Benchmarking Large Language Models for Diagnosing Students' Cognitive Skills from Handwritten Math Work

Anonymous ACL submission

Abstract

Diagnosing students' cognitive skills from their handwritten math work is essential for personalized learning, as such work captures intermediate reasoning beyond final answers. Yet interpreting it remains labor-intensive for teachers, and automating it with LLMs is far from straightforward. Unlike mathematical problem-solving, where LLMs generate solutions themselves, cognitive diagnosis requires models to infer latent reasoning from student work. To systematically investigate this underexplored task, we construct MATHCOG, a benchmark dataset of 3,036 diagnostic verdicts across 639 student responses to 110 math problems, annotated by teachers using TIMSS-grounded cognitive skill checklists with evidential strength labels (*Evident*/*Vague*). Using MATHCOG, we evaluate 17 open and proprietary LLMs and find that (1) all models underperform (F1 $<$ 0.521) regardless of capability, and (2) performance degrades sharply under vague evidence. Error analysis reveals systematic failure patterns: models misattribute vague evidence as evident, over-infer from minimal cues, and hallucinate nonexistent evidence. These findings highlight fundamental limitations in LLMs' ability to reason under implicit evidential conditions, with direct implications for evidence-aware, teacher-in-the-loop designs in LLM-based cognitive diagnosis.

1 Introduction

Diagnosing students' cognitive skills from their problem-solving work is a long-standing goal in mathematics education, as it can reveal where and how students' reasoning breaks down beyond final correctness (Livingston, 2009; Jin et al., 2024). Cognitive skills refer to the mental procedures and knowledge structures that underlie competent performance (Anderson, 1982). In mathematics education, these skills are operationalized through frameworks such as TIMSS (Mullis, 2017), which categorizes them into students' abilities to know, apply, and reason with mathematical concepts and procedures.

Handwritten math work provides rich insight into students' mathematical thinking, capturing intermediate reasoning steps, strategy choices, and partial understanding, which makes it particularly valuable for diagnosing cognitive skills. (Kheong, 1994; Jin et al., 2024). Despite its value, such handwritten work often provides incomplete, implicit, or unevenly expressed evidence of students' reasoning (Henderson et al., 2004; Rahbarnia et al., 2014), making reliable cognitive skill diagnosis inherently challenging. For example, students may omit some steps that could serve as clear evidence of certain skills (e.g., doing mental math and skipping writing out calculation steps); despite the absence of explicit evidence, one can infer that the students know how to calculate if they solved the later steps correctly. Despite recent advances in large language models (LLMs), including multimodal perception (Zhang et al., 2024b,a) and reasoning capabilities (Wang et al., 2023; Hao et al., 2024), their ability to diagnose and infer students' cognitive skills from handwritten work remains largely unexamined. While prior work has examined LLMs' own mathematical problem-solving abilities (Didolkar et al., 2024; Fang et al., 2024; Ahn et al., 2024), our work extends beyond problem solving to examine their ability to diagnose human problem-solving processes.

In this work, we systematically investigate how well existing LLMs diagnose students' cognitive skills in mathematics. Specifically, we focus on the degree of evidential strength in students' handwritten responses. We address the following research questions:

RQ1. How do different LLMs (varying in image input, reasoning, model size, and few-shot prompting) perform in diagnosing students'

cognitive skills?

RQ2. How does the evidential strength of student responses affect LLM's cognitive skill diagnosis performance?

To answer these questions, we (i) constructed MATHCOG, an expert-crafted benchmark dataset designed to evaluate cognitive skill diagnosis. In collaboration with 5 education experts and 15 middle school teachers, we curated 12 middle school math topics and 110 problems, each with 50+ student responses diagnosed by teachers based on a problem-specific diagnostic checklist grounded in the TIMSS cognitive framework (Mullis, 2017). Teachers provided binary judgements (*Yes*/*No*) for each skill and annotated the presence of supporting evidence as *Evident* or *Vague*, yielding 3,036 diagnostic items across 639 student responses in total. Unlike existing benchmarks that rely on synthetic or automatically generated data, all items in MATHCOG are grounded in real student responses and expert-agreed diagnostic judgments, which are inherently difficult to collect at scale in educational settings.

