The Eurasia Proceedings of
Educational and Social Sciences (EPESS)
ISSN: 2587-1730

The Eurasia Proceedings of Educational and Social Sciences (EPESS), 2026

Volume 49, Pages 228-244

ICRET 2026: International Conference on Research in Education and Technology

Guidelines For Automatic Grading of Student Essays Using Large Language Models

Diwakar Yalpi
Old Dominion University

Sruta Keerti Kasula
Old Dominion University

Ravi Mukkamala
Old Dominion University

Abstract: Automated essay evaluation using large language models (LLMs) has emerged as a promising approach to support scalable and consistent educational assessment. However, the effectiveness of LLM-based grading varies significantly across evaluation dimensions and is highly influenced by prompt design and model selection. In this study, we evaluate five state-of-the-art LLMs across five rubric-based categories: Relevance to Question, Reasoning and Critical Thinking, Evidence and Examples, Organization, and Clarity and Writing Quality. We systematically investigate the impact of three prompting strategies, including rubric-only prompting, exemplar-based prompting (with and without rubric guidance)(Original and Refined prompt designs) incorporating structured instructions. Additionally, a prompt ablation study is conducted to analyze the effect of instruction detail and reasoning guidance on grading accuracy. Our results demonstrate that no single model consistently outperforms others across all categories. Instead, each LLM exhibits strengths in specific evaluation dimensions, motivating a category-wise model selection approach. Furthermore, refined and structured prompting strategies significantly improve evaluation performance, with chain-of-thought-style instructions yielding the highest accuracy. These findings highlight the importance of both prompt engineering and task-specialized model allocation in developing robust LLM-based grading systems. The study provides evidence supporting multi-agent frameworks, where different LLMs can be assigned to distinct evaluation tasks to enhance overall grading reliability and alignment with human judgments.

Keywords: Higher education, Critical-thinking, Automated grading, Large language models, Task distribution, Ensemble LLM

Introduction

In recent years, higher education has placed increasing emphasis on writing across disciplines, significantly expanding the demands associated with evaluating student essays and providing timely, high-quality feedback (Nickelson and Schaf-fer, 2025). This shift has intensified the instructional burden on educators, particularly in large-scale or writing-intensive courses, where consistent and detailed assessment is both essential and resource-intensive.

Critical thinking represents a fundamental learning outcome in higher education and is commonly evaluated through student's analytical essays (Golden, B. 2023). In conventional assessment practices, expert evaluators typically rely on analytic rubrics that focus on the quality of reasoning, the adequacy and relevance of supporting evidence, the coherence of argumentation, and the extent to which responses address the assigned prompt.

- This is an Open Access article distributed under the terms of the Creative Commons Attribution-Noncommercial 4.0 Unported License, permitting all non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

- Selection and peer-review under responsibility of the Organizing Committee of the Conference

© 2026 Published by ISRES Publishing: www.isres.org

The emergence of Large Language Models (LLMs) offers new possibilities for supporting essay evaluation through their advanced natural language processing and generative capabilities. These systems have demonstrated strong performance in a variety of language understanding and generation tasks, suggesting their potential to assist in scoring, justification, and feedback generation. However, the application of LLMs in educational assessment remains challenging. Issues such as hallucination, output inconsistency, and sensitivity to prompt design raise significant concerns regarding the reliability, validity, and fairness of automated or semi-automated grading systems (Huang et al., 2025).

In collaborative grading scenarios, where multiple evaluators independently assess the same essay, large language models (LLMs) can be leveraged to enhance quality assurance processes. In particular, LLMs can assist in detecting discrepant or outlier scores, aggregating evaluative feedback, and producing consensus-oriented summaries (Yalpi et al., 2026). These capabilities can be integrated while preserving rubric fidelity and ensuring consistent application across different graders. Consequently, LLMs offer the potential to improve moderation efficiency and reduce inconsistencies in scoring decisions. Moreover, the interpretability and chain-of-reasoning capabilities of LLMs can improve assessment transparency by generating structured, coherent, and concise feedback (Wang, X et al., 2023). Such explanations can help clarify the relationship between rubric criteria and assigned evaluations, thereby strengthening both the accountability of grading processes and the formative value of feedback provided to students.

However, evaluating student responses is a multifaceted task that involves multiple dimensions such as relevance to the question, reasoning and critical thinking, use of evidence and examples, organization of ideas, and clarity of writing. Traditional single-model approaches often struggle to consistently capture these diverse aspects of assessment. Furthermore, the performance of LLMs is highly sensitive to prompt design, instruction clarity, and contextual guidance, which can significantly influence grading outcomes.

In this work, we investigate the performance of five different LLMs across multiple rubric-based evaluation categories using a structured experimental framework. We explore three distinct prompting strategies: rubric-only prompting, 3 exemplar-based prompting (with and without rubric guidance) (original and refined prompt design) incorporating improved instruction clarity. Additionally, we conduct a prompt ablation study to analyze the impact of prompt structure on evaluation accuracy.

Our experiments further examine category-wise variability in model performance, revealing that different LLMs excel in different evaluation dimensions. This motivates a category-specific model selection approach and supports the use of multi-agent frameworks, where specialized LLMs are assigned to different grading subtasks to improve overall evaluation accuracy and robustness.

The contributions of this paper are as follows: (i) a comparative analysis of multiple LLMs across five rubric-based evaluation categories, (ii) an investigation of the impact of different prompting strategies on grading performance, and (iii) evidence supporting category-wise LLM selection for improved automated essay evaluation systems. The primary contributions of this work are as follows:

- **Modular LLM-Based Assessment Architecture:** We present a modular system design that explicitly separates scoring, rationale generation, and feedback synthesis components. This decomposition enables improved interpretability, targeted validation, and flexible integration of heterogeneous models within the assessment pipeline.

- **Assessment Preparation and Calibration Protocols:** We formalize essential preparatory procedures, including rubric specification, anchor-based calibration, and systematic bias evaluation. These steps ensure consistent interpretation of grading criteria and promote fairness across diverse student responses.

- **Robust Prompt Engineering Strategies:** We develop structured prompt design techniques aimed at enhancing output stability, transparency, and alignment with predefined rubrics. These strategies incorporate standardized templates, explicit criteria grounding, and controlled output formats to reduce variability and improve reproducibility.

- **Comprehensive Evaluation Framework:** We introduce evaluation protocols to rigorously assess system performance, including measures of reliability, agreement with human raters, and robustness across prompt variations and model versions. This framework supports systematic validation of LLM-assisted grading systems under realistic deployment conditions.

Method

This paper employs a *Qualitative Content Analysis (QCA)* approach to systematically evaluate student written responses in analytical essay tasks designed to assess critical thinking. The primary objective is to map textual evidence in student responses to predefined analytic criteria, enabling structured and transparent assessment of reasoning quality, evidence use, argument coherence, and relevance to the prompt.

