DeepSeek, ChatGPT, and Gemini exhibit distinct performance characteristics across evaluation parameters, with each model demonstrating particular strengths in reasoning and multilingual understanding (Rahman et al., 2025). Previous studies examining AI models for educational assessment found that different models systematically produce different scores, with Gemini providing superior interrater reliability with human raters. At the same time, DeepSeek excels in consistency measures (Oğuz, 2025). These performance variations stem from divergent training objectives and architectural differences across models. Given these fundamental differences in model design and training approaches, this study hypothesizes significant differences in mean absolute difference scores among the three LLM types.

H4. The effect of prompt engineering technique depends on the utilized LLM model

Research on mathematical and logical tasks demonstrates that while few-shot prompting benefits standard models, zero-shot chain-of-thought prompting often proves superior for reasoning-oriented models (Cheng et al., 2025). This differential effectiveness suggests that prompting techniques interact with model architecture rather than producing uniform effects across all systems. Different models may respond divergently to the same prompting approach due to variations in their underlying attention mechanisms, training paradigms, and reasoning capabilities.

H5. The effect of prompt engineering technique depends on student group

Assessing proficient essays is considered a low-difficulty task because quality indicators are easily identifiable. In contrast, evaluating basic essays presents a high level of difficulty, as raters must separate grammatical errors from the underlying conceptual content. Chain-of-thought prompting has been shown to be increasingly effective as task complexity rises, offering only marginal benefits for simple tasks but yielding significant advantages for more complex ones (Wu et al., 2025). This relationship between task complexity and performance indicates that prompting techniques may vary in effectiveness depending on student ability levels.

H6. Performance differences between LLM models depends on student group

Models equipped with stringent safety filters, especially proprietary systems such as Gemini and GPT, often interpret incoherent text typical of Basic student essays as nonsensical or as potential safety violations. This interpretation can result in scoring refusals or outputs with low confidence. Research indicates that undesirable behaviors in LLMs occur disproportionately among users with lower English

proficiency, which makes these models unreliable for vulnerable populations (Poole-Dayan et al., 2024). The observed differential sensitivity to text quality implies that models are likely to demonstrate varying performance gaps across different student proficiency levels.

H7. The interaction between prompt engineering technique and LLM model varies depending on student group

This hypothesis integrates the preceding theoretical frameworks into a unified analytical model, proposing that no single optimal AI configuration exists for automated essay scoring. Instead, scoring reliability depends on alignment among prompt engineering techniques, LLM architecture, and student proficiency levels. The interaction between prompting methods and model types may vary depending on the proficiency of student essays, as models are influenced by both the cognitive scaffolding provided by prompts and the linguistic complexity present in student writing.

Methodology

This study employs a quantitative quasi-experimental design to examine the effectiveness of LLMs in evaluating student responses across varying conditions. The study manipulates three independent variables—student proficiency level, prompt engineering technique, and LLM model—while controlling question type and assessment criteria as constants to isolate treatment effects. The dependent variable measured is the score difference between LLM-generated evaluations and human rater assessments, expressed as numeric values derived from ordinal categorical ratings.

Quantitative Quasi-Experimental Design

This study adopts a quantitative approach wherein objective theories are tested by analyzing relationships among variables (Creswell, 2017). Variables are measured numerically using structured research instruments, and the data are analyzed using relevant statistical procedures to determine whether the collected evidence supports or rejects the formulated hypotheses. This study selected the quasi-experimental design because random assignment was not possible, as student groups are naturally formed based on pre-existing English proficiency levels within the language institution. While this lack of randomization may threaten internal validity, the design remains appropriate for testing causal relationships under practical conditions where random assignment is infeasible (Cook et al., 2002; Creswell, 2017).

Human raters serve as the control group, representing the established standard in educational assessment. This designation is based on the principle that human evaluators bring unique qualities—including deep contextual understanding, critical reasoning, and nuanced interpretation—that algorithms cannot fully replicate. Despite potential subjective biases, human raters function as the baseline or "traditional approach," enabling direct comparison of LLM performance against established assessment standards (Stasuik, 2025). LLMs constitute the experimental group, representing the novel "treatment" whose effects are measured (Siska Merrydian et al., 2024). The primary objective of the research is to evaluate whether LLMs can produce assessment results comparable to those of human instructors and to determine the impact of prompt engineering techniques in achieving this goal.

Internal Validity

To address selection bias—the most common threat in quasi-experimental designs where systematic pre-existing differences between groups may be mistakenly attributed to treatment effects (Siska Merrydian et al., 2024)— this study implemented several mitigation strategies. All students received identical instructions, equivalent completion times, and similar environmental conditions to reduce external-factor bias. Assessment processes randomized student evaluation order to minimize bias from sequential assessments of students with varying proficiency levels.

Instrumentation bias, occurring when measurement instruments change over time or differ between groups (Siska Merrydian et al., 2024), was addressed through multiple strategies. Test questions were sourced from the Language Learning Center curriculum, pre-validated by lecturers to ensure alignment

with targeted competencies and student proficiency levels. Scoring rubrics were designed with a detailed structure to minimize ambiguity and subjective interpretation, ensuring both human and AI raters employed identical criteria. This specificity ensures consistent construct measurement, minimizing undesired variation and attributing result differences to treatment rather than instrumental variance. However, a limitation persists: the Language Learning Center's natural division of students into three proficiency levels, each receiving different questions of the same type, creates a confound between proficiency level and question content.

Experimental Procedure

The complete experimental procedure, illustrated in Figure 1, was conducted centrally through the Language Learning Center's Moodle-based Learning Management System (LMS) operating within an intranet environment without internet connectivity, preventing students from accessing external assistance such as Generative AI tools. Students were grouped and assigned to classes according to English proficiency levels, with experimental questions available in courses corresponding to each level. Students completed assessments in the LMS according to standard examination protocols, with responses automatically collected by the system for subsequent human evaluation.

Following response collection, instructors serving as human raters evaluated answers based on established rubrics, with assessments stored systematically for further processing. Each response received evaluation from one randomly assigned human rater matched by proficiency level. Subsequently, student responses previously evaluated by a human rater were assessed by three LLM models, simulating conditions comparable to human evaluation.

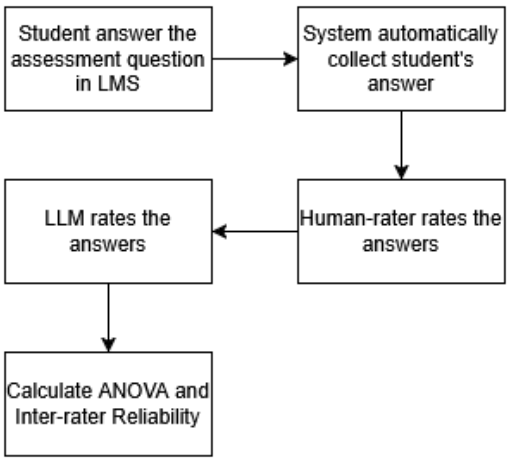

Figure 1. Experimental Procedure

Participant Selection and Case Study Approach

Student Selection

To calculate inter-rater reliability using Kappa statistics, student participants were selected using a whole-population method encompassing all 580 students at the English Language Learning Center. This method aligns with population characteristics—naturally stratified homogeneous groups based on proficiency levels—and manageable population size. Utilizing homogeneous population groups reduces within-strata variability, enabling more precise estimation of group characteristics and ultimately enhancing the accuracy of conclusions about the overall population. The resulting sample facilitates diagnostic analysis of LLM performance across proficiency levels, with proportionally distributed data enabling valid and meaningful cross-strata comparisons.

Human Rater Selection

Research findings regarding the relationship between rater experience and assessment quality present mixed conclusions. Highly experienced raters tend to produce more positive evaluations and higher scores than less experienced raters (Şahan, 2018), while also demonstrating superior adherence to established rubrics and greater resistance to contextual influences, resulting in more consistent scores over time (Zhao et al., 2017). However, other research suggests this relationship is neither linear nor straightforward, with differences between experienced and inexperienced raters (the latter appearing stricter with lower scores) proving statistically insignificant (Leckie & Baird, 2011).