Using MATHCOG, we evaluated 17 closed- and open-source LLMs spanning multiple model families, sizes, and capabilities. Results indicate that current LLMs struggle to reliably diagnose students' cognitive skills (all F1 scores < 0.521), with performance degrading sharply under vague-evidence conditions and frequently misattributing weak evidence as strong. Qualitative error analysis further reveals the model's failure patterns across model types. In particular, models often *misidentify* or *hallucinate* supporting evidence, or *over-infer* students' cognitive skills from incomplete work, showing that errors cascade throughout the diagnostic pipeline from recognizing student responses to interpreting them based on the rubric.

Our contributions are threefold. (1) We introduce MATHCOG, an expert-crafted benchmark dataset for cognitive skill diagnosis. (2) Using MATHCOG, we evaluate 17 large language models and analyze the effects of multimodality, reasoning, model size, and evidential strength on diagnosis performance. (3) We analyze diagnostic error patterns under vague-evidence conditions, identifying recurring failure modes across model types.

2 Related Work

We review existing frameworks for characterizing cognitive skills in learning and assessment, along with prior investigations into LLMs' performance in math-related contexts, and relevant benchmarks, highlighting their current gaps.

2.1 Cognitive Diagnosis in Mathematics Education

Several frameworks have been proposed to characterize cognitive skills in learning and assessment, including Cognitive Diagnostic Models (CDMs) (Leighton and Gierl, 2007), the PISA Assessment and Analytical Framework (OECD, 2023), and the TIMSS assessment framework (Mullis, 2017). Among these, the TIMSS is a comprehensive and math-specific framework for cognitive diagnosis, comprising content and cognitive domains. The cognitive domain, which evaluates knowledge application, is divided into three key areas: Knowing (*recalling* definitions, *recognizing* mathematical entities, *classifying or ordering* quantities, *computing*), Applying (*determining* strategies, *representing* problems, *implementing* solution procedures), and Reasoning (*analyzing*, *justifying*). This research aims to explore whether LLMs can substitute for the mapping to enable scalable and explainable cognitive diagnosis. To our knowledge, this is the first systematic investigation of LLMs in the context of the TIMSS framework.

2.2 LLM Capabilities in Mathematical Tasks

Recent advances have enabled LLMs to achieve remarkable performance in mathematical problem-solving (Guo et al., 2025; OpenAI, 2024), yet they show relative weaknesses in diagnosing student abilities (Macina et al., 2025; Weitekamp et al., 2025). These diagnostic challenges are further complicated by current models' struggles with multimodal inputs such as visual elements and handwritten content (Zhang et al., 2024b; Baral et al., 2025; Liu et al., 2024). While Ma et al. (2025) demonstrated LLMs' potential for cross-domain cognitive diagnosis, existing research has limited systematic exploration of LLMs' diagnostic capabilities for cognitive skills in math specifically. We address this critical gap in the field by providing the first comprehensive evaluation of LLMs' ability to diagnose cognitive skills in mathematical contexts using an established educational framework.

2.3 Benchmarks for Evaluating Mathematical Assessment Tasks

In educational settings, several benchmarks have been proposed to evaluate LLMs' performance in mathematical assessment tasks, including grading, skill recognition, and error analysis. For example, Lucy et al. (2024) introduced MathFish, which aligns 9,900 problems with 385 K–12 standards to assess models' ability to identify targeted mathematical skills and concepts. Automated scoring of open-ended constructed responses in math word problems has also been studied (Hellman et al., 2023). Other benchmarks focus on misconception detection in multiple-choice questions (Kaggle and Eedi, 2020) or handwriting recognition in student work (Baral et al., 2025). While prior research has focused on error detection and scoring, diagnosing complex cognitive skills from students' open-ended handwritten responses remains underexplored and under-resourced. To address this gap, we introduce a novel dataset for cognitive skill diagnosis that enables systematic evaluation of LLMs' ability to interpret implicit reasoning and diagnose students' cognitive skills beyond final answers.