A deductive coding framework is adopted to guide the analysis. In contrast to inductive approaches where categories emerge from the data, deductive coding in this work is driven by a predefined 0–5 analytic rubric and associated exemplar responses. The rubric operationalizes critical thinking into measurable dimensions, including (i) clarity and logical structure of arguments, (ii) appropriateness and sufficiency of supporting evidence, (iii) depth and relevance of reasoning, and (iv) alignment with the essay prompt. Each student response is coded by systematically matching segments of text to these rubric-defined performance levels.

The scoring process is implemented within a criterion-referenced assessment paradigm, where each response is evaluated against fixed performance standards rather than relative comparisons among peers. This ensures that evaluation reflects absolute levels of mastery for each rubric dimension, supporting fairness and interpretability in scoring outcomes.

To enhance consistency and reliability, the methodology incorporates an expert-calibrated scoring protocol. Evaluators either human raters or model-assisted graders are trained using anchor responses that represent each level of the rubric. These anchor examples serve as reference points to standardize interpretation of scoring criteria across evaluators and reduce subjective variability. Calibration sessions are used to align evaluators' judgments prior to formal scoring, ensuring consistent application of the rubric across all responses. In addition, the analysis process supports structured interpretive validation, whereby evaluators justify assigned scores with explicit references to textual evidence from student responses. This step strengthens transparency by linking each score to observable linguistic and argumentative features, rather than implicit judgment alone.

To ensure methodological rigor, the framework emphasizes consistency and reliability checks, including cross-evaluation of selected responses and comparison of scoring patterns across evaluators. This enables identification of scoring drift, inconsistencies, or systematic biases in rubric interpretation. Where applicable, discrepancies are resolved through moderation procedures or consensus discussion.

Overall, this methodological design integrates qualitative interpretive analysis with structured, rubric-driven evaluation principles. The combination of deductive coding, criterion-referenced assessment, and calibration-based scoring ensures that assessments are systematic, reproducible, and aligned with established standards in educational measurement research.

Experimental Setup

This paper investigates a decomposition-based approach to automated essay assessment, in which evaluation tasks are distributed across an ensemble of large language models (LLMs) according to explicitly defined critical-thinking criteria. The primary objective is to examine whether different LLMs exhibit differential strengths across rubric dimensions and to identify which models are most effective for specific aspects of evaluation. To this end, we employ five heterogeneous LLMs, namely ChatGPT, Qwen, Gemini, Llama, and Mistral, as independent evaluators within the proposed framework. Each model is assigned to assess student responses along predefined rubric categories related to critical thinking, including reasoning quality, evidence usage, argument coherence, and relevance to the prompt. This decomposition enables a fine-grained analysis of model-specific performance across distinct evaluative dimensions.

The dataset consists of student responses to the essay prompt: "What are the signs of cyberbullying?" These responses serve as the input corpus for evaluating the effectiveness and consistency of the multi-model assessment framework. By analyzing model outputs across rubric-aligned dimensions, the study aims to characterize the relative strengths and limitations of each LLM in educational assessment tasks and to inform the design of more reliable ensemble-based grading systems. The experimental setup is shown in Figure 1.

The experimental setup assumes a class size of 100 students, with student performance modeled using a normal distribution centered at a mean score of 3.5. To ensure consistency in response length and controllability of the evaluation task, each student response is constrained to a maximum of 125 words. For synthetic data generation,

a large language model (Sonnet) is prompted with a structured instruction set that specifies its role, the input requirements, task constraints, and expected outputs. The prompt explicitly defines the LLM's function within the system, including adherence to rubric-based evaluation principles and compliance with output formatting constraints.

In addition, the evaluation rubric (Figure 2) is provided directly to the synthetic data generation model to ensure alignment between generated responses and predefined assessment criteria. This design enables controlled generation of student-like responses while maintaining consistency with the scoring framework used in subsequent evaluation stages. In prior work utilizing different architectures, we observed that the evaluation LLM frequently exhibited difficulty in distinguishing between adjacent score bands, particularly grades 3 and 4. This indicated ambiguity in the rubric interpretation and suggested a need for improved exemplar design and prompt specification.

To address this limitation, we developed a refined version of the evaluation exemplars where we provide 3 exemplars for each grade level for each category. In this revision, the characteristics of each score band were defined more explicitly to reduce overlap between adjacent levels. The associated "why this fits" explanations were redesigned to be more diagnostic and instructionally oriented, thereby improving their utility for both calibration and evaluation. In addition, the language used in the exemplars was tightened to minimize ambiguity and reduce interpretive drift across scoring levels. Special attention was given to ensuring that exemplars for lower score bands were aligned with the revised standards, thereby improving overall grading consistency. We refer to the two configurations as the "Original" and "Refined" versions. Results from both settings are reported to highlight the impact of prompt and exemplar engineering on the stability and discriminative capability of LLM-based evaluation systems.