Based on these findings, human rater selection criteria include: (1) minimum five years active English teaching experience at relevant proficiency levels; (2) demonstrated experience evaluating academic essays, such as involvement as standardized language test raters (IELTS or TOEFL) or institutional final examination chief raters; (3) relevant academic qualifications, including degrees in English Language Teaching, Applied Linguistics, or English Literature; and (4) participation in rater training and calibration sessions involving sample questions, rubrics, and responses, with raters evaluating identical answers, discussing scores, and aligning rubric interpretations until achieving consensus.

LLM Model Selection

During the research design, Generative AI using LLMs represented a rapidly evolving technology with limited research on their capacity for response evaluation, particularly for open-ended questions in English courses. Model selection, therefore, relied on multiple reference sources. ChatGPT, Gemini, and DeepSeek were identified as leading AI models in an article from Swiss German University (SGU, 2025), supported by GlobalStats data showing ChatGPT's highest market share (StatCounter, 2025). Academic research applications were further supported by Google's program, which allows free academic access to Gemini Pro (Google, 2025).

Multiple comparative studies have examined these three models across various dimensions. Rahman et al., (2025) conducted a comprehensive comparison of ChatGPT, Gemini, and DeepSeek, evaluating their features, architectural techniques, performance metrics, and prospects, using different model versions than those employed in this research. Similarly, Oğuz, (2025) compared these models specifically for essay scoring applications, albeit with a methodological approach distinct from the present study. These comparative analyses support the selection of these three platforms as representative of current LLM capabilities.

Official OpenAI documentation indicates that chain-of-thought prompt engineering techniques are less relevant for reasoning models, which perform reasoning internally without enabling user oversight or specific reasoning direction (OpenAI Team, 2025b, 2025a). Consequently, this research employs the most recent non-reasoning models from each platform: ChatGPT-4.1 (CG), Gemini-Flash (GF), and DeepSeek-Chat (DC).

Data Collection Method

Data collection proceeded through three sequential stages: collecting student responses, obtaining human rater assessments, and conducting LLM-based evaluations using prompt engineering techniques.

Student Response Collection

After receiving institutional approval and consent, the Language Learning Center used Moodle to deliver research instruments to participants. This system supports computer-based exams, presents open-ended questions, stores responses, and verifies participant identities, aligning with our focus on MOOC platform analysis. We collected responses during one mid-semester week, coordinating with the institution's exam schedule. We then masked identifying information to ensure anonymity. This process produced a comprehensive, de-identified dataset of participant responses.

Human Rater Assessment

Subsequently, student responses underwent evaluation by human raters, with each response assessed by one designated rater. Human evaluation generated an assessment list containing participant masked identifiers and corresponding ordinal categorical scores, as illustrated in Table 2.

Table 2. Rater Assessment Results

Participant	Score
Participant 1	3
Participant 2	4
Participant 3	2
...	...
Participant N	4

LLM-Based Assessment

Following human evaluation, responses were assessed using LLMs that employed three prompt engineering techniques, which are Zero-shot (ZS), Few-shot (FS), and Chain-of-Thought (CoT). AI-rater assessments produced evaluation lists analogous to Table 2.

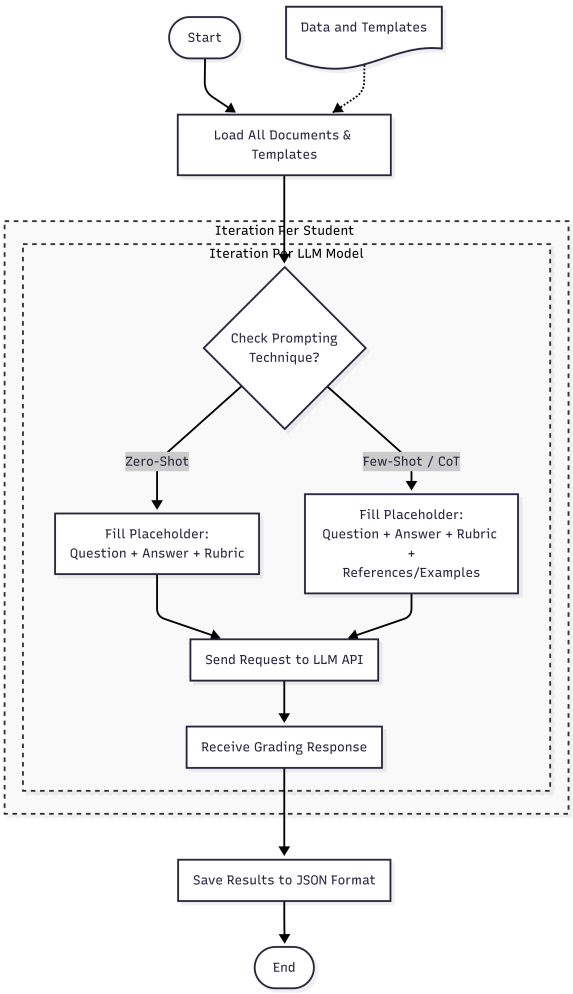

Figure 2. Assessment Software Workflow using LLM

Assessment implementation used a custom Java console application, with the operational workflow depicted in Figure 2. The application begins by loading all necessary documents and templates, including student responses, questions, rubrics, and prompt templates. For each student answer and LLM model combination, the system determines the appropriate prompting technique. Zero-shot prompts fill placeholders with the question, student answer, and scoring rubric only. Few-shot and Chain-of-Thought prompts additionally incorporate reference examples as reasoning guides. The application then sends the constructed prompt to the respective LLM API, receives the grading response, and stores results in JSON format for subsequent analysis.

The complete application source code is available on GitHub (see Appendix A for repository link). Appendix B provides the prompt templates used for each technique, Appendix C presents example questions from one of the proficiency level, Appendix D contains the scoring rubric structure, and Appendix E demonstrates the reasoning guide format. This implementation ensures consistent prompt construction and systematic data collection across all experimental conditions.

Data Analysis Method

Data analysis proceeded through three sequential stages, implemented using Python with specialized statistical packages. In the first stage, Human-AI reliability between human raters and LLM models was computed using Quadratic Weighted Kappa (QWK) from sklearn.metrics. Next, the second stage calculated inter-rater reliability within the LLM group (Inter-AI reliability), also employing QWK from sklearn.metrics. Subsequently, the final stage employed ANOVA to identify factors influencing LLM assessment outcomes, utilizing the statsmodels.stats.anova module. Throughout all stages, data manipulation and organization leveraged the pandas library for efficient dataframe operations. The complete Python implementation, including detailed analysis notebooks, is available on GitHub (see Appendix F).

Human-AI Reliability

Human-AI reliability was calculated using Quadratic Weighted Kappa through pairwise comparisons between human raters and LLMs. Inter-rater reliability values from each pairing were averaged to produce the final Human-AI reliability result, with the calculation process exemplified in Table 3.

Table 3 Calculation Example for Reliability

Human Rater	**AI Rater**	**Human-AI Reliability**
Human Rater	AI Rater 1	~0.45
Human Rater	AI Rater 2	~0.33
Human Rater	AI Rater 3	~0.42
Human-AI Reliability		~0.40

Inter-AI Reliability

Inter-AI reliability was calculated using Quadratic Weighted Kappa. This calculation examines consistency within the LLM group, with higher Inter-AI reliability values indicating greater LLM consistency in conducting assessments, particularly consistency in producing identical or similar outputs when provided identical or similar inputs.

ANOVA

ANOVA calculations were based on mean absolute differences in scores between human raters and LLM outputs, examining score discrepancies between LLM-generated assessments and human evaluations. This analysis identifies factors influencing score differences by comparing results across combinations of independent variable groups, employing a three-way ANOVA design. The significance level was determined using Fisher's standard at 0.05, representing a 1:20 probability threshold for

determining whether results are statistically significant for further investigation. This threshold indicates the research accepts a 5% probability of Type I error—rejecting the null hypothesis when it is actually true (false positive)—concluding an effect exists when none is present (Fisher, 1925).