3 Dataset: MATHCOG

To evaluate LLMs' performance on cognitive skill diagnosis, we created a benchmark dataset comprising secondary-school math problems, handwritten student responses, diagnostic checklists, and teacher-generated verdicts (Table 1). The math problem and student response data are from AI-Hub[1] , which provides OCR transcriptions of handwritten work. Each topic includes multiple isomorphic problems that share the same problem-solving procedures but differ in numerical values (e.g., "Solve $x^2 + 2x - 3 = 0$" and "Solve $x^2 + x - 2 = 0$") (Reed et al., 1990; Morrison et al., 2015). From this data, we focused on topics from grades 7–9, where problems are sufficiently complex to elicit students' cognitive processes. We further excluded topics with fewer than 50 student responses to ensure sufficient data for reliable evaluation. Through this filtering process, the resulting dataset contains 15 topics, 137 problems, and 796 student responses.

[1]This research used datasets from 'The Open AI Dataset Project (AI-Hub, S. Korea)'. All data information can be accessed through AIHub.

3.1 Diagnostic Checklist

To systematically diagnose cognitive skills reflected in student responses, we developed a diagnostic checklist commonly applicable to isomorphic problems. Each checklist consists of binary question items mapped to one of the 15 cognitive skills defined in the TIMSS 2019 assessment framework. The checklist items were adapted from TIMSS skill descriptions and refined through expert review by five mathematics curriculum and evaluation experts with PhD degrees in education and practical experience in making math assessment guidelines. Experts gave feedback on the clarity, granularity, and validity of the checklist items. The experts also commented that cognitive skills could be reliably assessed from the given problem types. They noted that the problems predominantly require numerical computation and application of known procedures, which makes them well-suited for assessing *knowing* and *applying* skills. In contrast, higher-order *reasoning* skills (e.g., justifying, analyzing, generalizing) are less consistently observable from students' written responses in this dataset. Based on this observation, we intentionally scoped our checklist to the *knowing* and *applying* domains to ensure reliable and evidence-grounded diagnosis. The checklists were refined through two iterative rounds, with each version independently reviewed by two experts.

3.2 Teacher-generated Verdict

We recruited 15 middle school math teachers to evaluate 796 student responses based on predefined diagnostic checklists. Teachers had an average of 6.1 years of teaching experience (SD = 4.3; min = 2.5, max = 20). Each check item was assessed along two dimensions: evidential strength (*Evident*/*Vague*) and correctness (*Yes*/*No*)."*Evident*" meant there was clear evidence to support the judgment, whereas "*Vague*" signified insufficient evidence. A "*Yes*" response indicated that the student fully demonstrated the cognitive action specified, while a "*No*" indicated otherwise, including partially demonstrated responses. To account for the subjectivity in cognitive diagnosis, each student's response was evaluated by three teachers with overlapping assignments, allowing us to measure inter-rater agreement (see Table 6 in Appendix. Following the threshold established in Graham et al. (2012), we excluded topics with agreement below 70%, resulting in a final dataset of 12 topics, 110

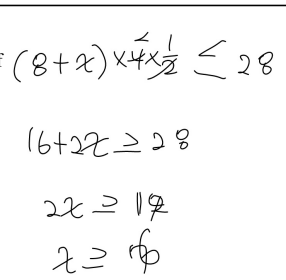
Problem	Student Response	Diagnostic Checklist	Verdict
There is a trapezoid with a lower side length of 8 cm and a height of $4cm$. If the area of this trapezoid is not less than 28 cm^2, find how much more cm the length of the upper side of the trapezoid must be.		**Recall**: Does the student remember the formula for the area of a shape correctly?	Evident Yes
		Compute: Has the student calculated the linear and constant terms correctly?	Vague No
		Determine: Did the student know the need to set up an equation and then solve it to find the range of solutions that meet the conditions?	Evident Yes
		Represent: Has the given situation been expressed correctly?	Evident No
		Implement: When simplifying an expression, does the student keep the expression correct by performing the same operation on both sides?	Vague Yes

Table 1: A sample from MATHCOG. Each data point is composed of a math problem, student response, relevant diagnostic checklist, and verdict for each check item.