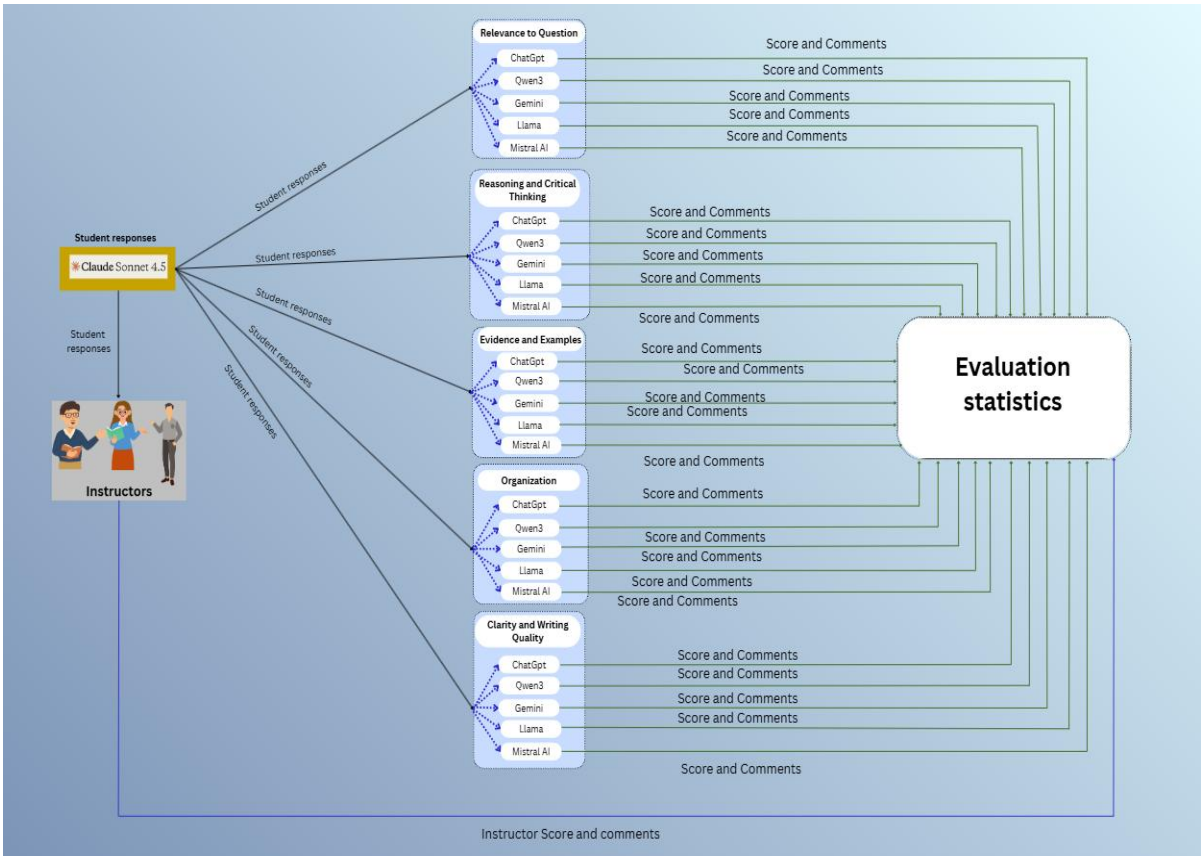

Figure 1. Experimental setup

Rubric

An analysis of the experimental setup provides a comprehensive understanding of the system's architectural design (Figure 1). There are four main components in the proposed architecture.

Category	Excellent [5]	Proficient[4]	Developing [3]	Emerging [2]	Beginning [1]	No Evidence [0]
Relevance to the Question	Fully understands and addresses all parts of the question with nuance and accuracy.	Addresses the question clearly with minor gaps.	Basic understanding of the question; may miss or misinterpret elements.	Limited understanding. Addresses the question superficially.	Minimal understanding; mostly off-topic.	Does not address the question. Completely off-topic
Reasoning and Critical Thinking	Insightful, logical, and well-developed analysis; strong critical thinking	Mostly logical reasoning with some depth.	Reasoning is present but uneven or superficial	Weak reasoning; ideas lack development.	Little to no reasoning; mostly unsupported claims.	No analysis or reasoning
Evidence and Examples	Evidence is relevant, accurate and well-integrated; clearly supports claims.	Evidence is appropriate but may lack depth or explanation	Some evidence was provided but weakly connected.	Minimal or unclear evidence.	Evidence is irrelevant or incorrect.	No evidence provided.
Organization	Clear structure, strong flow, effective transitions.	Generally well-organized with minor issues	Some organizational structure but inconsistent.	Poor organization; ideas feel disconnected.	Very disorganized; difficult to follow.	No coherent structure.
Clarity & Writing Quality	Clear, concise, and polished writing with no errors.	Mostly clear writing; minor errors do not impede meaning.	Writing is understandable but contains noticeable errors	Frequent errors that interfere with clarity.	Writing is unclear or error ridden.	Writing is incomprehensible.

Figure 2. Rubric

A. Data Generation

Student response data were synthetically generated using Claude Sonnet to ensure controlled variation in response quality while maintaining alignment with realistic educational settings. The use of synthetic generation enabled systematic control over response length, complexity, and distributional characteristics, which is particularly important for evaluating rubric-based assessment systems under consistent conditions. Given that the study focuses on critical thinking evaluation, the assessment rubric was designed to capture multiple complementary dimensions of analytical writing quality.

The rubric was operationalized into five categories: (i) Relevance to Question, (ii) Reasoning and Critical Thinking, (iii) Evidence and Examples, (iv) Organization, and (v) Clarity and Writing Quality. These categories collectively represent key constructs of critical thinking as manifested in written discourse, ensuring that evaluation spans both content-level reasoning and structural writing quality. To ensure consistency in generation,

the LLM was prompted with explicit constraints regarding response length, task framing, and adherence to rubric-aligned expectations. This design allowed for controlled variability across responses while preserving meaningful differences in quality levels.

B. Human Evaluation

To establish a reliable evaluation benchmark, multiple domain-experienced instructors independently evaluated the generated student responses. Each instructor evaluated responses using the predefined analytic rubric and assigned category-specific scores on a standardized scale, accompanied by qualitative justifications grounded in rubric descriptors. These instructor-generated evaluations were treated as the ground truth reference standard for subsequent model comparison. The use of multiple instructors enabled mitigation of individual bias and improved the robustness of the reference labels.

All evaluators completed the evaluation process independently, without access to other evaluator's scores or system outputs, thereby ensuring independence and reducing potential anchoring effects. Given their extensive experience in essay grading and educational assessment, instructor judgments are considered a high-fidelity approximation of expert human scoring. This human baseline serves as the primary standard against which automated and LLM-based scoring systems are evaluated.

C. Ensemble-Based LLM Evaluation Framework

The proposed system adopts a decomposition-driven ensemble architecture, wherein the essay evaluation task is partitioned across multiple large language models (LLMs) according to rubric dimensions. Each rubric category is assigned a dedicated LLM ensemble to enable specialization in distinct aspects of critical thinking assessment. Since the rubric consists of five categories, the system employs five parallel evaluation pipelines, each corresponding to one rubric dimension as shown in Figure 1. This design allows each LLM to focus on a narrower cognitive and evaluative task, thereby reducing cognitive overload and improving interpretability of model outputs.

Each LLM is prompted to produce a structured output consisting of: (i) a score on a scale of 0–5, (ii) a justification explaining the rationale behind the score with explicit reference to rubric criteria, and (iii) a self-reported confidence score indicating uncertainty in the evaluation. This multi-output formulation enables both quantitative comparison with human scores and qualitative inspection of reasoning consistency. By structuring outputs in this way, the framework supports transparency in decision-making and facilitates fine-grained diagnostic analysis of model behavior across different evaluation dimensions.

D. Evaluation Procedure and Statistical Analysis

Following evaluation and model inference, a category-wise comparative evaluation was conducted to assess alignment between LLM-generated scores and human instructor judgments. This evaluation was performed independently for each of the five rubric categories, enabling fine-grained analysis of model strengths and weaknesses across distinct dimensions of critical thinking. To quantify performance, LLM outputs were systematically compared against ground truth annotations using category-level aggregation. This allowed for identification of systematic deviations, bias patterns, and dimension-specific performance differences across models.

All results were consolidated into an Evaluation Statistics module, which served as a centralized analytical layer for cross-model and cross-category comparison. In addition to numerical comparisons, visual analytics were employed to examine score distributions, alignment trends, and discrepancies between human and model evaluations.

The complete set of outputs from all participating LLMs was further aggregated into a unified performance profile, enabling holistic comparison across models and rubric dimensions. This structured analysis facilitates interpretation of which LLMs are most effective for specific components of critical thinking assessment and provides insights into the behavior of ensemble-based evaluation systems.

You are a calibrated essay scorer. Your job is to evaluate a student's response to the question "How can individuals identify signs of cyberbullying?" and each student's response is limited to 125 words. You will receive one category (with score levels, Rubric and grading notes) and student responses follows.