Effect Size Interpretation

Effect size provides a standardized metric measuring the strength of relationships between variables or differences among treatment groups, revealing practical importance beyond mere statistical significance (Cohen, 2013). This study employs Cohen's d, which expresses mean differences between groups in standard deviation units, making results independent of original measurement scales. Cohen (2013) established conventional interpretation benchmarks (see Table 4), which enable evaluation of whether statistically significant ANOVA findings possess meaningful practical implications for automated essay scoring applications.

Table 4. Interpretation of Cohen's d

Cohen's *d*	Interpretation
0.20	Small
0.50	Moderate
0.80	Large

Quadratic Weighted Kappa Interpretation

Quadratic Weighted Kappa (QWK) measures inter-rater agreement while adjusting for chance agreement, providing more accurate reliability assessments than simple percentage agreement (Cohen, 1960; Fleiss, 1971). QWK specifically addresses ordinal categorical data by assigning differential weights to disagreements based on their magnitude (Cohen, 1968), thus penalizing larger rating discrepancies more severely (Kvålseth, 2018; Rau & Shih, 2021). Landis and Koch (1977) provide widely accepted interpretation benchmarks for kappa values (see Table 5).

Table 5. Interpretation of Kappa value

Kappa value	Interpretation
< 0	*Poor 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*
0.80 – 1.00	*Almost perfect agreement*

Results

Interaction Between Prompt Engineering Technique, Large Language Model, and Student Group

H1: Significant Differences in Mean Absolute Score Differences Across Student Groups

Significant differences were observed in mean absolute score differences across student groups ($p = 2.4\times10^{-13}$, $F = 29.24$), though the effect size was small ($\eta^2p = 0.0125$). Post-hoc pairwise comparisons with Bonferroni correction revealed that LLM assessment discrepancies were significantly higher for Advanced and Proficient students ($M = 0.816$, $SD = 0.684$) compared to both Intermediate ($M = 0.564$, $SD = 0.615$, $p\_bonf < 0.001$) and Basic and Elementary students ($M = 0.587$, $SD = 0.600$, $p\_bonf <$

0.001). No significant difference existed between Basic and Elementary and Intermediate groups (p_bonf > 0.999), as shown in Table 6 and Table 7. These findings indicate that LLMs demonstrate higher accuracy when evaluating Basic and Elementary and Intermediate student responses, with accuracy declining for Advanced and Proficient students.

Table 6. Mean Absolute Score Difference of Student Groups

Student Group	Mean	Std Dev
Advanced Proficient (AP)	0,816	0,684
Intermediate (I)	0,564	0,615
Basic Elementary (BE)	0,587	0,600

Table 7. Post-Hoc Pairwise Comparisons (Bonferroni) Between Student Group

Comparison	t-value	p-value	p_bonf	Mean diff
Basic Elementary vs Intermediate	0,505	0,6138	1,0000	0,0221
Basic Elementary vs Advanced Proficient	-4,215	0,0000	0,0001	-0,2294
Intermediate vs Advanced Proficient	-5,161	0,0000	0,0000	-0,2515

H2: Significant Differences in Mean Absolute Score Differences Across Prompt Engineering Techniques

Significant differences emerged across prompt engineering techniques ($p = 5.5\times10^{-3}$, $F = 5.22$), with a small effect size ($\eta^2p = 0.0023$). Post-hoc analysis with Bonferroni correction indicated that Chain-of-Thought prompting ($M = 0.606$, $SD = 0.613$) produced smaller discrepancies from human rater assessments compared to both Zero-shot ($M = 0.653$, $SD = 0.631$, p_bonf = 0.0027) and Few-shot techniques ($M = 0.637$, $SD = 0.654$, p_bonf = 0.0137). No significant difference was found between Zero-shot and Few-shot approaches (p_bonf = 0.8017), as presented in Table 8 and

Table 9.

Table 8. Mean Absolute Score Differences by Prompt Engineering Technique

Prompt	Mean	Std Dev
Chain-of-Thought (CoT)	0,606	0,613
Few-shot (FS)	0,637	0,654
Zero-shot (ZS)	0,653	0,631

Table 9. Post-Hoc Pairwise Comparisons (Bonferroni) between Prompt Techniques

Comparison	t-value	p-value	p_bonf	Mean diff
ZS vs FS	1,110	0,2672	0,8017	0,0161
ZS vs CoT	3,333	0,0009	0,0027	0,0471
FS vs CoT	2,848	0,0046	0,0137	0,0310

These results suggest that techniques incorporating explicit reasoning steps facilitate LLM assessments more closely aligned with human evaluation, while merely providing examples (Few-shot) offers minimal advantage over providing no examples (Zero-shot).

H3: Significant Differences in Mean Absolute Score Differences Across LLM Models

Significant differences were identified across LLM models ($p = 3.5\times10^{-4}$, $F = 7.98$), with a small effect size ($\eta^2p = 0.0034$). Post-hoc analysis revealed that ChatGPT-4.1 (M = 0.603, SD = 0.585) produced smaller discrepancies from human assessments compared to Gemini-Flash (M = 0.662, SD = 0.662, p_bonf = 0.0097). No significant differences were found between ChatGPT-4.1 and DeepSeek-Chat (p_bonf = 0.576) or between DeepSeek-Chat and Gemini-Flash (p_bonf = 0.427), as shown in Table 10 and

Table 11. These findings indicate that ChatGPT-4.1 closely approximates human assessment patterns, establishing it as the most accurate model in this study.

Table 10. Mean Absolute Score Differences by LLM Model

LLM	Mean	Std Dev
ChatGPT-4.1 (CG)	0,603	0,585
DeepSeek-Chat (DC)	0,631	0,648
Gemini-Flash (GF)	0,662	0,662

Table 11. Post-Hoc Pairwise Comparisons (Bonferroni) Between LLM Models

Comparison	t-value	p-value	p_bonf	Mean diff
CG vs GF	-2,958	0,003	0,0097	-0,0592
DC vs CG	-1,306	0,192	0,5760	-0,0276
GF vs DC	1,469	0,142	0,4273	0,0316

H4: Effect of Prompt Engineering Technique Depends on LLM Model

A statistically significant, though small, interaction was observed between prompt engineering technique and LLM model ($p = 3.7\times10^{-2}$, $F = 2.55$, $\eta^2p = 0.0022$). Table 12 demonstrates that Chain-of-Thought achieved optimal performance with ChatGPT-4.1 (M = 0.590), producing the lowest mean absolute difference compared to other techniques on the same model—Few-shot (M = 0.621) and Zero-shot (M = 0.600)—and compared to other models using the same technique—DeepSeek-Chat (M = 0.622) and Gemini-Flash (M = 0.607).

Table 12. Mean Absolute Score Differences by Model × Prompt Interaction

LLM	Chain-of-Thought	Few-shot	Zero-shot
ChatGPT-4.1	0,590	0,621	0,600
DeepSeek-Chat	0,622	0,628	0,643
Gemini-Flash	0,607	0,664	0,717

Simple effect analysis—shown in Table 13—revealed that Chain-of-Thought produced no statistically significant differences across models ($p > 0.05$). However, with Zero-shot prompting, Gemini-Flash exhibited the highest discrepancies (M = 0.0741 and M = 0.1172) compared to ChatGPT-4.1 and DeepSeek-Chat (M = 0.0431). These results demonstrate that ChatGPT-4.1 generally performs better regardless of prompt technique, while Chain-of-Thought equalizes performance across models. Chain-

of-Thought particularly enhances Gemini-Flash accuracy, supporting the hypothesis that prompt engineering technique effects depend on the LLM model employed.

