A Human-in-the-Loop Benchmark for Assessing LLM Instruction Adherence in Qualitative Education Rubrics

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are increasingly used for automated educational assessment, yet their ability to follow structured qualitative rubrics remains poorly understood. We introduce a human-in-the-loop benchmarking framework to evaluate LLM adherence to hierarchical competency-based grading criteria. The benchmark is constructed from a Grade 10 mathematics assessment aligned with the Nepal national curriculum and evaluated by expert teachers, producing a high-reliability ground truth ($\kappa_w = 0.8652$). Four heterogeneous LLMs, including open-weight dense models and proprietary Mixture-of-Experts (MoE) models, are tasked with assigning competency levels across a multi-dimensional rubric covering comprehension, knowledge, operational fluency, and behavioral correlation. Results show that current LLMs struggle to replicate expert qualitative evaluation, with the best model achieving only fair agreement with human graders ($\kappa_w \approx 0.385$). Dense models demonstrate particularly poor adherence to hierarchical scoring logic, while MoE architectures exhibit more consistent rubric compliance. These findings highlight a significant gap between LLM reasoning capabilities and structured educational evaluation, suggesting that reliable deployment requires transparent human-in-the-loop oversight.

1 Introduction

Evaluating whether Large Language Models (LLMs) can follow structured multi-step instructions remains an open challenge, particularly in tasks that require hierarchical reasoning and strict rubric adherence (He et al., 2024; Zhang et al., 2024, 2025). Educational assessment provides a natural testbed for such evaluation problems (Cao et al., 2024; Doewes, 2026; Hou et al., 2024). In many school systems, student performance is traditionally measured through numerical scores rather than holistic competencies (Radu et al., 2024; Hochstetter-Diez et al., 2025; Huang, 2025). This creates a gap in assessment systems, as marks alone often fail to capture a student's understanding of complex multi-step reasoning (Arinaitwe and Kazaara, 2025). In mathematics, for example, a student may demonstrate conceptual insight yet fail due to minor procedural slips, or conversely apply memorized formulas correctly without understanding underlying logic (Umar et al., 2025). Competency-Based Education (CBE) attempts to address these limitations by evaluating deeper reasoning processes rather than only final answers (Chappell et al., 2020; Hernández-de Menéndez and Morales-Menendez, 2016).

Recent advances in LLM reasoning capabilities suggest that these models could assist educators in evaluating qualitative reasoning processes rather than only final answers (Wangwiwattana and Tongvivat, 2023). However, the reliable deployment of LLMs in educational assessment remains challenging (Doewes, 2026; Li et al., 2024). A key limitation is the lack of empirical evaluation frameworks that measure whether models can follow structured grading rubrics and replicate expert judgment (Doewes, 2026; Cao et al., 2024). Existing studies often propose theoretical AI-assisted grading systems but rarely evaluate them against expert human assessments in controlled experiments (Owan et al., 2023). Moreover, rubric-based competency evaluation requires models to follow hierarchical instructions and synthesize multiple reasoning dimensions, which are capabilities that remain largely unexplored in current LLM benchmarks (Setälä et al., 2025).

To address this gap, we introduce a Human-in-the-Loop (HITL) benchmarking framework designed to evaluate LLM instruction adherence in competency-based educational assessment. Instead of using LLMs as opaque grading tools, our framework tests whether models can follow a hierarchical rubric and assign qualitative competency

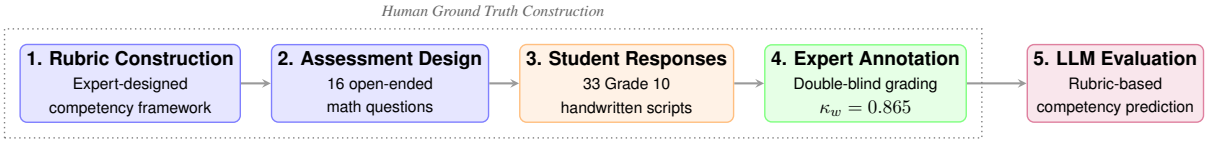

Figure 1: Overview of the benchmark construction and evaluation pipeline. Human experts design the rubric, create the assessment, and produce the ground-truth annotations before LLMs are evaluated on the same rubric-based competency prediction task.