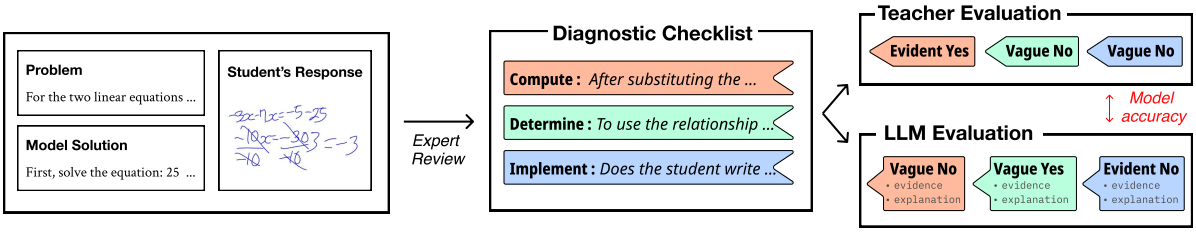


Figure 1: Diagnostic pipeline of MATHCOG. Each instance includes a math problem and student response from an open dataset with a TIMSS-aligned diagnostic checklist developed with mathematics curriculum experts. Math teachers evaluate each checklist item for correctness and evidential strength, and LLM predictions are compared with these teacher verdicts.

problems, and 639 student responses. As a result, the benchmark focuses on diagnostic cases with stable expert agreement, providing a reliable basis for evaluating model performance. These topics cover three-fourths of the content domains defined in TIMSS 2019 (see Table 5 in Appendix), and the dataset remains reasonably sized given the substantial human effort required for expert annotation.

4 Experimental Setting

Using MATHCOG as a benchmark dataset, we evaluated a diverse set of LLMs with varying input modalities, reasoning capabilities, model sizes, and prompting strategies to address RQ1, and analyzed how the evidential strength of student responses affects diagnosis performance to address RQ2.

Prompting. We designed our prompting protocol rigorously to ensure controlled and reproducible evaluation. LLMs were instructed to evaluate each diagnostic check item, given a math problem, its solution, a student's response, and a diagnostic checklist. Student's response image inputs were provided as OCR transcriptions, with mathematical formulas and visual cues (e.g., strikethroughs) represented in LaTeX. For multimodal settings, we additionally supplied images of student responses. Since the original inputs were in Korean, we machine-translated them into English to prevent possible performance degradation due to language (Achiam et al., 2023); we used Google Translate API for batch translation and manually verified them. We employed Chain-of-Thought prompting (Liu et al., 2024) to guide LLMs to systematically address each check item by first restating its content, identifying relevant evidence, providing an explanation, and delivering a final verdict. The verdict followed one of the four categories used by teachers in MATHCOG. To further examine the robustness of our prompt design, we additionally evaluated five representative models under few-shot prompting using two exemplars covering all four verdict types (Table 8). The few-shot experiment results and full system and user prompts are provided in the Appendix.

Models. We evaluated 17 open and proprietary LLMs from multiple model families, selected to analyze the effects of multimodality, reasoning capability, and model size. For reasoning effort level, SOTA models with controllable reasoning budgets were each evaluated under both low and high effort conditions. For multimodality, selected models were tested under both text-only and image-

augmented conditions. For model size, models were grouped into large, medium, and small categories.

Metrics. We evaluated LLM outputs against teacher-provided ground-truth diagnostic labels. Overall diagnosis performance was assessed using **macro F1 score** and **accuracy** over the four verdict categories. We report macro F1 to mitigate the effect of label imbalance across verdicts (see Table 7 in Appendix). To address **RQ2**, we analyzed how the *evidential strength* of student responses affects diagnosis performance. To formalize this analysis, we treat each diagnosis instance as a (*student response*, *cognitive skill*) pair, indexed by i. For each instance i, the ground-truth labels consist of a skill judgment $y_i \in \{\text{Yes}, \text{No}\}$ and evidential strength $e_i \in \{\text{Evident}, \text{Vague}\}$, and the corresponding model predictions $\hat{y}_i$ and $\hat{e}_i$. To characterize how models handle evidential strength, we define two complementary metrics that capture distinct failure modes. Prior work has shown that model-generated explanations can be plausible yet fail to faithfully reflect their actual reasoning basis (Jacovi and Goldberg, 2020; Turpin et al., 2023; Lyu et al., 2024). In educational diagnosis, an analogous failure occurs when a model falsely attributes evidential support to its verdicts, either by overstating vague evidence, or by asserting evidence for incorrect diagnoses. We define two metrics to quantify each form of this failure. First, we define **evidence over-attribution** (**OverAttr**) as the frequency with which a model assigns "Evident" to cases where the ground-truth evidential strength is "Vague":

$$\text{OverAttr} = \frac{|\{i \mid e_i = \text{Vague} \wedge \hat{e}_i = \text{Evident}\}|}{|\{i \mid e_i = \text{Vague}\}|} \quad (1)$$