Table 13. Simple Effect Analysis of Model × Prompt

Prompt Technique	Model Comparison	p-value	Mean diff
Zero-shot	ChatGPT vs Gemini	0,0000	-0,1172
Zero-shot	ChatGPT vs DeepSeek	0,0473	-0,0431
Zero-shot	Gemini vs DeepSeek	0,0063	0,0741
Few-shot	ChatGPT vs Gemini	0,1233	-0,0431
Few-shot	ChatGPT vs DeepSeek	0,8249	-0,0069
Few-shot	Gemini vs DeepSeek	0,2154	0,0362
Chain-of-Thought	ChatGPT vs Gemini	0,4864	-0,0172
Chain-of-Thought	ChatGPT vs DeepSeek	0,2076	-0,0328
Chain-of-Thought	Gemini vs DeepSeek	0,5472	-0,0155

H5: Effect of Prompt Engineering Technique Depends on Student Group

No statistically significant interaction was found between prompt engineering technique and student group ($p = 1.4\times10^{-1}$, $F = 1.72$). Even if this null hypothesis acceptance were erroneous, the potential interaction would demonstrate minimal impact ($\eta^2p = 0.0015$).

H6: LLM Performance Differences Depend on Student Group

A statistically significant interaction emerged between LLM model and student group ($p = 8.5\times10^{-6}$, $F = 7.23$), though with small effect size ($\eta^2p = 0.0062$). Table 14 shows that all LLMs produced higher mean absolute differences for Advanced and Proficient students (ChatGPT-4.1: M = 0.719, DeepSeek-Chat: M = 0.868, Gemini-Flash: M = 0.861). For Advanced and Proficient students, ChatGPT-4.1 performed best, while DeepSeek-Chat showed the largest discrepancy from human assessments, though comparable to Gemini-Flash. For other student groups, all three LLMs exhibited similar performance with 2-3% variation, except Gemini-Flash which reached 7% difference.

Table 14. Mean Absolute Score Differences by Model × Student Group

LLM	Student Group	*Mean*	Std Dev
ChatGPT-4.1	Advanced Proficient	0,719	0,551
ChatGPT-4.1	Basic Elementary	0,575	0,612
ChatGPT-4.1	Intermediate	0,560	0,573
DeepSeek-Chat	Advanced Proficient	0,868	0,686
DeepSeek-Chat	Basic Elementary	0,546	0,598
DeepSeek-Chat	Intermediate	0,564	0,634
Gemini-Flash	Advanced Proficient	0,861	0,787
Gemini-Flash	Basic Elementary	0,638	0,630
Gemini-Flash	Intermediate	0,569	0,582

Simple effect analysis (Table 15) indicated that comparisons involving Advanced and Proficient students achieved high significance across all models ($p < 0.05$). All models demonstrated performance

decline when evaluating Advanced and Proficient students, while maintaining comparable accuracy for other groups. ChatGPT-4.1 exhibited the best performance with the lowest increase in mean absolute differences (0.560→0.719, 28.4%; 0.575→0.719, 25%) compared to other LLMs (DeepSeek-Chat: 53.6% and 49.9%; Gemini-Flash: 51.1% and 34.6%), supporting the hypothesis that LLM performance varies significantly across student groups.

Table 15. Simple Effect Analysis - Model × Student Group

LLM	Student Group Comparison	*p-value*	Mean diff
ChatGPT-4.1	Basic Elementary vs Intermediate	0.7724	0.0143
ChatGPT-4.1	Basic Elementary vs Advanced Proficient	0.0159	-0.1446
ChatGPT-4.1	Intermediate vs Advanced Proficient	0.0018	-0.1589
DeepSeek-Chat	Basic Elementary vs Intermediate	0.7239	-0.0187
DeepSeek-Chat	Basic Elementary vs Advanced Proficient	0.0000	-0.3222
DeepSeek-Chat	Intermediate vs Advanced Proficient	0.0000	-0.3035
Gemini-Flash	Basic Elementary vs Intermediate	0.1620	0.0707
Gemini-Flash	Basic Elementary vs Advanced Proficient	0.0015	-0.2215
Gemini-Flash	Intermediate vs Advanced Proficient	0.0000	-0.2922

H7: Three-Way Interaction Between Prompt Engineering Technique, LLM Model, and Student Group

A statistically significant three-way interaction was observed between prompt engineering technique, LLM model, and student group ($p = 4.7\times10^{-2}$, $F = 1.96$, $\eta^2p = 0.03$). Although small, this effect size was larger than other interactions due to interaction complexity. This finding demonstrates that interactions among independent variable factors collectively influence dependent variable outcomes.

Inter-Rater Reliability Analysis

Inter-AI Reliability

Within-model consistency testing revealed that ChatGPT-4.1 achieved the highest reliability with QWK values ranging from 0.720 to 0.868 (M = 0.794), classified as substantial agreement. Using Few-shot and Chain-of-Thought prompting, ChatGPT-4.1 reached almost perfect agreement ($\kappa = 0.868$), demonstrating consistent performance regardless of prompt engineering technique. Conversely, DeepSeek-Chat exhibited greater variability across prompt techniques with QWK values ranging from 0.632 to 0.809 (M = 0.698), particularly with Zero-shot and Chain-of-Thought combinations ($\kappa = 0.632$), as shown in Table 16. These results indicate ChatGPT-4.1's stability across different prompting strategies, while DeepSeek-Chat demonstrates sensitivity to prompt technique selection, reinforcing the previously identified interaction between prompt engineering technique and LLM model.

Table 16. Within-Model Consistency

LLM	Comparison	QWK	Std Dev
ChatGPT-4.1	Few-shot vs Chain-of-Thought	0,8675	0,4115
ChatGPT-4.1	Zero-shot vs Chain-of-Thought	0,7968	0,4775
ChatGPT-4.1	Zero-shot vs Few-shot	0,7196	0,5495
Gemini-Flash	Few-shot vs Chain-of-Thought	0,7538	0,5048

LLM	Comparison	QWK	Std Dev
Gemini-Flash	Zero-shot vs Chain-of-Thought	0,7070	0,5466
Gemini-Flash	Zero-shot vs Few-shot	0,6687	0,5708
DeepSeek-Chat	Few-shot vs Chain-of-Thought	0,8087	0,4530
DeepSeek-Chat	Zero-shot vs Chain-of-Thought	0,6524	0,5593
DeepSeek-Chat	Zero-shot vs Few-shot	0,6324	0,5716

Cross-model consistency analysis—employing identical prompt techniques across different models—revealed variation dependent on prompt engineering technique. Chain-of-Thought produced the highest cross-model reliability ($M_\kappa = 0.644$), with all pairwise comparisons achieving substantial agreement ($\kappa = 0.614$-0.666). Zero-shot demonstrated the lowest consistency ($M_\kappa = 0.573$), particularly between Gemini-Flash and DeepSeek-Chat ($\kappa = 0.502$), as presented in Table 17. These findings indicate that Chain-of-Thought generates consistent outputs across different LLMs, while Zero-shot—lacking examples or reasoning guidance—produces divergent outputs, reinforcing the significant effect of prompt engineering technique selection.

Table 17. Cross-Model Consistency

Prompt Tech.	Comparison	QWK	Std Dev
Chain-of-Thought	ChatGPT-4.1 vs Gemini-Flash	0,6660	0,6186
Chain-of-Thought	ChatGPT-4.1 vs DeepSeek-Chat	0,6510	0,5797
Chain-of-Thought	Gemini-Flashvs DeepSeek-Chat	0,6139	0,6045
Few-shot	ChatGPT-4.1 vs Gemini-Flash	0,6048	0,6835
Few-shot	ChatGPT-4.1 vs DeepSeek-Chat	0,5733	0,6261
Few-shot	Gemini-Flashvs DeepSeek-Chat	0,5292	0,6781
Zero-shot	ChatGPT-4.1 vs Gemini-Flash	0,6483	0,5117
Zero-shot	ChatGPT-4.1 vs DeepSeek-Chat	0,5659	0,6464
Zero-shot	Gemini-Flashvs DeepSeek-Chat	0,5024	0,6258

Overall inter-rater agreement across LLMs with 36 pairwise comparisons achieved substantial agreement ($M_\kappa = 0.612$, $SD = 0.095$, range: 0.439-0.868). Nearly half (47.2%) of pairwise comparisons produced substantial agreement ($\kappa > 0.60$), with remaining comparisons achieving moderate agreement ($0.41 \leq \kappa \leq 0.60$). No comparisons fell below moderate agreement, indicating reasonable consistency—not poor—for LLM assessment regardless of model or prompt engineering technique differences.