labels consistent with expert teachers. We operationalize this evaluation using a Grade 10 Optional Mathematics assessment aligned with the Nepal national curriculum (Curriculum Development Centre, 2019). Student responses are independently evaluated by two senior mathematics faculty members, producing a high-reliability ground truth ($\kappa_w = 0.8652$). We then task four heterogeneous LLMs with performing the same rubric-based evaluation, allowing us to quantitatively measure their agreement with expert human judgment. Our main contributions are:

- A human-in-the-loop benchmark for evaluating LLM adherence to hierarchical competency rubrics in mathematics assessment.
- A curated dataset of student responses with expert-annotated competency labels achieving high inter-rater reliability ($\kappa_w = 0.8652$).
- An empirical analysis of heterogeneous LLM architectures, revealing significant limitations in rubric adherence and highlighting architectural differences in instruction compliance.

2 Benchmark Construction and Experimental Setup

Evaluating LLM adherence to qualitative grading rubrics requires a structured benchmarking pipeline that enables direct comparison between model predictions and expert human judgment. To construct this benchmark, domain experts designed the competency rubric, authored the assessment questions, and performed the ground-truth evaluation of student responses. This expert-driven process ensures that the dataset reflects authentic pedagogical reasoning tasks rather than simple answer verification. The benchmark pipeline consists of five stages, as illustrated in Figure 1: (1) rubric construction, (2) assessment design, (3) student data collection, (4) expert annotation, and (5) LLM evaluation.

2.1 The Hierarchical Evaluation Rubric

We formalized a multi-dimensional rubric to shift assessment from binary correctness toward qualitative reasoning analysis. The framework evaluates four cross-cutting cognitive competencies: **Comprehension** (problem interpretation), **Knowledge** (conceptual depth), **Operational Fluency** (procedural accuracy), and **Behavior & Correlation** (interdisciplinary synthesis). Student proficiency within each competency is classified into four progressive levels: Awareness (1), Application (2), Mastery (3), and Influence (4).

Unlike traditional automated grading, our system evaluates the *process trajectory* of a student's reasoning rather than only the final answer. This structure allows the evaluation to distinguish between types of reasoning errors, such as mechanical calculation mistakes (Operational Fluency) and conceptual misunderstandings (Knowledge).

To operationalize the rubric, we established specific measurement criteria for each competency. Knowledge is evaluated sequentially across the four questions within each topic. Solving the first question corresponds to the Awareness level, the second to Application, and progressively higher levels for subsequent questions. Behavior and Correlation are assessed using the final, most complex question in each topic, which requires interdisciplinary reasoning. In contrast, Comprehension and Operational Fluency are evaluated holistically across the student's entire response. Appendix Table 3 summarizes the criteria used to determine levels for each competency.

This rubric provides the link between raw student responses and the competency labels used in our evaluation framework. The hierarchical structure ensures logical consistency between levels; for example, a student cannot be categorized at the *Influence* level without first demonstrating *Mastery* of the underlying procedural operations. Using this rubric, expert annotators evaluated the collected student responses to produce the ground truth dataset used for benchmarking LLM performance.

2.2 Assessment Design and Student Dataset

To empirically evaluate the proposed rubric, we designed a mathematics assessment aligned with the Grade 10 Optional Mathematics curriculum in Nepal. The assessment consisted of 16 open-ended questions, with four questions for each of four core topics: Matrix Algebra, Coordinate Geometry, Trigonometry, and Functions. The full question paper is provided in Appendix B.

We conducted an experimental study with 33 Grade 10 students who completed this assessment under exam conditions. Open-ended questions were intentionally selected over multiple-choice formats to capture sequential reasoning steps, reveal partial misconceptions, and allow partially correct solutions.

Student responses were handwritten and later digitized using an OCR system, followed by manual correction to ensure transcription accuracy. Student responses were assigned anonymized codes, and pages containing noise, empty sections, crossed-out answers, doodles, or scribbles were excluded to maintain transcription quality. The dataset will be made publicly available.

2.3 Establishing the Expert Ground Truth

Each of the 33 student submissions was independently evaluated by two subject matter experts with more than five years of teaching experience. The experts analyzed the four questions per topic and assigned a competency level (1–4) for each of the four cross-cutting skills defined in the rubric. This process produced 16 topic-level competency labels per student (four topics × four competencies). The experts then synthesized these topic-level evaluations to assign four global competency labels, resulting in 20 qualitative labels per student, as illustrated in Figure 2. A double-blind grading protocol was employed, achieving a high inter-rater reliability of $\kappa_w = 0.8652$, establishing a reliable human ground truth. In cases of disagreement, a third senior expert (the Lead Adjudicator) resolved the final label.