1. Grading Instructions:
Step 1: Read the given student response.
Step 2: Use the Rubric to assign the score.
Step 3: Focus only on the specific evaluation category you have been assigned.
Step 3: Assign a score from 0–5 based solely on this category.
Step 4: Provide a clear explanation to the student describing the strengths and weaknesses (if any) of their response to justify the score.
Step 5: Provide a confidence percentage indicating how certain you are about the accuracy of your scoring

Rubric for Relevance to Question:

Category Excellent [5] Proficient[4] Developing [3] Emerging [2] Beginning [1] No Evidence [0]
Relevance to the Question Fully understands and addresses all parts of the question with nuance and accuracy. Addresses the question clearly with minor gaps. Basic understanding of the question; may miss or misinterpret elements. "Limited understanding. Addresses the question superficially." Minimal understanding; mostly off-topic. Does not address the question. Completely off-topic

Use these ranges:
0–40% = Low confidence (uncertain or borderline)
41–70% = Medium confidence (some ambiguity)
71–100% = High confidence (score clearly fits the rubric).

2. Guidelines for Grading:
- Use the rubric to assign 0-5 grades.
- Follow the scoring notes provided for the category.
- Do not evaluate anything outside the assigned category (e.g., grammar, style, organization) unless the category explicitly includes it.

Figure 3. Prompt for strategy (i) - Only rubric provided to LLMs

You are a calibrated essay scorer. Your job is to evaluate a student's response to the question "How can individuals identify signs of cyberbullying?" and each student's response is limited to 125 words. You will receive one category (with score levels, exemplars, and grading notes) and student responses follows.

1. Grading Instructions:
Step 1: Read the given student response.
Step 2: Focus only on the specific evaluation category you have been assigned.
Step 3: Use the two provided Exemplars for this category along with the scoring guidelines.
Step 4: Assign a score from 0–5 based solely on this category.
Step 5: Provide a clear explanation to the student describing the strengths and weaknesses (if any) of their response to justify the score.
Step 6: Provide a confidence percentage indicating how certain you are about the accuracy of your scoring
Use these ranges:
0–40% = Low confidence (uncertain or borderline)
41–70% = Medium confidence (some ambiguity)
71–100% = High confidence (score clearly fits the rubric).

2. Guidelines for Grading the Exemplars:
- Use the exemplars to understand what performance looks like at different score levels.
- Compare the student response to the exemplar patterns, not to your personal expectations.
- Follow the scoring notes provided for the category.
- Do not evaluate anything outside the assigned category (e.g., grammar, style, organization) unless the category explicitly includes it.

3. Explanation of Why Each Exemplar Fits Its Category:
For each score level, use the provided explanation to understand:
- Why the exemplar belongs in that category
- What qualities distinguish it from higher or lower levels
- What specific features (or missing features) define the score band

Figure 4. Prompt for strategy (ii) - 3 Exemplars without rubric provided to LLMs

You are a calibrated essay scorer. Your job is to evaluate a student's response to the question "How can individuals identify signs of cyberbullying?" and each student's response is limited to 125 words. You will receive one category (with score levels, Rubric, exemplars and grading notes) and student responses follows.

1. Grading Instructions:
Step 1: Read the given student response.
Step 2: Use the Rubric to assign the score.
Step 3: Focus only on the specific evaluation category you have been assigned.
Step 3: Use the two provided Exemplars for this category along with the scoring guidelines.
Step 4: Assign a score from 0–5 based solely on this category.
Step 5: Provide a clear explanation to the student describing the strengths and weaknesses (if any) of their response to justify the score.
Step 6: Provide a confidence percentage indicating how certain you are about the accuracy of your scoring
Use these ranges:
0–40% = Low confidence (uncertain or borderline)
41–70% = Medium confidence (some ambiguity)
71–100% = High confidence (score clearly fits the rubric).

2. Guidelines for Grading responses using the Exemplars:
- Use the exemplars to understand what performance looks like at different score levels.
- Compare the student response to the exemplar patterns, not to your personal expectations.
- Follow the scoring notes provided for the category.
- Do not evaluate anything outside the assigned category (e.g., grammar, style, organization) unless the category explicitly includes it.

3. Explanation of Why Each Exemplar Fits Its Category:
For each score level, use the provided explanation to understand:
- Why the exemplar belongs in that category
- What qualities distinguish it from higher or lower levels
- What specific features (or missing features) define the score band

Figure 5. Prompt for strategy (iii) - 3 Exemplars without rubric provided to LLMs

Results

We employed three distinct strategies to evaluate student responses. In Strategy (i) (Figure 3), Only Rubric, the ensemble of large language models (LLMs) was provided solely with the task instructions and the corresponding rubric criterion. For instance, the model was given the prompt along with the rubric entry for a specific category, such as Relevance to Question.

For Strategies (ii) and (iii), two configurations were considered: Original and Refined.

In the Original configuration, the LLM was explicitly assigned a role and guided through a structured, step-by-step procedure (e.g., chain-of-thought prompting) to mitigate ambiguity in complex grading tasks. Additionally, 3 exemplars were provided for each score level, including explanations detailing how to grade student responses and why alternative score levels were not appropriate. Given six score levels (0–5), this resulted in a total of 18 exemplars per evaluation category. In the Refined configuration, the 3 exemplars and prompts were enhanced to improve precision, consistency, and diagnostic clarity. The revisions ensured tighter alignment with score band definitions and minimized ambiguity in grading.

In Strategy (ii) (Figure 4), Exemplars without Rubric, each LLM was provided with 3 exemplars corresponding to each score level (0–5), along with detailed guidance on grading. Each exemplar included justification explaining why it belonged to a specific score and not adjacent scores. For example, an exemplar assigned a score of 4 included reasoning clarifying why it did not qualify for scores of 3 or 5. This approach was applied in both the Original and Refined configurations.

In Strategy (iii) (Figure 5), Exemplars with Rubric, the same 3 exemplar-based guidance from Strategy (ii) was supplemented with the corresponding rubric criteria. In summary, each LLM was given three distinct prompting strategies and the output of LLM had grade and Justification comment which is compared with the instructor grade and comment.

LLM Suitability to Tasks

To determine the suitability of different large language models (LLMs) for the evaluation task, we conducted experiments using five selected LLMs:

Relevance to Question

To analyze the variability in evaluation accuracy across different large language models (LLMs), we conducted experiments focusing on the Relevance to Question category of the rubric. As shown in Table 1, the dataset corresponds to this category, wherein the LLMs assess whether each student response is relevant to the given question based on the provided rubric. In parallel, instructors independently evaluated the same responses, and their assessments were treated as the ground truth for comparison. The highest accuracy values are highlighted in yellow to distinguish them from other reported values.