Human-AI Reliability

Human-AI agreement demonstrated considerable variation across LLM and prompt engineering technique combinations, ranging from $\kappa = 0.399$ to $\kappa = 0.629$. Only ChatGPT-4.1 using Chain-of-Thought ($\kappa = 0.629$) and Few-shot ($\kappa = 0.602$) achieved substantial agreement with human raters. Other configurations attained moderate agreement ($\kappa = 0.543$ to $\kappa = 0.570$) or fair agreement, notably Gemini-Flash with Zero-shot ($\kappa = 0.399$), as shown in Table 18.

Table 18. Human-AI Reliability

LLM	Prompt Tech.	QWK	*Std Dev*
ChatGPT-4.1	Chain-of-Thought	0,629	0,802
ChatGPT-4.1	Few-shot	0,602	0,819
ChatGPT-4.1	Zero-shot	0,570	0,826
DeepSeek-Chat	Chain-of-Thought	0,533	0,866
Gemini-Flash	Chain-of-Thought	0,522	0,880
DeepSeek-Chat	Few-shot	0,495	0,904
Gemini-Flash	Few-shot	0,464	0,939
DeepSeek-Chat	Zero-shot	0,453	0,895
Gemini-Flash	Zero-shot	0,399	0,978

Model-level analysis—using different prompt techniques within the same LLM—revealed substantial differences. Average agreement within ChatGPT-4.1 ($M_\kappa = 0.600$) approached substantial agreement, while other models demonstrated fair agreement (DeepSeek-Chat: $M_\kappa = 0.494$; Gemini-Flash: $M_\kappa = 0.461$). ChatGPT-4.1's average exceeded Gemini-Flash by 0.139 points (30% improvement), indicating significant practical difference. Gemini-Flash's low average—as shown in Table 19—agreement suggests limited validity for response evaluation, reinforcing previous findings regarding LLM model effects.

Table 19. Human-AI Reliability by LLM Model

LLM	QWK	Std Dev
ChatGPT-4.1	0,600	0,815
DeepSeek-Chat	0,494	0,889
Gemini-Flash	0,461	0,932

Prompt strategy-level analysis—employing different LLMs with identical prompt techniques—indicated significant influence on Human-AI agreement. Chain-of-Thought's average reliability ($M_\kappa = 0.561$) exceeded Few-shot ($M_\kappa = 0.520$) by 9.7% and Zero-shot ($M_\kappa = 0.474$) by 18.4%, as presented in Table 20. These findings confirm prompt engineering technique effects, with Chain-of-Thought emerging as the most effective technique for aligning LLM assessments with human judgment.

Table 20. Reliability by Prompt Engineering Technique

Prompt Tech.	QWK	Std Dev
Chain-of-Thought	0,561	0,849
Few-shot	0,520	0,887
Zero-shot	0,474	0,890

LLM sensitivity to prompt engineering techniques varies considerably. ChatGPT-4.1 demonstrated stable performance with 10.5% variation ($\Delta\kappa = 0.059$) from lowest reliability using Zero-shot ($\kappa = 0.570$) to highest using Chain-of-Thought ($\kappa = 0.629$). DeepSeek-Chat showed higher variation (17.6%, $\Delta\kappa = 0.080$) between Zero-shot ($\kappa = 0.453$) and Chain-of-Thought ($\kappa = 0.533$). Gemini-Flash emerged as the most sensitive model with 30.8% variation ($\Delta\kappa = 0.123$) between Zero-shot ($\kappa = 0.399$) and Chain-of-Thought ($\kappa = 0.522$).

These findings reinforce previous ANOVA results, demonstrating that both LLM model and prompt engineering technique produce statistically significant effects that, despite small statistical effect sizes, generate meaningful practical impacts on assessment performance.

Human Rater Quality as Gold Standard

Comparison between average Inter-AI Agreement (between models) and Human-AI Agreement raised questions regarding internal validity. When considering all models, average Inter-AI Agreement consistently exceeded Human-AI Agreement, as shown in Table 21. This systematic pattern indicates that human raters potentially exhibit greater subjectivity compared to internal LLM consistency.

Table 21. Within-Model Agreement vs Human-AI Agreement

LLM	Within-model Agreement	Human-AI Agreement	Diff
ChatGPT-4.1	0,794	0,600	-0,194
DeepSeek-Chat	0,710	0,461	-0,249
Gemini-Flash	0,698	0,494	-0,204

Cross-model agreement analysis revealed that employing different LLMs with identical prompt engineering techniques still produced higher Inter-AI Agreement than Human-AI Agreement, as presented in Table 22. This suggests human raters potentially employ assessment patterns differing from the consensus achieved by LLMs—despite using different models—in interpreting scoring rubrics. The optimal Inter-AI Agreement configuration—ChatGPT-4.1 using Few-shot and Chain-of-Thought—achieved almost perfect agreement. Human-AI Agreement for the same configuration reached $\kappa = 0.602$ for Few-shot and $\kappa = 0.629$ for Chain-of-Thought, both classified as substantial agreement but lower than Inter-AI Agreement. These results clarify that human raters employ assessment patterns similar to LLMs while retaining inherent subjectivity in evaluation processes.

Table 22. Cross-Model Agreement vs Human-AI Agreement

Prompt Technique	Inter-AI Agreement	Human-AI Agreement	Diff
Chain-of-Thought	0,644	0,561	-0,083
Few-shot	0,569	0,520	-0,049
Zero-shot	0,573	0,474	-0,099

Discussion

Summary of Findings

This research demonstrates that interactions among independent variables collectively influence LLM assessment outcomes. Student proficiency group emerged as the variable with the highest impact, while prompt engineering technique demonstrated the smallest effect. LLMs exhibited higher accuracy when evaluating "Basic" and "Elementary and Intermediate" student responses, with performance declining significantly for Advanced and Proficient students. These findings may align with a previous study where LLMs struggled more on higher cognitive-level tasks (Murali et al., 2024).

Among tested models, ChatGPT-4.1 has the most similar result with human rater assessments, establishing it as the most accurate model in this study, as proven by the lowest mean absolute score differences from human evaluations. Prompt engineering techniques incorporating reasoning steps consistently improved accuracy, with Chain-of-Thought demonstrating the capacity to equalize

performance across models. This technique, in combination with ChatGPT-4.1, achieved optimal accuracy while also enhancing performance for Gemini-Flash.

Inter-AI within-model reliability testing revealed that ChatGPT-4.1 using Few-shot and Chain-of-Thought achieved the highest reliability, showing internal consistency. Inter-AI cross-model reliability supported previous findings that reasoning steps in prompts (Few-shot and Chain-of-Thought) enhance LLM output accuracy, reinforcing that ChatGPT-4.1 with Chain-of-Thought represents the most accurate configuration when measured by mean absolute score differences from human assessments.

Human-AI reliability varied considerably with ChatGPT-4.1 using Chain-of-Thought achieving substantial agreement ($\kappa = 0.629$) and Few-shot approaching this threshold ($\kappa = 0.602$), while Gemini-Flash demonstrated the lowest average agreement. These results are way lower than the previous study, where Human-AI reliability resides in 92-95% when tested using ChatGPT-3.5 and ChatGPT-4. That said, the previous study was able to compare Inter-Human reliability, which have 96% result, when the Inter-AI reliability only have 60% and 80% for ChatGPT-3.5 and ChatGPT-4.0, respectively (Tate et al., 2024). Another study that also uses Few-shot prompting with a simple rubric achieved reliability of 0.7 while having Inter-Human reliability of 0.75 (Henkel et al., 2024). While this study is unable to compare Inter-Human reliability, a systematic pattern emerged wherein Human-AI reliability consistently fell below Inter-AI agreement. Compared to combinations that achieved almost perfect inter-AI within-model reliability, Human-AI reliability for identical configurations reached only substantial agreement, indicating that humans may employ similar but more subjective assessment patterns compared to LLMs.