2.4 LLM Prompting and Comparative Setup

With the expert ground truth established, we evaluated the capability of multiple LLMs to replicate the same qualitative assessment process. Four heterogeneous LLMs were individually prompted to perform the identical evaluation task using a single master evaluation prompt (Appendix C.1). The standardized prompt initialized the model as an expert mathematics educator responsible for evaluating student reasoning.

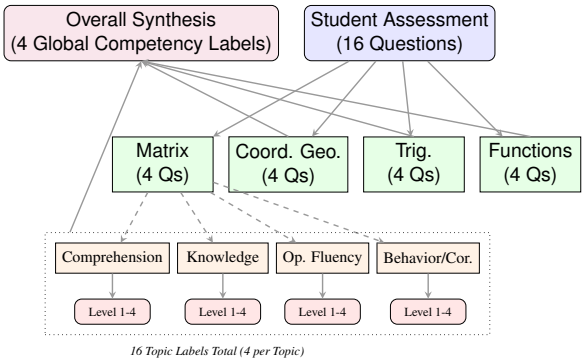

Figure 2: Hierarchical competency evaluation rubric. Student responses from the 16-question assessment are mapped to 4 mathematical topics and evaluated across 4 competencies. Each competency receives an ordinal level (1–4), producing 16 topic-level labels which are then synthesized into four global competency scores.

The prompt design followed three principles. First, models were prohibited from assigning numerical scores, forcing them to produce qualitative competency labels aligned with the rubric. Second, the output format required structured tables capturing both unit-level reasoning and global competency assessments. Third, models were required to cite explicit evidence from student responses when assigning competency levels, improving interpretability and reducing unsupported judgments.

2.5 LLM Selection and Evaluation Metrics

The assessment utilizes a multi-provider suite of models designed to balance reasoning capacity with architectural diversity. As shown in Table 1, the models are categorized by their structural frameworks and tier-based accessibility, which is an important consideration for scalable implementation in the Global South.

Model ID	Base Model	Tier	Primary Role
EAGLE	Llama 3.1-8B (Grattafiori et al., 2024)	Free	Latency baseline and speed testing.
ORION	Llama 3.3-70B (Grattafiori et al., 2024)	Free	Complex reasoning for multi-step proofs.
NOVA	Gemini 2.5 Flash (Team et al., 2023)	Paid	High-throughput Sparse MoE logic.
LYRA	Gemini 3 Pro (Team et al., 2023)	Paid	Expert Arbiter and CoT reasoning.

Table 1: Model Configuration and Deployment Details.

The core assessment was conducted at a uniform temperature of 0.1 to maximize determinism and reproducibility. EAGLE and ORION serve as open-weight dense baselines, while NOVA and

LYRA represent proprietary, frontier-class Mixture-of-Experts (MoE) architectures. This heterogeneous setup allows us to examine whether architectural differences affect rubric adherence.

Statistical Control: Model performance was evaluated using Quadratic Weighted Cohen's Kappa (κ_w), which accounts for the ordinal structure of competency levels. The weighting scheme penalizes larger disagreements more heavily than minor level differences. Agreement was measured across three axes: Human–Human (baseline reliability), Human–LLM (instruction adherence), and Inter-Model Consensus.

Human Ground Truth (κ_w)			Inter-Model Consensus		
Evaluator Pair	**Score**	**Level**	**Model Pair**	**Score**	**Level**
T1 vs. T2	0.8652	Perfect	Lyra vs Nova	0.5597	Mod.
NOVA (MoE)	0.3852	Fair	Lyra vs Eagle	0.2791	Fair
LYRA (MoE)	0.2693	Fair	Eagle vs Nova	0.2466	Fair
EAGLE (8B)	0.1030	Slight	Nova vs Orion	0.0417	Slight
ORION (70B)	-0.0261	None	Eagle vs Orion	0.1234	Slight

Table 2: Kappa Results (κ_w) and Model Consensus

3 Results

The evaluation reveals a substantial gap between human agreement and LLM rubric adherence. As shown in Table 2, the baseline comparison between the two human experts yielded high agreement ($\kappa_w = 0.8652$), indicating that the hierarchical rubric produces consistent evaluations across annotators and provides a stable ground truth for model comparison.