Table 1. Overall evaluation accuracy for the "Relevance to Question" category across different LLMs and strategies

LLMs	Strategy (i)	Strategy (ii)		Strategy (iii)	
		(a) Original	(b) Refined	(a) Original	(b) Refined
ChatGPT	40%	67%	76%	86%	95%
Qwen3	37%	67%	63%	80%	91%
Gemini	55%	66%	67%	86%	95%
Llama	34%	72%	72%	78%	81%
Mistral	50%	65%	66%	77%	90%

Figure 6 illustrates the performance of different large language models (LLMs) for the Relevance to Question category. The x-axis represents the evaluated LLMs, while the y-axis denotes accuracy in percentage. Under Strategy (i) (Only Rubric), Gemini achieved the highest accuracy of 55% among all models. For Strategy (ii) (Exemplars without Rubric), in the Original configuration, Llama performed best with an accuracy of 72%. In contrast, in the Refined configuration, ChatGPT achieved the highest accuracy of 76%.

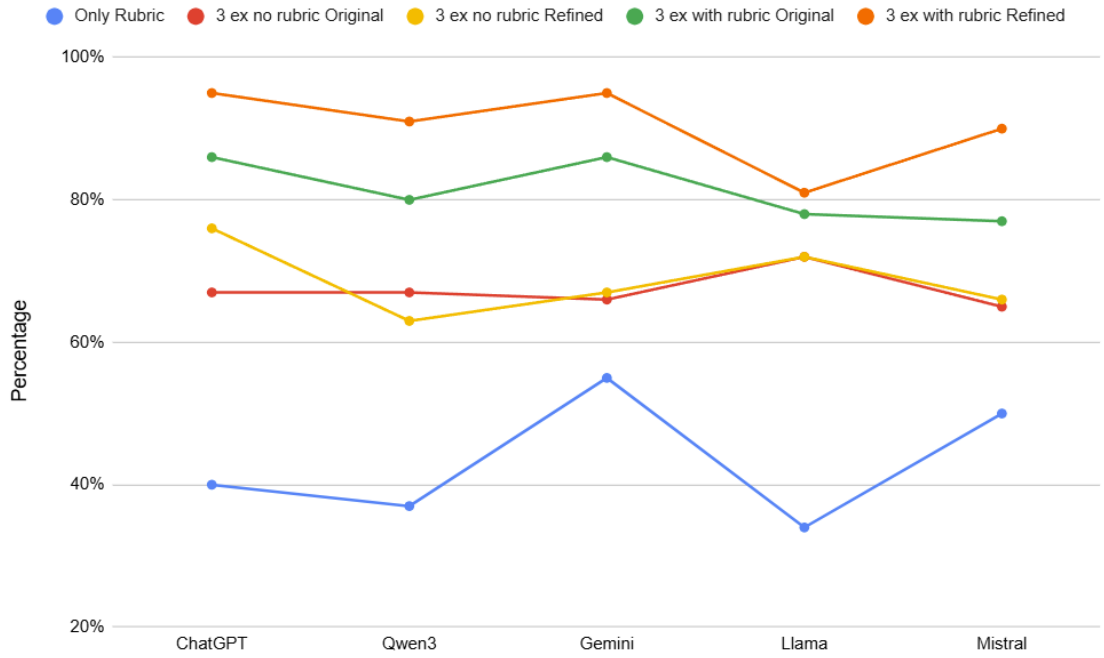

Figure 6. Accuracy of LLMs for the category "Relevance to Question"

For Strategy (iii) (Exemplars with Rubric), strong performance was observed across multiple models. Qwen3 and Mistral achieved accuracies of 91% and 90%, respectively. However, ChatGPT and Gemini outperformed all other models, each attaining an accuracy of 95%, demonstrating greater consistency with instructor-assigned grades, as also reflected in Figure 4. Llama showed comparatively lower performance in this setting, with an accuracy of 81%. These results suggest that, for the Relevance to Question category, ChatGPT and Gemini are the most suitable candidates for inclusion in the ensemble model when adopting a category-wise task division approach.

Reasoning and Critical Thinking

Table 2. Overall evaluation accuracy for the "Reasoning and Critical Thinking" category across different LLMs and strategies

LLMs	Strategy(i)	Strategy(ii)		Strategy(iii)	
		(a)Original	(b)Refined	(a)Original	(b)Refined
ChatGPT	33%	48%	44%	85%	85%
Qwen3	31%	71%	86%	84%	93%
Gemini	75%	88%	73%	78%	89%
Llama	34%	73%	72%	86%	85%
Mistral	36%	43%	54%	62%	64%

To analyze the variability in evaluation accuracy across different large language models (LLMs), we conducted experiments on the Reasoning and Critical Thinking category of the rubric. Table 2 presents the overall evaluation accuracy for this category across multiple LLMs and evaluation strategies. In this setting, the LLMs assess whether each student response satisfies the Reasoning and Critical Thinking criteria based on the provided rubric. In parallel, instructors independently evaluated the same responses, and their assessments were treated as the ground truth for comparison. The highest accuracy values are highlighted in yellow to distinguish them from other reported values.

Figure 7, Under Strategy (i) (Only Rubric), Gemini achieved the highest accuracy of 75%. For Strategy (ii) (Exemplars without Rubric), in the Original configuration, Gemini again performed best with an accuracy of 88%, whereas in the Refined configuration, Qwen3 achieved the highest accuracy of 86%. Under Strategy (iii) (Exemplars with Rubric), in the Original configuration, Llama performed best with an accuracy of 86%.