Theoretical Implications

LLM Reliability and Internal Consistency in Automated Essay Scoring

Inter-AI cross-model reliability findings reinforce theoretical claims that prompt engineering techniques integrating reasoning steps—such as Few-shot and Chain-of-Thought—enhance transparency in model cognitive processes (Wei et al., 2024). From a cognitive load theory perspective, providing explicit reasoning steps in prompts facilitates systematic information organization, thereby improving output consistency (Brown et al., 2020; Sweller et al., 2011). Inter-AI within-model reliability testing demonstrated that ChatGPT-4.1 with Chain-of-Thought and Few-shot exhibited the highest internal consistency, achieving almost perfect inter-rater reliability (Cohen, 1960; Landis & Koch, 1977). This high internal consistency indicates consistent output generation when identical inputs receive varied prompting, suggesting that LLM risks—bias, inaccuracy, and inconsistency—while present as evidenced by inter-rater reliability below perfect scores (Kasneci et al., 2023; Yan et al., 2024), can be mitigated through prompt engineering techniques that direct LLM reasoning processes (Brown et al., 2020).

Implementation Framework for LLM-Based Automated Essay Scoring

Results indicate that the optimal configuration—ChatGPT-4.1 with Chain-of-Thought prompting—achieved substantial agreement ($\kappa > 0.60$) with human raters, approaching but not meeting the threshold for automated essay scoring implementation. Williamson et al. (2012) established minimum inter-rater agreement of 0.70 for automated essay scoring systems suitable for high-stakes assessment contexts (Williamson et al., 2012). These findings demonstrate that LLMs cannot yet fully replace human raters in such contexts. The systematic pattern wherein Inter-AI agreement exceeds Human-AI agreement underscores human raters' essential role as validity anchors—referred to as gold standards in this research. Automated essay scoring should be viewed as complementary rather than substitutional to human evaluation in high-stakes assessment (Williamson et al., 2012). This research provides empirical evidence supporting hybrid scoring models wherein LLMs serve for preliminary assessment or as supplements to human evaluation.

Comparing Untrained LLM with Human Rater

This research compares LLMs receiving only rubric information and limited prior assessment examples (when employing Few-shot and Chain-of-Thought)—without training on human-scored response datasets—against human raters, reflecting a deliberate experimental design that prioritizes generalizable results over prediction accuracy for specific purposes. Conventional automated essay scoring systems are typically trained on human-scored datasets to optimize agreement, treating human scores as targets for learning (Williamson et al., 2012). However, this paradigm assumes the availability of training data that—in this case study context—proves unfeasible. This research design aligns more closely with Brown et al.'s (2020) perspective that LLMs possess task-agnostic capabilities enabling performance on novel tasks—in this research, evaluating diverse questions—without model updates or fine-tuning, relying purely on text-based interaction (prompts) with minimal reference examples (few-shot) (Brown et al., 2020).

This evaluation methodology reflects validity for resource-constrained educational environments where extensive training data remains unavailable. Systems trained on specific human rater assessment patterns also risk overfitting, potentially reproducing assessment biases (Williamson et al., 2012). In contrast, this research evaluates whether LLMs can apply reasoning based solely on explicit criteria (scoring rubrics)—a capability transferable across questions and institutions. This approach enables research findings to benefit institutions beyond the Language Learning Center case study context, supporting broader implementation without requiring institution-specific training.

Conclusion, Limitations, and Future Work

This research is an initial exploratory case study, and the conclusions derived from findings should be understood as preliminary indications requiring further verification through subsequent studies with more comprehensive designs and scales. In this study, interactions among independent variables significantly influence assessment outcomes. The student proficiency group has the greatest impact on accuracy, followed by LLM-prompt engineering technique combinations. LLMs exhibited high internal reliability in evaluating open-ended responses, demonstrating consistent capability in interpreting questions, scoring rubrics, and student answers, although they still cannot yet be fully relied upon to replace human raters entirely. This resulted in an optimal strategy suggested by the author for leveraging LLMs, which involves utilizing them as preliminary assessors or supplements to human evaluation. It can provide feedback based on the input, potentially helping students to improve their capabilities. This research shows that ChatGPT-4.1 with Chain-of-Thought prompting has the most accurate configuration for such applications.

While this study provides initial contributions in this matter, several limitations require attention. The quasi-experimental design was chosen due to the inability of random assignment. Research subjects consist of naturally formed groups within the academic context, limiting the implementation ability of findings to broader populations with more diverse conditions. The student proficiency groups represent a combination of two independent variables—student proficiency level and assigned open-ended questions—from the institutional practice of providing different questions to each proficiency level despite identical question types. This creates a confounding variable whose specific influence on outcomes remains undetermined. Inter-rater reliability (or agreement) among human raters was also not measured, limiting generalization of the conclusions. The systematic pattern wherein Human-AI agreement consistently falls below Inter-AI agreement suggests potentially low human-human agreement reliability. This possibility could result in different research conclusions. Without establishing human-human agreement baselines, the true ceiling for automated scoring performance remains unknown. LLM model selection was also constrained by the availability of non-reasoning models during the research design phase. The rapid evolution of LLM technology means findings specific to ChatGPT-4.1, Gemini-Flash, and DeepSeek-Chat may differ for subsequent model versions or alternative architectures. This research also focused only on English language courses within a university language learning center context, limiting generalizability to other disciplines or educational levels where assessment criteria and construct definitions differ substantially.

This research opens possibilities for subsequent studies to deepen understanding regarding LLM and prompt engineering technique utilization in evaluating open-ended question responses, particularly in English language learning contexts. However, it should conduct internal validation through inter-human reliability measurement to ensure that gold standard benchmarks possess high internal reliability. This can be achieved through paired assessment designs wherein each response receives evaluation from at least two human raters, establishing baseline human-human agreement necessary for valid automated scoring performance assessment. Prompt engineering techniques also emerged as a significant factor influencing LLM outputs. Since scoring rubrics directly constitute prompt components, more detailed scoring rubrics directly influence prompt effectiveness and assessment quality. Future research should also investigate human rater behavior when responses have undergone preliminary LLM assessment, examining whether human raters exhibit tendencies toward anchoring on initial LLM evaluations without conducting independent, thorough reviews. Such an investigation would inform optimal human-AI collaboration frameworks and identify potential biases introduced through hybrid scoring models. Iterative learning approaches should also be explored, where LLMs' outputs would undergo continuous human review, creating a feedback loop for progressive improvement. This may answer whether iteratively learned systems maintain generalizability to novel question types—essential for MOOC contexts with evolving curricula—and determine minimal human review levels enabling effective learning while preserving cost efficiency. Finally, expanding research scope to include diverse educational contexts, additional LLM models, alternative rubric designs, and varied question types would enhance generalizability and provide comprehensive guidelines for LLM implementation in educational assessment across disciplines and proficiency levels. (Brown et al., 2020)

References

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *Proceedings of the 34th International Conference on Neural Information Processing Systems, Nips '20*.

Cheng, X., Pan, C., Zhao, M., Li, D., Liu, F., Zhang, X., Zhang, X., & Liu, Y. (2025). Revisiting Chain-of-Thought Prompting: Zero-shot Can Be Stronger than Few-shot. *Findings of the Association for Computational Linguistics: EMNLP 2025*, 13533–13554. https://doi.org/10.18653/v1/2025.findings-emnlp.729

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213–220. https://doi.org/10.1037/h0026256

Cohen, J. (2013). *Statistical Power Analysis for the Behavioral Sciences* (0 ed.). Routledge. https://doi.org/10.4324/9780203771587

Cook, T. D., Campbell, D. T., & Shadish, W. (2002). *Experimental and quasi-experimental designs for generalized causal inference* (Vol. 1195). Houghton Mifflin Boston, MA.