When compared to this human baseline, the evaluated LLMs exhibited substantially lower agreement levels. The highest performing model was NOVA (Gemini 2.5 Flash), achieving "Fair" agreement with human evaluators ($\kappa_w = 0.3852$). LYRA (Gemini 3 Pro) followed with a similar but lower score ($\kappa_w = 0.2693$). In contrast, the dense models demonstrated minimal alignment with the expert annotations. Notably, the 70B dense model ORION produced a negative Kappa score ($\kappa_w = -0.0261$), indicating disagreement with the human ground truth. The inter-model agreement analysis reveals additional structure. Agreement was highest between NOVA and LYRA ($\kappa_w = 0.5597$), suggesting that the two models produce relatively consistent rubric interpretations compared to the other models in the evaluation.

4 Discussion

The results highlight the difficulty current LLMs face when applying hierarchical evaluation rubrics to multi-step mathematical reasoning. Although the human annotators achieved strong agreement, all evaluated models showed substantially lower alignment with the expert ground truth.

Architectural Differences. An interesting pattern emerges across model architectures. The two Mixture-of-Experts (MoE) models achieved higher agreement with human annotations than the dense models evaluated in this study. While the benchmark size is limited, this observation suggests that architectural design may influence how well models follow complex evaluation instructions.

Inter-Model Agreement. The MoE models demonstrated higher agreement with each other ($\kappa_w = 0.5597$) than with human evaluators. This pattern indicates that the models may apply internally consistent evaluation strategies that do not fully align with human pedagogical reasoning. Further work is needed to understand how models interpret rubric-based grading criteria.

Implications for Automated Assessment. With the highest Human–LLM agreement reaching only $\kappa_w = 0.3852$, the results suggest that current LLMs are not yet reliable as fully autonomous graders for qualitative, multi-dimensional mathematics assessments. However, their ability to extract structured reasoning evidence indicates potential value as assistive tools within Human-in-the-Loop evaluation workflows.

5 Conclusion

This paper introduces a Human-in-the-Loop benchmarking framework for evaluating LLM adherence to hierarchical competency-based assessment rubrics in secondary mathematics. Using expert-annotated student responses aligned with the Nepal Grade 10 Optional Mathematics curriculum, we conducted a controlled comparison between human evaluators and four heterogeneous LLMs. The results show that while human annotators achieve high agreement under the proposed rubric ($\kappa_w = 0.8652$), current LLMs exhibit substantially lower alignment with expert judgments. Among the evaluated models, MoE architectures demonstrated relatively stronger performance, although agreement levels remain insufficient for autonomous grading. These findings suggest that rubric-based qualitative assessment remains a challenging task for current language models.

6 Limitations and Future Work

This study has several limitations. First, the evaluation relies on ordinal competency labels measured using Quadratic Weighted Cohen's Kappa (κ_w). While this metric captures agreement in ordinal judgments, it does not fully reflect qualitative differences in the reasoning explanations produced by the models. Future work could incorporate more detailed analysis of the reasoning evidence generated during rubric-based evaluation.

Second, the dataset consists of responses from 33 students drawn from a single institution and curriculum context. Although the high human inter-rater agreement suggests that the rubric is internally consistent, the dataset size and geographic scope limit the generalizability of the findings. Expanding the benchmark across multiple schools, curricula, and linguistic contexts will be necessary to validate the robustness of the framework.

Future work will also explore methods for improving model reliability in rubric-based evaluation tasks. This includes investigating alternative prompting strategies, structured reasoning verification, and multi-model evaluation approaches. Additionally, we plan to extend the benchmark to other subject areas in order to study whether similar challenges arise in competency-based assessment beyond mathematics.

7 Ethical Considerations

This study was conducted with attention to data privacy and responsible use of educational data. Student responses were collected with permission from the participating institution, and all handwritten scripts were anonymized prior to digitization. Personal identifiers were removed and replaced with randomized codes to protect student identity.

The results also highlight potential risks associated with automated educational assessment. Current LLMs demonstrate limited agreement with expert human evaluators in qualitative grading tasks. As a result, autonomous deployment of such systems in high-stakes educational settings could introduce grading errors or reinforce automation bias, where human evaluators place undue trust in machine-generated judgments.