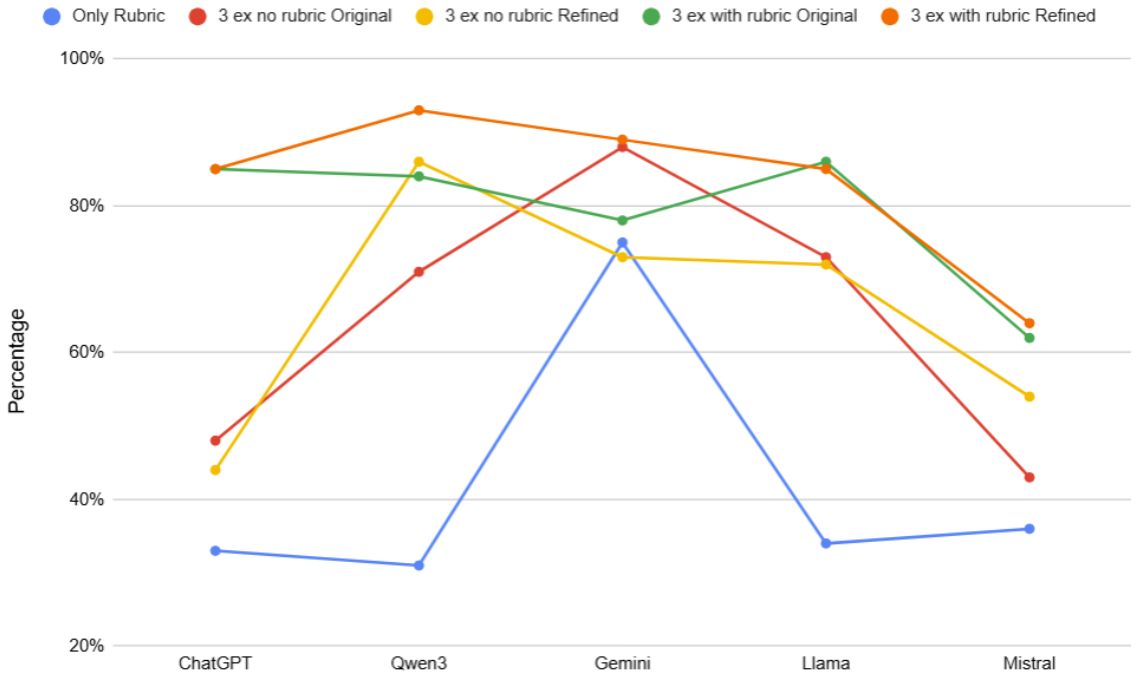

Figure 7. Accuracy of LLMs for the category "Reasoning and Critical Thinking"

Overall, Qwen3 achieved the highest accuracy of 93%, followed by Gemini with 89%. ChatGPT and Llama demonstrated comparable intermediate performance, while Mistral exhibited relatively lower accuracy for this category. These results indicate that Qwen3 is the most suitable model for the Reasoning and Critical Thinking category and should be prioritized for inclusion in the ensemble to achieve improved evaluation performance.

Evidence and Examples

The next category evaluated was Evidence and Examples, following the same experimental procedure as described previously. Table 3 presents the accuracy percentages across different large language models (LLMs) and evaluation strategies. In this category, the LLMs assess whether student essays include appropriate evidence and examples, and assign grades based on the provided rubric criteria. In parallel, instructors independently evaluated the same essays, and their assessments were used as the ground truth for comparison. The highest accuracy values are highlighted in yellow to distinguish them from other reported values.

Table 3. Overall evaluation accuracy for the "Evidence and Examples" category across different LLMs and strategies

LLMs	Strategy(i)	Strategy(ii)		Strategy(iii)	
		(a)Original	(b)Refined	(a)Original	(b)Refined
ChatGPT	74%	62%	75%	71%	70%
Qwen3	77%	89%	81%	58%	56%
Gemini	78%	50%	76%	69%	94%
Llama	73%	90%	88%	64%	64%
Mistral	74%	83%	84%	64%	75%

The chart as shown in Figure 8 shows the results correspond to the Evidence and Examples category. Under Strategy (i) (Only Rubric), all LLMs achieved moderate performance, with accuracies around 70%, while Gemini attained the highest accuracy of 78%. For Strategy (ii) (Exemplars without Rubric), Llama demonstrated the best performance in both configurations, achieving accuracies of 90% in the Original configuration and 88% in the Refined configuration, indicating that the Original exemplars were slightly more effective in this case. Under Strategy (iii) (Exemplars with Rubric), in the Original configuration, ChatGPT achieved the highest accuracy of 71%.

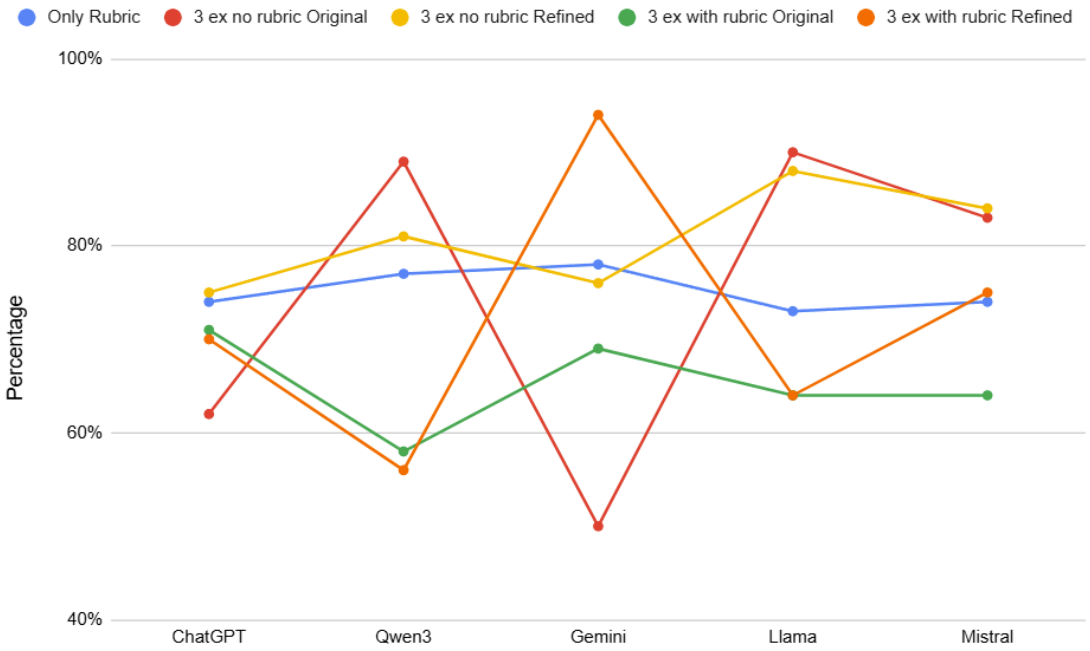

Figure 8. Accuracy of LLMs for the category "Evidence and Examples"

Notably, Gemini achieved the highest overall accuracy of 94% under Strategy (iii) (Refined configuration), while other LLMs showed comparatively lower performance in this setting. Furthermore, for this category, Strategy (i) yielded better accuracy than Strategy (iii) for most LLMs, highlighting an unexpected trend. Based on these observations, Gemini is identified as the most suitable model for the Evidence and Examples category and is recommended for inclusion in the ensemble to achieve improved evaluation performance.

Organization

The next category evaluated was Organization, following the same experimental procedure described previously. Table 4 presents the accuracy percentages across different large language models (LLMs) and evaluation strategies. In this category, the LLMs assess whether student essays exhibit proper organization of content and assign grades based on the provided rubric criteria. In parallel, instructors independently evaluated the same essays for this category, and their assessments were used as the ground truth for comparison. The highest accuracy values are highlighted in yellow to distinguish them from other reported values.

Table 4. Overall evaluation accuracy for the "Organization" category across different LLMs and strategies

LLMs	Strategy(i)	Strategy(ii)		Strategy(iii)	
		(a)Original	(b)Refined	(a)Original	(b)Refined
ChatGPT	32%	78%	72%	60%	73%
Qwen3	64%	50%	49%	76%	78%
Gemini	74%	56%	56%	81%	78%
Llama	43%	86%	90%	78%	95%
Mistral	35%	80%	76%	77%	78%

Looking at the chart in Figure 9, for the Organization category, under Strategy (i) (Only Rubric), Gemini achieved the best performance with an accuracy of 74%. For Strategy (ii) (Exemplars without Rubric), Llama performed well in both configurations, attaining accuracies of 86% in the Original setting and 90% in the Refined setting. Under Strategy (iii) (Exemplars with Rubric), in the Original configuration, Gemini achieved an accuracy of 81%.