Creswell, J. W. (2017). *Research design. Qualitative, quantitative, and mixed methods approaches* (5th edition (international student edition)). SAGE Publications.

Fisher, R. A. (1925). *Statistical methods for research workers, 11th ed. Rev.* Edinburgh.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378–382. https://doi.org/10.1037/h0031619

Google. (2025). Gemini for Student. *Gemini for Student*. https://gemini.google/students/

Haefner, N., Wincent, J., Parida, V., & Gassmann, O. (2021). Artificial intelligence and innovation management: A review, framework, and research agenda☆. *Technological Forecasting and Social Change*, *162*, 120392. https://doi.org/10.1016/j.techfore.2020.120392

Henkel, O., Hills, L., Boxer, A., Roberts, B., & Levonian, Z. (2024). Can Large Language Models Make the Grade? An Empirical Study Evaluating LLMs Ability To Mark Short Answer Questions in K-12 Education. *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, 300–304. https://doi.org/10.1145/3657604.3664693

Hodgkinson-Williams, C. (2014). *Degrees of ease: Adoption of OER, open textbooks and MOOCs in the Global South*. OpenUCT. http://hdl.handle.net/11427/1188

Howell, D. C. (2007). *Statistical methods for psychology*. Thomson Wadsworth. https://books.google.co.id/books?id=-bmMPwAACAAJ

Isaacs, T., Zara, C., Herbert, G., Coombs, S., & Smith, C. (2013). *Key Concepts in Educational Assessment*. SAGE Publications Ltd. https://doi.org/10.4135/9781473915077

J. Kay, P. Reimann, E. Diebold, & B. Kummerfeld. (2013). MOOCs: So Many Learners, So Much Potential ... *IEEE Intelligent Systems*, *28*(3), 70–77. https://doi.org/10.1109/MIS.2013.66

Jalil, S., Rafi, S., LaToza, T. D., Moran, K., & Lam, W. (2023). ChatGPT and Software Testing Education: Promises & Perils. *2023 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*, 4130–4137. https://doi.org/10.1109/ICSTW58534.2023.00078

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kvålseth, T. O. (2018). An Alternative Interpretation of the Linearly Weighted Kappa Coefficients for Ordinal Data. *Psychometrika*, *83*(3), 618–627. https://doi.org/10.1007/s11336-018-9621-1

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159. https://doi.org/10.2307/2529310

Leckie, G., & Baird, J.-A. (2011). Rater Effects on Essay Scoring: A Multilevel Analysis of Severity Drift, Central Tendency, and Rater Experience. *Journal of Educational Measurement*, *48*, 399–418. https://doi.org/10.2307/41427532

Lee, G.-G., Latif, E., Wu, X., Liu, N., & Zhai, X. (2024). Applying large language models and chain-of-thought for automatic scoring. In *Computers and Education: Artificial Intelligence* (Vol. 6). Elsevier B.V. https://doi.org/10.1016/j.caeai.2024.100213

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks* (Version 4). arXiv. https://doi.org/10.48550/ARXIV.2005.11401

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns*, *4*(7), 100779. https://doi.org/10.1016/j.patter.2023.100779

Liu, X., Wang, J., Sun, J., Yuan, X., Dong, G., Di, P., Wang, W., & Wang, D. (2023). *Prompting Frameworks for Large Language Models: A Survey* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2311.12785

Mendonça, N. C. (2024). Evaluating ChatGPT-4 Vision on Brazil's National Undergraduate Computer Science Exam. *ACM Trans. Comput. Educ.*, *24*(3), 37:1-37:56. https://doi.org/10.1145/3674149

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., & Zettlemoyer, L. (2022). Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 11048–11064. https://doi.org/10.18653/v1/2022.emnlp-main.759

Murali, R., Dhanalakshmy, D. M., Avudaiappan, V., & Sivakumar, G. (2024). Towards Assessing the Credibility of Chatbot Responses for Technical Assessments in Higher Education. *2024 IEEE Global Engineering Education Conference (EDUCON)*, 1–9. https://doi.org/10.1109/EDUCON60312.2024.10578934

Oğuz, E. (2025). Can generative AI figure out figurative language? The influence of idioms on essay scoring by ChatGPT, Gemini, and Deepseek. *Assessing Writing*, *66*, 100981. https://doi.org/10.1016/j.asw.2025.100981

OpenAI Team. (2024, September 14). *How ChatGPT and our language models are developed*. OpenAI Help Center. https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed

OpenAI Team. (2025a). Prompt Engineering Guide. *Prompt Engineering Guide*. https://platform.openai.com/docs/guides/prompt-engineering

OpenAI Team. (2025b). Reasoning Best Practice. *Reasoning Best Practice*. https://platform.openai.com/docs/guides/reasoning-best-practices

Poole-Dayan, E., Roy, D., & Kabbara, J. (2024). *LLM Targeted Underperformance Disproportionately Impacts Vulnerable Users* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2406.17737

Rahman, A., Mahir, S. H., Tashrif, Md. T. A., Karim, Md. A., Aishi, A. A., Kundu, D., Debnath, T., Moududi, Md. A. A., Eidmum, Md. Z. A., Miah, A. S. M., Farid, F. A., & Karim, H. A. (2025). Comparative analysis based on DeepSeek, ChatGPT, and Google Gemini: Features, techniques, performance, future prospects. *Systems and Soft Computing*, *7*, 200396. https://doi.org/10.1016/j.sasc.2025.200396

Rau, G., & Shih, Y.-S. (2021). Evaluation of Cohen's kappa and other measures of inter-rater agreement for genre analysis and other nominal data. *Journal of English for Academic Purposes*, *53*, 101026. https://doi.org/10.1016/j.jeap.2021.101026

Şahan, Ö. (2018). *The impact of rater experience and essay quality on rater behavior and scoring* [Phd].

Sarim, M., Masood, F., Maheshwari, M., Faridi, A. R., & Shamsan, A. H. (2025). Generating reliable software project task flows using large language models through prompt engineering and robust evaluation. *Scientific Reports*, *15*(1), 35194. https://doi.org/10.1038/s41598-025-19170-9

SGU. (2025, February 7). A Comparison of Leading AI Models: DeepSeek AI, ChatGPT, Gemini, and Perplexity AI. *A Comparison of Leading AI Models: DeepSeek AI, ChatGPT, Gemini, and Perplexity AI*. https://sgu.ac.id/a-comparison-of-leading-ai-models-deepseek-ai-chatgpt-gemini-and-perplexity-ai/

Shi, L., Cristea, A., Toda, A., & Oliveira, W. (2020, August 12). *Revealing the Hidden Patterns: A Comparative Study on Profiling Subpopulations of MOOC Students*. https://doi.org/10.48550/arXiv.2008.05850

Siska Merrydian, Indah Mustika Rini, Wardani Rahayu, & Riyadi. (2024). Systematic Literature Review: Validitas Internal dan Eksternal dalam Desain Ekperimen. *EduInovasi: Journal of Basic Educational Studies*, *4*(3), 987–1001. https://doi.org/10.47467/edu.v4i3.2734

Sivarajkumar, S., Kelley, M., Samolyk-Mazzanti, A., Visweswaran, S., & Wang, Y. (2024). An Empirical Evaluation of Prompting Strategies for Large Language Models in Zero-Shot Clinical Natural Language Processing: Algorithm Development and Validation Study. *JMIR Medical Informatics*, *12*, e55318. https://doi.org/10.2196/55318

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A., … Wu, Z. (2023). *Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models* (arXiv:2206.04615). arXiv. https://doi.org/10.48550/arXiv.2206.04615

Stasuik, N. C. (2025). *Evaluating LLM Performance in Essay Assessment: A Comparative Analysis of AI Grading and Feedback Systems for University English Courses*. https://doi.org/10.14288/1.0448868

StatCounter. (2025). AI Chatbot Market Share. *AI Chatbot Market Share*. https://gs.statcounter.com/ai-chatbot-market-share

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory*. Springer.