For this reason, we frame the proposed system as a Human-in-the-Loop assistive tool designed to support educators rather than replace them. Human oversight remains essential when applying AI systems to pedagogical evaluation tasks.

References

Julius Arinaitwe and Asiimwe Isaac Kazaara. 2025. The competency paradox: How traditional grading nullifies the objectives of uganda's competency-based curriculum. *International Journal of Academic Multidisciplinary Research (IJAMR)*, 9(11):370–379.

Boxi Cao, Mengjie Ren, Hongyu Lin, Xianpei Han, Feng Zhang, Junfeng Zhan, and Le Sun. 2024. Structeval: Deepen and broaden large language model assessment via structured evaluation. In *Findings of the Association for Computational Linguistics, ACL 2024*.

Clive Chappell, Andrew Gonczi, and Paul Hager. 2020. Competency-based education. In *Understanding adult education and training*, pages 191–205. Routledge.

Curriculum Development Centre. 2019. *Optional Mathematics Grade 10*. Ministry of Education, Science and Technology, Government of Nepal, Sanothimi, Bhaktapur, Nepal.

Afrizal Doewes. 2026. *Rethinking Automated Essay Scoring: Agreement, Fairness, and Feedback*. Ph.D. thesis, Eindhoven University of Technology.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Qianyu He, Jie Zeng, Wenhao Huang, Lina Chen, Jin Xiao, Qianxi He, Xunzhe Zhou, Jiaqing Liang, and Yanghua Xiao. 2024. Can large language models understand real-world complex instructions? In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence*.

Marcela Hernández-de Menéndez and Ruben Morales-Menendez. 2016. Competency based education – current global practices. *World Journal of Engineering and Technology*, 4(3):193–199.

Jorge Hochstetter-Diez, Marlene Negrier-Seguel, Mauricio Diéguez-Rebolledo, Esteban Candia-Garrido, and Elizabeth Vidal. 2025. From mapping to action: Smartrubrics, an ai tool for competency-based assessment in engineering education. *Sustainability*, 17(13):6098.

Jinchang Hou, Chang Ao, Haihong Wu, Xiangtao Kong, Zhigang Zheng, Daijia Tang, Chengming Li, Xiping Hu, Ruifeng Xu, Shiwen Ni, and Min Yang. 2024. E-eval: A comprehensive chinese k-12 education evaluation benchmark for large language models. In *Findings of the Association for Computational Linguistics, ACL 2024*.

Xin Huang. 2025. Designing human-ai orchestrated classrooms: Mechanisms, protocols, and governance for competency-based education. *Artificial Intelligence Education Studies*, 1(3):14–27.

Xiaoyuan Li, Wenjie Wang, Moxin Li, Junrong Guo, Yang Zhang, and Fuli Feng. 2024. Evaluating mathematical reasoning of large language models: A focus on error identification and correction. In *Findings of the Association for Computational Linguistics, ACL 2024*.

Valentine Joseph Owan, Kinsgley Bekom Abang, Delight Omoji Idika, Eugene Onor Etta, and Bassey Asuquo Bassey. 2023. Exploring the potential of artificial intelligence tools in educational measurement and assessment. *EURASIA Journal of Mathematics, Science and Technology Education*, 19(8).

Cătălina Radu, Carmen Nadia Ciocoiu, Cristina Veith, and Răzvan Cătălin. 2024. Artificial intelligence and competency-based education: A bibliometric analysis. *Amfiteatru Economic*, 26(65):220–240.

Mika Setälä, Pieta Sikström, Ville Heilala, and Tommi Kärkkäinen. 2025. Assessment of evolving large language models in upper secondary mathematics. *arXiv preprint*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Umar, Ahmad Budi Sutrisno, Andi Kaharuddin, and Vishal Gupta. 2025. Analysis of student difficulties in solving problems on statistics material. *Edu-Transform: Multidisciplinary International Journal*, 1(1):27–38.

Chatchai Wangwiwattana and Yuwaree Tongvivat. 2023. Automating academic assessment: a large language model approach. In *2023 7th International Conference on Information Technology (InCIT)*, pages 330–334. IEEE.

Li Zhang, Matthias Grabmair, Morgan Gray, and Kevin Ashley. 2025. Thinking longer, not always smarter: Evaluating llm capabilities in hierarchical legal reasoning. *arXiv preprint arXiv:2510.08710*.