This metric reflects the model's tendency to overstate evidential support under vague evidence cases, independent of diagnosis correctness. Second, we define **evidence false-attribution** (**FalseAttr**) as the frequency with which a model assigns Evident to incorrect diagnoses:

$$\text{FalseAttr} = \frac{|\{i \mid \hat{y}_i \neq y_i \wedge \hat{e}_i = \text{Evident}\}|}{|\{i \mid \hat{y}_i \neq y_i\}|} \quad (2)$$

This metric captures a particularly misleading failure mode, where erroneous diagnoses are accompanied by strong evidential claims.

5 Results

This section reports results addressing our research questions.

5.1 RQ1. How do different LLMs perform in diagnosing students' cognitive skills?

Overall, all evaluated LLMs showed limited performance in cognitive skill diagnosis, with most macro F1 scores below 0.5 (Figure 7, Table 2). The only model to exceed F1 of 0.5 was `Gemini-3.1-Pro`, which achieved F1 scores of 0.509 and 0.521 under low and high reasoning effort conditions, respectively, with the high-effort setting also yielding the highest accuracy (0.778) across all evaluated conditions. While overall accuracy was relatively moderate ($M = .683$, $SD = .054$), it obscures important diagnostic failures. Models frequently over-attributed evidential strength (OverAttr; $M = .544$, $SD = .112$) and falsely attributed evidence (FalseAttr; $M = .579$, $SD = .118$), often asserting strong evidence even when diagnoses were incorrect. Notably, even the best-performing models in terms of F1 score and accuracy showed high OverAttr and FalseAttr values, suggesting that performance gains often co-occur with unreliable evidential judgement.

Analysis of skill-specific performance reveals further insights into the limitations of current LLMs. **When averaged across all models, no skill category achieved an F1 score above 0.5**, indicating that fine-grained diagnosis remains challenging even at the individual skill level (Figure 4, Figure 9). While `Gemini-3.1-Pro` occasionally exceeded this threshold on individual skills such as *Implement* and *Recognize*, this was not consistent across models. Despite this overall limitation, we observed a performance gap between the *Knowing* and *Applying* cognitive skills. Contrary to our expectation that *Knowing* skills (e.g., *Recall*, *Recognize*) would be easier due to their surface-level nature, models performed better on *Applying* skills in both F1 score ($t = 6.09$, $p < .001$) and accuracy ($t = 4.50$, $p < .001$). However, this performance gain came with increased evidential errors. *Applying* skills exhibited higher evidence over-attribution ($t = 5.42$, $p < .001$) and false-attribution ($t = 5.85$, $p < .001$) than *Knowing* skills, indicating a tendency to overstate evidential support when diagnosing higher-level cognitive skills. Notably, *Recall* stood out as a particularly difficult skill, showing the lowest 1-OverAttr and