Tang, H., & Qian, Y. (2022). Designing MOOCs with LITTLE. *Cogent Education*, *9*(1), 2064411. https://doi.org/10.1080/2331186X.2022.2064411

Tate, T. P., Steiss, J., Bailey, D., Graham, S., Moon, Y., Ritchie, D., Tseng, W., & Warschauer, M. (2024). Can AI provide useful holistic essay scoring? *Computers and Education: Artificial Intelligence*, *7*, 100255. https://doi.org/10.1016/j.caeai.2024.100255

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). *LLaMA: Open and Efficient Foundation Language Models* (arXiv:2302.13971). arXiv. https://doi.org/10.48550/arXiv.2302.13971

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., & Le, Q. V. (2022). *Finetuned Language Models Are Zero-Shot Learners* (arXiv:2109.01652). arXiv. http://arxiv.org/abs/2109.01652

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2024). Chain-of-thought prompting elicits reasoning in large language models. *Proceedings of the 36th International Conference on Neural Information Processing Systems, Nips '22*.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A Framework for Evaluation and Use of Automated Scoring. *Educational Measurement: Issues and Practice*, *31*(1), 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

Wu, Y., Wang, Y., Ye, Z., Du, T., Jegelka, S., & Wang, Y. (2025). *When More is Less: Understanding Chain-of-Thought Length in LLMs* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2502.07266

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, *55*(1), 90–112. https://doi.org/10.1111/bjet.13370

Yu, H., Miao, C., Leung, C., & White, T. J. (2017). Towards AI-powered personalization in MOOC learning. *Npj Science of Learning*, *2*(1), 1–5. https://doi.org/10.1038/s41539-017-0016-3

Zhao, H., Andersson, B., Guo, B., & Xin, T. (2017). Sequential Effects in Essay Ratings: Evidence of Assimilation Effects Using Cross-Classified Models. *Frontiers in Psychology*, *8*, 933. https://doi.org/10.3389/fpsyg.2017.00933

Zhong, Y., Hao, J., Fauss, M., Li, C., & Wang, Y. (2025). *AI-generated Essays: Characteristics and Implications on Automated Scoring and Academic Integrity* (arXiv:2410.17439). arXiv. https://doi.org/10.48550/arXiv.2410.17439

How to cite:

Wilmar, F. & & Putra, P. O. H. (2026). Evaluation of Open-ended Question's Answers using Large Language Model (LLM): A Case Study of a Language Learning Center in University. *Jurnal Sistem Informasi (Journal of Information System)*, *22*(1), 48–75.

Appendices

Appendix A
Java Console Repository

The automated assessment application utilizing LLMs is available in the following repository: https://github.com/faisalwilmar/generative-ai-scoring. The repository includes references and sample datasets, enabling replication and adaptation for future research applications.

Appendix B
Prompt Template

Template	Prompt
System Message	You are a teacher for English Course, currently grading a writing assessment. The question you give to your students is: {{question}}. You have scoring guide as follows: {{scoring guide}}. Score the students' answer and give your reason as feedback. Return your final evaluation strictly as valid JSON with exactly two keys: 'llm_grade' (an integer for the score) and 'llm_feedback' (a string for your feedback). Do not include any extra text or your reasoning process.
User Message Zero-shot	Score this student's answer: {{student answer}}
User Message Few-shot	Here is a student's answer: {{example answer}}. That answer got {{example point}} point.
User Message Chain-of-Thought	Here is a student's answer: {{example answer}}. That answer got {{example point}} point because of the following reasons: {{example reason}}.

Appendix C
Example Question

This example question is from Advanced Proficient Proficiency level, *Responding to a text* with the following instruction:

Directions: Respond to the e-mail as if you were Anna Kyarsi. In your e-mail, ask for TWO pieces of information.

From : Kay Poan

To : Anna Kyarsi

Subject : How to answer

Sent : February 22, 4:32 P.M.

I know you always score high on TOEIC tests, so I'm reaching out to ask for a favor. In the Reading section – especially in the Incomplete Sentences and Incomplete Texts – I often feel frustrated because I don't know how to approach the questions. Each one requires extensive grammar knowledge, and I struggle to make an intelligent guess.

Could you share some tips or strategies that have worked for you? I'd really appreciate your help!

Thank you.

Appendix D
Scoring Rubric

Score	Description of a Typical Test taker's Response
4	The response effectively addresses all the tasks in the prompt using multiple sentences that clearly convey the information, instructions, questions, etc., required by the prompt. The response uses organizational logic or appropriate connecting words or both to create coherence among sentences. The one and register of the response is appropriate for the intended audience. A few isolated errors in grammar or usage may be present but do not obscure the writer's meaning.
3	The response is mostly successful but falls short in addressing one of the tasks required by the prompt. The response may omit, respond unsuccessfully, or respond incompletely to ONE of the required tasks. The response uses organizational logic or appropriate connecting words in at least part of the response. The response shows some awareness of audience. Noticeable errors in grammar or usage may be present; ONE sentence may contain errors that obscure meaning.
2	The response is marked by several weaknesses. The response may address only ONE of the required tasks or may unsuccessfully or incomplete address TWO or THREE of the required tasks. Connections between ideas may be missing or obscure. The response may show little awareness of audience. Errors in grammar and usage may obscure meaning in MORE THAN ONE sentence.
1	The response is seriously flawed and conveys little or no information, instructions, questions, etc., required by the prompt. The response addresses NONE of the required tasks, although it may include some content relevant to the stimulus. Connections between ideas are missing or obscure. The tone or register may be inappropriate for the audience. Frequent errors in grammar or usage obscure the writer's meaning most of the time.
0	A response at this level merely copies words from the prompt or stimulus, rejects the topic or is otherwise not connected to the topic, is written in a language other than English, consist of keystrokes characters that convey no meaning, or is blank.

Appendix E
Example Reasoning Guide

Answer	Score	Score Reasoning
From : Anna Kyarsi To : Kay Poan Subject : Tips to answer Thank you for trusting me to ask a favor, Kay. But first, do you have other struggle like short-attention span or else? Because other factor such as thing can affect the test. For this, we can meet up to discuss it more. When do you have time? please let me know, my time is flexible. As for my tips and strategies in the reading section, all you need to do is focusing on certain keywords and expand your vocabulary. For incomplete texts or sentences section, I usually focused on the answer sheet	4	Your response effectively addresses all parts of the prompt. You provide clear tips and strategies for the TOEIC Reading section, and you ask for two pieces of information: whether Kay has other struggles (like short attention span) and when Kay is available to meet. The email is organized, uses appropriate connecting words, and maintains a suitable tone for the audience. There are a few minor grammar errors, but they do not obscure your meaning.

Answer	Score	Score Reasoning
first to get to know which one is more suitable for the texts or sentences. Remember, find the keywords.		
FROM : Anna Kyarsi TO : Kay Poan SUBJECT : How To Answer Hello Kay Poan, I appreciate it. First of all, I felt your struggle. Reading English book is the key. What you have to do to the Incomplete Sentences is memory English vocabulary. If you understand some word, you can understand the written. Also for the Incomplete Texts, the important thing I do is I read English book even though I dont understand. It helps your feeling about the true or false word to be add to the sentence. I hope this could help you. Please, contact me if you need a private class with me.	2	Your response addresses the request for tips and strategies, but it does not ask for two pieces of information as required by the prompt. The organization is somewhat logical, but there are several grammar and usage errors that make some sentences difficult to understand. Additionally, the response does not show clear awareness of the audience, and the tone is not fully appropriate for an email reply. To improve, make sure to ask for two pieces of information from Kay and use clearer, more accurate English.

Appendix F
Python Notebook Repository

The Python implementation analysis notebook is available in the following repository: https://gist.github.com/faisalwilmar/f67fbd1b45b05e57d1d56ec24cafb3b7.