Overall, Gemini performed well under Strategy (i) and Strategy (iii) (Original); however, Llama consistently outperformed other models across the remaining strategies and achieved the highest accuracy of 95% under Strategy (iii) (Refined). The remaining LLMs showed comparatively lower performance across this category. These results indicate that Llama demonstrates the strongest overall performance for the Organization category.

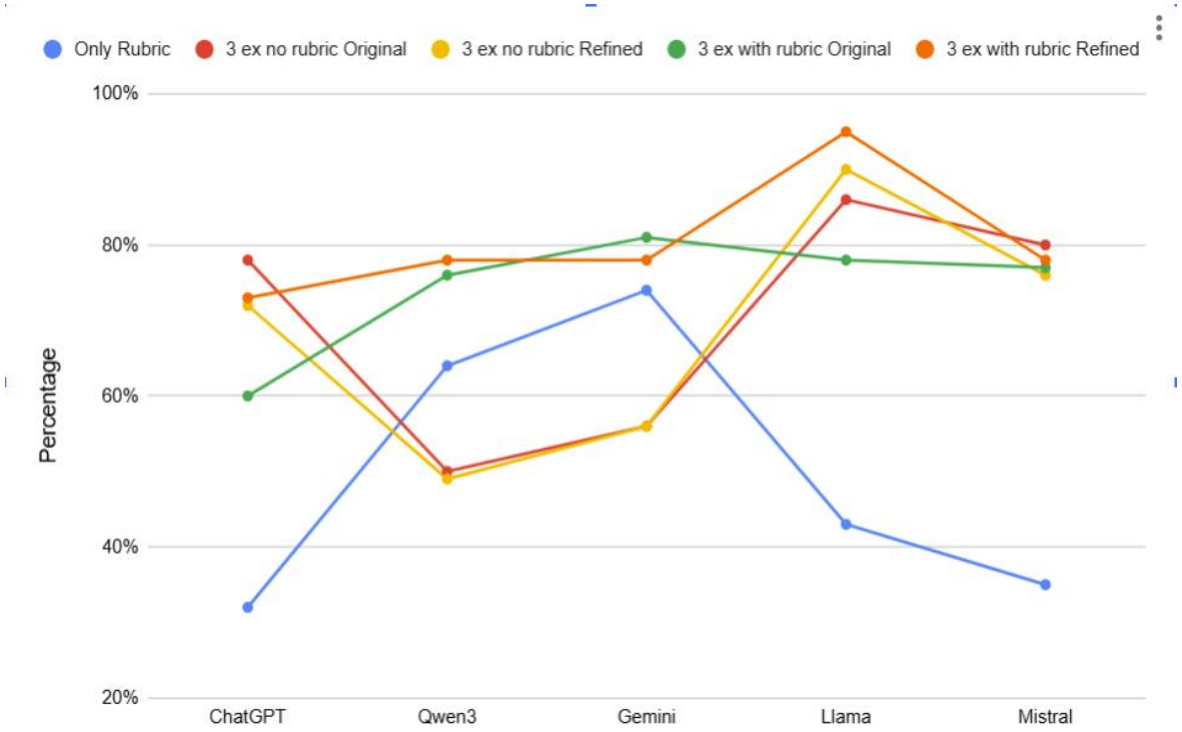

Figure 9. Accuracy of LLMs for the category "Organization"

Clarity and Writing Quality

The final category evaluated was Clarity and Writing Quality, following the same experimental procedure described previously. Table 5 presents the accuracy percentages across different large language models (LLMs) and evaluation strategies. In this category, the LLMs assess whether student essays demonstrate clarity in writing, including clear, concise, and well-polished expression, and assign grades based on the provided rubric criteria. In parallel, instructors independently evaluated the same essays for this category, and their assessments were used as the ground truth for comparison. The highest accuracy values are highlighted in yellow to distinguish them from other reported values.

Table 5. Overall evaluation accuracy for the "Clarity and Writing Quality" category across different LLMs and strategies

LLMs	Strategy(i)	Strategy(ii)		Strategy(iii)	
		(a)Original	(b)Refined	(a)Original	(b)Refined
ChatGPT	25%	64%	58%	84%	81%
Qwen3	36%	82%	79%	72%	79%
Gemini	63%	75%	81%	67%	54%
Llama	43%	61%	59%	73%	74%
Mistral	47%	87%	91%	77%	96%

From the chart (see Figure 10), for the Clarity and Writing Quality category, under Strategy (i) (Only Rubric), Gemini achieved the best performance with an accuracy of 63%. For Strategy (ii) (Exemplars without Rubric), Mistral AI performed best in both configurations, achieving accuracies of 87% in the Original setting and 91% in the Refined setting. Under Strategy (iii) (Exemplars with Rubric), in the Original configuration, ChatGPT achieved an accuracy of 84%. Overall, Mistral AI demonstrated the highest performance for this category, with an overall accuracy of 96%, followed by ChatGPT with 81%. The remaining models showed comparatively lower performance across the evaluated strategies. These results highlight that selecting task-specific LLMs significantly improves evaluation performance. This finding is particularly important for multi-agent frameworks such as Ensemble LLMs, where different subtasks can be assigned to specialized models to enhance overall system effectiveness.

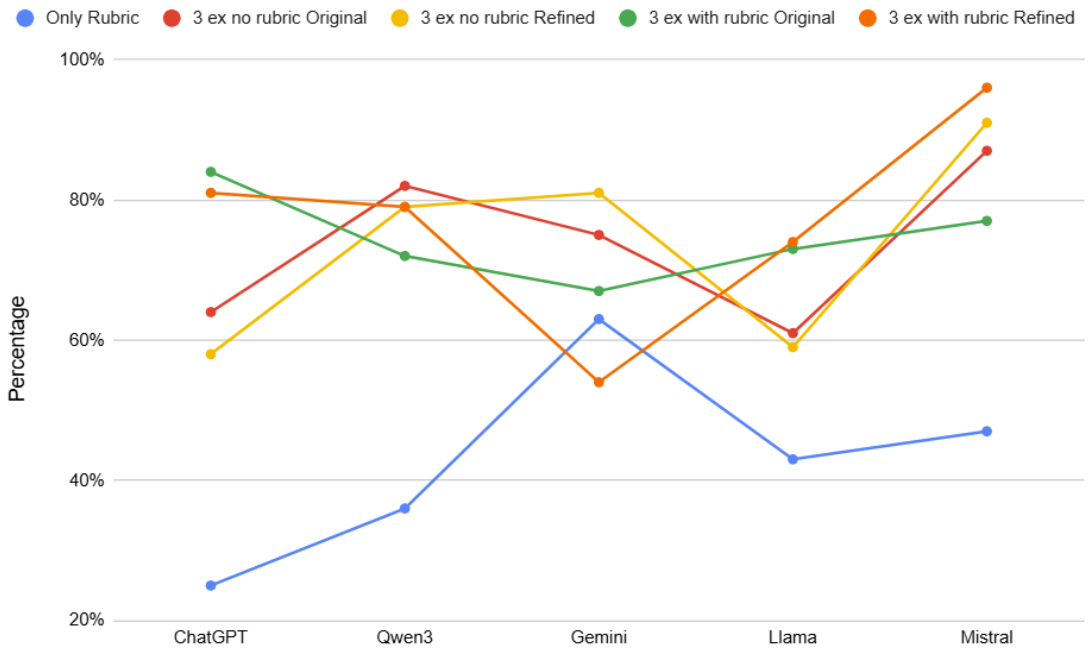

Figure 10. Accuracy of LLMs for the category "Clarity and Writing Quality"