Family	R-Low				R-High				+Image				Large				Medium				Small			
	F1	Acc	OA	FA	F1	Acc	OA	FA	F1	Acc	OA	FA	F1	Acc	OA	FA	F1	Acc	OA	FA	F1	Acc	OA	FA
GPT	*GPT-5.2*				*GPT-5.2*				*GPT-4o*				*GPT-4o*								*GPT-4o-mini*			
	.428	.664	**.410**	.482	**.434**	.667	**.326**	**.471**	**.448**	**.743**	**.656**	**.681**	.412	.672	**.648**	**.658**	-	-	-	-	**.352**	.622	**.767**	.784
Claude	*Claude-4.6-Sonnet*				*Claude-4.6-Sonnet*				*Claude-3.5-Sonnet*				*Claude3.5-Sonnet*											
	.432	**.646**	.493	**.570**	.459	**.662**	.419	.586	**.413**	**.691**	.626	.559	.416	.656	.551	.574	-	-	-	-	-	-	-	-
Gemini	*Gemini-3.1-pro*				*Gemini-3.1-pro*				*Gemini-1.5-Flash*				*Gemmini-1.5-Pro*				*Gemmini-1.5-Flash*				*Gemmini-1.5-Flash-8b*			
	.509	**.775**	.480	**.628**	**.521**	**.778**	.454	.628	.429	.702	**.471**	**.471**	**.440**	.706	.520	.635	**.432**	.679	.502	.465	.348	**.559**	**.537**	**.525**
DeepSeek					*DeepSeek-R1*								*DeepSeek-V3*											
	-	-	-	-	.442	.773	.767	**.862**	-	-	-	-	.417	**.714**	.612	.595	-	-	-	-	-	-	-	-
Llama	*Llama-4-Maverick*				*Llama-4-Maverick*								*Llama-3.1-405B*				*Llama-3.3-70B*				*Llama-3.1-8B*			
	.468	.758	**.551**	.623	.478	.763	**.555**	.622	-	-	-	-	**.385**	**.624**	**.467**	**.307**	**.397**	**.635**	**.436**	**.409**	**.323**	.654	.705	.626
Qwen																	*Qwen-2.5-72B*				*Qwen-2.5-7B*			
	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	.416	**.711**	**.568**	**.500**	.339	**.705**	.758	**.798**

Table 2: Performance of LLMs on cognitive skill diagnosis, grouped by model family. **OA** and **FA** denote OverAttr and FalseAttr (lower is better). **R-Low** and **R-High** reflect two reasoning effort levels of the latest frontier models (e.g., thinking budget or inference compute); **+Image** denotes models prompted with student response images in addition to OCR text; and **Large**, **Medium**, and **Small** correspond to model size tiers based on parameter count. Specific models are indicated in italics above each value group. **Blue**: best; **Red**: worst value per column.

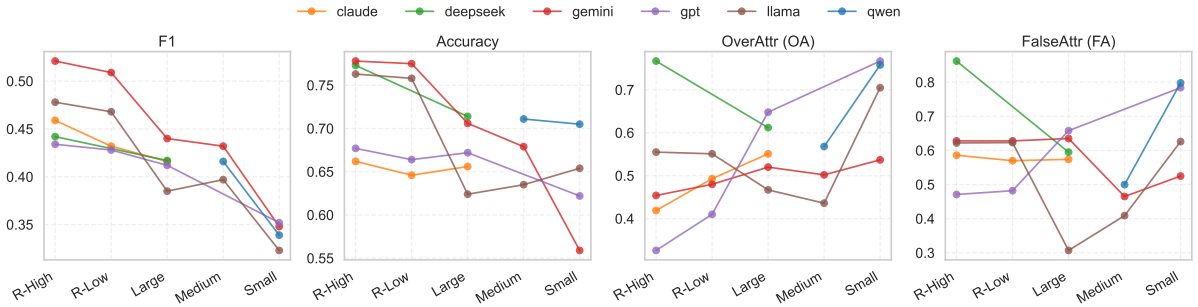


Figure 2: Performance of six LLM families across five evaluation conditions. R-High and R-Low denote high and low reasoning effort levels of the latest frontier models; Large, Medium, and Small refer to model size tiers. Missing points indicate no available model for that condition.

1-FalseAttr values among all skills (Figure 4), suggesting that models were most prone to misattributing evidence when diagnosing this higher-level cognitive skill.