Zhihan Zhang, Shiyang Li, Zixuan Zhang, Xin Liu, Haoming Jiang, Xianfeng Tang, Yifan Gao, Zheng Li, Haodong Wang, Zhaoxuan Tan, Yichuan Li, Qingyu Yin, Bing Yin, and Meng Jiang. 2024. Iheval: Evaluating language models on following the instruction hierarchy. In *Findings of the Association for Computational Linguistics, ACL 2024*.

A Criteria for Competency Level Determination

Table 3 provides the complete, granular rubric utilized by both the human experts and the LLMs to assign competency levels. To operationalize the assessment, specific measurement criteria were established for each cognitive function. Knowledge was evaluated sequentially across the four questions per topic, where solving the first question indicated an Awareness level, and progressing to the subsequent questions demonstrated higher levels of mastery. Behavior and Correlation were evaluated specifically using the final, most complex question in each unit. Conversely, Comprehension and Operational Fluency were assessed holistically based on the entirety of the student's response.

B Grade 10 Optional Mathematics Assessment

The following questions comprise the 16-question assessment utilized in the evaluation dataset. Students were encouraged to use answers/references from the prior problems wherever possible.

B.1 Matrix

1. Determine the determinant of the following matrix. Also find its adjoint. $A = \begin{bmatrix} 3 & 2 \\ -3 & k \end{bmatrix}$

2. What is a singular matrix? For what value of 'k' is the following matrix A singular? $A = \begin{bmatrix} 3 & 2 \\ -3 & k \end{bmatrix}$

3. Solve the following simultaneous equation using matrix method:

   $3x + 2y = 5$

   $-3x + 7y = 4$

4. What do you mean by the solution of a system of linear equation? Determinants of both of the following matrices A and B are 0. Explain how the linear system of equations (i) and (ii) differs from one another in terms of solution.

   $A = \begin{bmatrix} 4 & 8 \\ 3 & 6 \end{bmatrix}, B = \begin{bmatrix} 2 & 4 \\ 3 & 6 \end{bmatrix}$

   System (i):

   $4x + 8y = 12$

   $3x + 6y = 9$

   System (ii):

   $2x + 4y = 6$

   $3x + 6y = 10$

B.2 Co-ordinate Geometry

1. Determine the slope of the following lines.

   I. L1: line joining points (1,1) and (3,3)

Level	Comprehension	Knowledge	Operational Fluency	Behavior & Correlation
Awareness (1)	Can only follow basic direction of question.	Prior foundational knowledge to at least begin the topic.	BODMAS.	Understands the topic from which the problem is from.
Application (2)	Ability to understand the direction of question with minimal information.	Has developed foundational knowledge in topic.	Linear equation with one variable.	Understands the possible hurdles and direction of the question.
Mastery (3)	Can relate prior questions in the solution of current problem (if possible).	Has developed sufficient knowledge to understand the topic.	Fractional equations.	Uses proper mathematical tools (formulas in general) to solve the problems.
Influence (4)	Ability to comprehend mathematical literature to mathematical form.	Can correlate with other units.	Quadratic equations.	Ability to solve questions completely or to an appreciably close degree integrating proper steps, and connecting topics.

Table 3: Summarized criteria for competency level determination across the four cognitive functions.

II. L2: line which makes angle 135 degrees clockwise with X-axis with y-intercept 1.

III. L3: $3x + 3y = 3$ Also, in each case find the equation of straight lines.

2. A. Find the angles between L1 and L2. What does the result state?

   B. Find the angles between L1 and L3. What does the result state?

   C. What is the acute angle between two lines with angles $60°$ and $120°$ with X-axis?

3. Find the equation of straight line which is perpendicular to L1: $2x + 3y = 5$ and passes through (1,1). Also find the equation of straight line parallel to L1 but passes through (3,3).

4. Consider the points A(0,0), B(2,2) and C(4,0). Find the angle of lines AB and AC. Hence find the center and radius of the circle passing through these points.

B.3 Function

1. Write the composition $fg(x)$ in the following case. $f(x) = \frac{2x+3}{x+1}, g(x) = x + 2$. For what value of x is $fg(x)$ not defined.