The Effect of Prompts

To analyze the effect of prompt design on evaluation performance, three different prompting strategies were considered.

Prompt 1 provided minimal instructions, where the LLM was asked to grade 100 student responses (each limited to 125 words) using the provided rubric for the question "How can individuals identify signs of cyberbullying?". The model was instructed only to return the results in a table format.

Prompt 2 specified the role of the LLM as an academic grader and retained the same task constraints (100 responses, 125-word limit, and use of the provided rubric for the same question). In addition, a structured output format was provided, including a predefined table template to guide the response formatting.

Prompt 3 incorporated a step-by-step reasoning procedure (chain-of-thought style instructions) prior to grading. The LLM was guided through explicit grading steps along with detailed grading instructions and the rubric for the same question before evaluating the student responses.

Table 6. Accuracy calculated for different types of prompting

Effect of prompt	1st Prompt	2nd Prompt	3rd Prompt
Accuracy	0%	32%	95%

For Prompt 1, the model generated outputs but failed to correctly interpret that averaging across categories was required, and that the final values needed to be rounded. As a result, although instructor grades were represented as numeric scores, the model produced outputs in formats such as 23/25, leading to incompatibility with the expected evaluation scale. Consequently, the overall accuracy achieved for this prompt was 0%. For Prompt 2, the model correctly inferred that averaging was required; however, performance remained limited due to inconsistencies in numerical formatting, particularly the presence of unrounded decimal values. This resulted in an overall accuracy of 32%. In contrast, Prompt 3, which explicitly included step-by-step instructions and detailed grading guidelines, enabled the model to correctly interpret all requirements, including averaging and rounding. This led to a significant improvement in performance, achieving an accuracy of 95%.

This experiment demonstrates that prompt design plays a critical role in enabling large language models (LLMs) to correctly interpret and execute evaluation tasks. In particular, the final prompt, which incorporated chain-of-

thought style step-by-step grading instructions, significantly improved performance by aligning the model's reasoning process with human-like evaluation behavior.

Conclusion

This study investigated the effectiveness of large language models (LLMs) for automated essay evaluation across multiple rubric-based categories, including Relevance to Question, Reasoning and Critical Thinking, Evidence and Examples, Organization, and Clarity and Writing Quality. We systematically evaluated different prompting strategies (Only Rubric, Exemplars without Rubric, and Exemplars with Rubric) under both original and refined configurations, along with a prompt ablation study.

The results demonstrate that no single LLM performs best across all evaluation dimensions. Instead, model performance varies significantly by category, indicating that task-specific model selection is essential. For example, ChatGPT and Gemini performed best for Relevance to Question, Qwen3 achieved the highest accuracy for Reasoning and Critical Thinking, Gemini was most suitable for Evidence and Examples, Llama excelled in Organization, and Mistral AI performed best for Clarity and Writing Quality. These findings highlight the effectiveness of a category-wise ensemble approach for improving overall grading accuracy. Additionally, the effect of prompting study shows that prompt design has a substantial impact on performance. Minimal prompting resulted in poor or inconsistent outcomes, while structured prompts improved accuracy, and chain-of-thought-based prompts achieved the highest performance by aligning model reasoning with human grading behavior.

Overall, the study confirms that combining task-specific LLM selection with well-designed structured prompts significantly enhances automated essay evaluation performance. These findings are particularly relevant for multi-agent frameworks, where different LLMs can be assigned to specialized evaluation tasks to improve robustness, consistency, and alignment with human grading standards.

Scientific Ethics Declaration

* The authors declare that the scientific ethical and legal responsibility of this article published in EPESS journal belongs to the authors.

Conflict of Interest

* The authors declare that they have no conflicts of interest

Funding

* No funds have been received from outside sources supporting this research.

Acknowledgements or Notes

* This article was presented as an oral presentation at the International Conference on Research in Education and Technology (www.icret.net) held in Konya/Türkiye on April 02-05, 2026.

References

Golden, B. (2023). Enabling critical thinking development in higher education through the use of a structured planning tool. *Irish Educational Studies, 42*(4), 941–969.

Huang, Y., & Wilson, J. (2025). Evaluating LLM-based automated essay scoring: Accuracy, fairness, and validity. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con)* (pp. 71–83).

Nickelson, N., & Schaffer, T. (2025). How does the use of contract grading affect the quality of student writing?:

A comparative study of essays from first-year writing courses. *College Teaching*, 1–12.

Wang, X., Wei, J., Schuurmans, D., Pfeifer, J., Tay, Y., Libman, M., … & Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models. *International Conference on Learning Representations (ICLR)*.

Yalpi, D., Kasula, S. K., & Mukkamala, R. (2026). LLM consortium assessment of critical thinking in student. In *Proceedings of the Fifth International Conference on Innovations in Computing Research (ICR'26)*. Springer Nature.

Yalpi, D., Kasula, S. K., & Mukkamala, R. (2026). Multi-Agent LLMs for Automated Critical Thinking Assessment: A Scalable and Reliable Essay Scoring Framework. In *Proceedings of the Fifth International Conference on Innovations in Computing Research (ICR'26).* Springer Nature.

Author(s) Information

Diwakar Yalpi
Old Dominion University, Norfolk, Virginia, 23529, USA
ORCID ID: https://orcid.org/0009-0007-2205-9723

Sruta Keerti Kasula
Old Dominion University, Norfolk, Virginia, 23529, USA
ORCID ID: https://orcid.org/0009-0001-9024-1873

Ravi Mukkamala
Old Dominion University, Norfolk, Virginia, 23529, USA
ORCID iD: https://orcid.org/0000-0001-6323-9789
*Contact e-mail: rmukkama@odu.edu

*Corresponding author(s) contact email

To cite this article:

Yalpi, D., Kasula, S.K. & Mukkamala, R. (2026). Guidelines for automatic grading of student essays using large language models. *The Eurasia Proceedings of Educational and Social Sciences (EPESS), 49,* 228-244.