2. In the above question, find $(fg)^{-1}(x)$.

3. If $g(x) = 2x - 3$, $g \circ f(x) = 2x + 5$ If $f^{-1}(x) = 2x - 6$ find the value of x.

4. Ram is planning to invest a sum of money in companies A and B, but he can only invest in one company for 1 year and at most for a year in that company. This means he cannot invest the money twice in the same company. Company A doubles the money invested in a year whereas company B ensures that he earns 1,000 in profits. Suppose his initial investment is 'x'. What will be the final money after 2 years he will have if:

   i. He invests first in company A?

   ii. He invests first in company B?

   (For extra credit, can you express the above schemes in using functional notations, say $f(x)$ and $g(x)$?)

   Through which scheme (i) or (ii) will he have more money at the end of two years? Also find the initial investment through scheme(ii) so that the money after 2 years is 3000.

B.4 Trigonometry

1. a. In the given right triangle OAC, find the values of $\sin A$ and $\cos A$.

   b. In the given right triangle OBD, find the values of $\sin B$ and $\cos B$.

2. Find the value of $\sin(A + B)$, $\cos(A - B)$, $\sin 2A$, $\cos 2A$ and $\tan 2A$ from the information in (1).

3. Solve for x: $(0 < x < 90°)$

   $\cos 3x + \cos x = \cos 2x$.

4. Ram and Shyam are planning to measure the height of a building. Shyam, who is on the top of the building, observes that the distance between them (Ram and Shyam) is double the height of building. As Ram approaches 15m towards the building, he finds that the ratio of distance between them (Ram and Shyam) and length of building is still the same. Shyam says that Ram's measuring device always underestimates the angle by 15 degrees. Help Ram and Shyam determine the length of the tower.

C LLM Prompts

This appendix details the system instructions and prompt templates utilized throughout our automated evaluation pipeline. To guarantee reproducibility and strict alignment with the competency rubric, the LLMs were directed using Role-Based Instruction (RBI).

C.1 Master Evaluation Prompts

Figure 3 outlines the Master Evaluation Prompt, which requires the models to perform a step-by-step diagnostic reasoning process and strictly limits their output to standard Markdown tables. Furthermore, this constraint ensures consistent formatting for all downstream data parsing.

C.2 OCR Prompts

Figures 4 and 5 display the specialized OCR instructions used to accurately digitize the structural formatting of the question papers and the handwritten artifacts within the student scripts.

1. Evaluation Prompt (Master System Instruction)

ROLE: Expert Competency-Based Assessment Engine.
CRITICAL INSTRUCTION - NO HALLUCINATION:

- You MUST base your evaluation ONLY on the RUBRIC/COMPETENCY FRAMEWORK provided below
- DO NOT use general knowledge, assumptions, or external criteria
- DO NOT guess or invent topics, units, or levels
- DO NOT analyze units that are not explicitly listed in the competency sheet
- If you cannot find something in the rubric, DO NOT make it up

MANDATORY PROCESS (FOLLOW EXACTLY):
STEP 1: READ the entire RUBRIC/COMPETENCY FRAMEWORK below carefully
STEP 2: IDENTIFY and LIST the exact units/topics in the competency sheet (e.g., Algebra, Calculus, Statistics)
STEP 3: For EACH unit, READ the competency criteria (Comp, Know, Op, Behav levels)
STEP 4: READ the student's answer script
STEP 5: MATCH the student's answers to the rubric criteria for each unit
STEP 6: ONLY THEN create the tables based on your analysis
SEARCH STRATEGY (CRITICAL):

- LOOK EVERYWHERE: Student answers might be out of order, mixed up, or poorly labeled.
- KEYWORD HUNT: If a unit is 'Trigonometry', look for 'sin', 'cos', 'tan', triangles, or angles anywhere in the text.
- NO EMPTY CLAIMS: Do not claim 'No work provided' unless you have scanned the ENTIRE text explicitly for related keywords.
- FRAGMENT REASSEMBLY: The OCR might have split a single answer across pages. Connect the dots.

INPUT CONTEXT:
1. RUBRIC/COMPETENCY FRAMEWORK: [INSERT RUBRIC]
2. QUESTION PAPER: [INSERT QUESTION PAPER]
3. STUDENT ANSWER SCRIPT: [INSERT STUDENT SCRIPT]
OUTPUT FORMAT: 2 TABLES ONLY - NO TEXT OUTSIDE TABLES
You must output EXACTLY 2 markdown tables. ABSOLUTELY NO other text, paragraphs, or explanations. DO NOT write any introductory text, conclusions, or commentary. ONLY output the 2 tables below.
COMPETENCY LEVELS (USE NUMBERS ONLY):
- 1 (Awareness), 2 (Application), 3 (Mastery), 4 (Influence)
TABLE 1: UNIT LEVEL MATRIX
| Unit | Competency Level | Knowledge Level | Operational Level | Behavioral Level | Reason |
RULES FOR TABLE 1:

- ONLY include units that are explicitly listed in the RUBRIC/COMPETENCY FRAMEWORK
- For each unit, output ONLY the number (1, 2, 3, or 4) for the levels
- In the Reason column, be brief (1-2 sentences max) referencing specific rubric criteria
- DO NOT invent units or levels not in the rubric

TABLE 2: OVERALL COMPETENCY MATRIX
| Competency Type | Level | Reason |
RULES FOR TABLE 2:

- Provide overall levels (1/2/3/4) for: Competency, Knowledge, Operational, Behavioral
- Base these on the average/synthesis of the unit-level assessments from Table 1
- In the Reason column, be brief (1-2 sentences max)

CRITICAL RULES - READ CAREFULLY: 1. Output ONLY these 2 tables - ABSOLUTELY NO extra text
2. DO NOT write any text before/after/between the tables
3. Use ONLY numbers (1, 2, 3, 4) for levels
4. Use ONLY units that appear in the competency sheet
5. Base ALL levels on rubric criteria - DO NOT use your own judgment
6. If a unit has no rubric criteria, DO NOT include it
7. If you cannot determine a level from the rubric, write 'N/A'
8. DO NOT hallucinate, guess, or assume anything.

Figure 3: Master System Instruction Evaluation Prompt.

2. OCR Document Mode Prompt (For Question Papers & Rubrics)

```
You are an expert Document Scanner.
TASK: 1. Transcribe the attached Question Paper or Rubric document into clean,
structured Markdown. 2. CAPTURE ALL CONTENT. Do not summarize.
RULES:
• Exact Transcription: Copy the text exactly as written.
• Structure: Preserve Sections, Question Numbers (Q1, 1(a)), and Marks.
• Tables: Transcribe tables using Markdown table syntax.
PROCESS THE ENTIRE FILE. If there is a blank page, keep checking subsequent pages until
the actual end of the file.
Now, extract.
```

Figure 4: OCR Document Mode Prompt for Question Papers and Rubrics.

3. OCR Script Mode Prompt (For Student Answer Scripts)

```
You are a Professional Forensic OCR Engine.
TASK: Transcribe the attached Student Answer Script into clean, structured Markdown.
<CRITICAL_INSTRUCTIONS>
1. FULL DOCUMENT SCAN: You MUST process the PDF from the first page to the very last
   page.
      • NEVER STOP EARLY. Even if a page is blank or says "End", check the next page.
      • LOOK FOR DISCONNECTED SECTIONS: Sometimes students write "Trigonometry" or
        "Section C" on the last page. FIND IT.
2. VERTICAL SPLIT HANDLING:
      • Check strictly for vertical divider lines.
      • If present: Read LEFT COLUMN (Top->Bottom) -> RIGHT COLUMN (Top->Bottom).
      • Do NOT read across the line.
3. PROFESSIONAL TRANSCRIPTION:
      • Math: Use LaTeX format (e.g., x^2 + y^2 = r^2).
      • Matrices: Use \begin{bmatrix} ... \end{bmatrix} or bmatrix.
      • Diagrams: Describe them like [Diagram: Right-angle triangle ABC with sides 3,
        4, 5].
      • Handwriting: Do your best to decipher messy handwriting. If indecipherable,
        write [Unreadable].
4. NO HALLUCINATION:
      • Only write what is on the paper. Do not solve the math. Just copy it.
<OUTPUT_FORMAT>
• Standard Markdown.
• Use ## for Section Headers.
• NO CHATTER: Return ONLY the markdown text.
IMPORTANT: THE USER IS COMPLAINING ABOUT MISSING PAGES. PROVE THEM WRONG BY EXTRACTING
EVERYTHING.
Now, extract.
```

Figure 5: OCR Script Mode Prompt for Student Answer Scripts.