We further examined the impact of **multimodality**, **reasoning effort level**, and **model size** as factors that may influence diagnostic performance (Table 2). **Multimodal input provided modest but inconsistent benefits (Figure 3).** Models supporting image input consistently outperformed their text-only counterparts in accuracy, indicating that access to handwritten layouts and visual cues helps mitigate errors introduced by OCR transcription. However, improvements in F1 score were inconsistent, suggesting that visual input can sometimes introduce noise and does not uniformly improve fine-grained skill diagnosis. **Higher reasoning effort generally yielded better diagnostic performance.** Among sota models evaluated under both reasoning effort conditions, high effort consistently outperformed low effort in F1 and accuracy across model families (Figure 2). Evidence over-attribution (OverAttr) was also somewhat lower under higher reasoning effort, while false-attribution (FalseAttr) remained comparable across the two conditions. These results suggest that increased reasoning compute offers modest but consistent benefits for cognitive diagnosis, though the gains were not large enough to substantially close the performance gap observed across all evaluated models. **Model size showed a strong positive relationship with F1 performance** (Spearman's $\rho = .716, p = .009$), indicating that larger models more robustly interpret student responses (Figure 2). However, size was not significantly associated with accuracy ($\rho = .399, p = .199$), OverAttr ($\rho = -.417, p = .177$), or FalseAttr ($\rho = -.246, p = .441$), though smaller models tended to show higher rates of overconfident evidential claims. These results suggest that scaling up model size improves diagnostic judgment but does not guarantee more reliable evidential reasoning.

Figure 3: Effect of image input on diagnostic performance across three multimodal models (`Claude-3.5-Sonnet`, `Gemini-1.5-Flash`, `GPT-4o`).

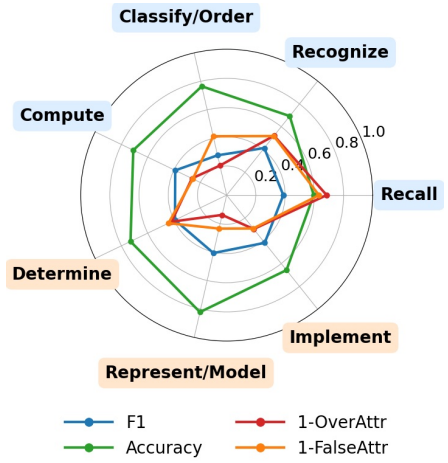


Figure 4: Radar chart showing model performance across seven cognitive skills. The metrics include F1, Accuracy, 1–OverAttr, and 1–FalseAttr (higher is better). Skills are grouped by cognitive domain: Knowing (blue) and Applying (orange).

Taken together, these findings suggest that model capability and scale alone are insufficient to ensure reliable cognitive skill diagnosis.

5.2 RQ2. How does the evidential strength of student responses affect LLM's cognitive skill diagnosis performance?

Our results reveal a **strong dependency of LLM diagnostic performance on the evidential strength** of student responses. Across all models, both F1 score ($t = 64.60$, $p < .001$) and accuracy ($t = 17.80$, $p < .001$) were substantially higher when student responses contained *evident* evidence for a cognitive skill than when evidence was *vague* (Fig. 5). The magnitude of this gap was substantial (mean $\Delta F1 = .51$; mean $\Delta Acc = .41$), indicating that current LLMs rely heavily on explicit evidence and exhibit limited robustness when reasoning under weak or implicit evidential conditions.

Beyond degraded performance, **models also exhibited over-attribution bias under vague evidence.** Specifically, they frequently labeled responses as *Evident* even when the ground-truth evidential strength was *Vague*, resulting in consistently high **OverAttr** values. More concerningly, models also showed a strong tendency toward **FalseAttr**, in which they asserted *evident* evidence despite arriving at incorrect cognitive skill diagnoses. In such cases, models produced unsupported explanations that appeared plausible yet were not grounded in students' actual responses. Notably, these two tendencies were strongly correlated across models (Spearman's $\rho = .837$, $p < .001$), indicating that models prone to over-attributing evidential strength are also more likely to produce incorrect diagnoses with asserted evidence. This behavior is particularly problematic for educational use, as it can mislead teachers and students by presenting incorrect diagnoses with seemingly well-justified rationales (Kim et al., 2025).

To better understand how such errors arise, we conducted a qualitative error analysis of model outputs under incorrectly diagnosed vague evidence cases. We examined all components generated in the models' CoT responses, including the identified evidence, explanations, and final verdicts, focusing on how the models misinterpreted students' answers, generated unsupported reasoning, and ultimately reached incorrect diagnoses. Three authors collaboratively analyzed 372 cases (13.2%) sampled from a total of 2,816 error instances, covering 17 model responses, and iteratively derived five recurring error types with substantial inter-rater reliability (Fleiss' $\kappa = 0.74$). The remaining cases were then independently labeled using the agreed error

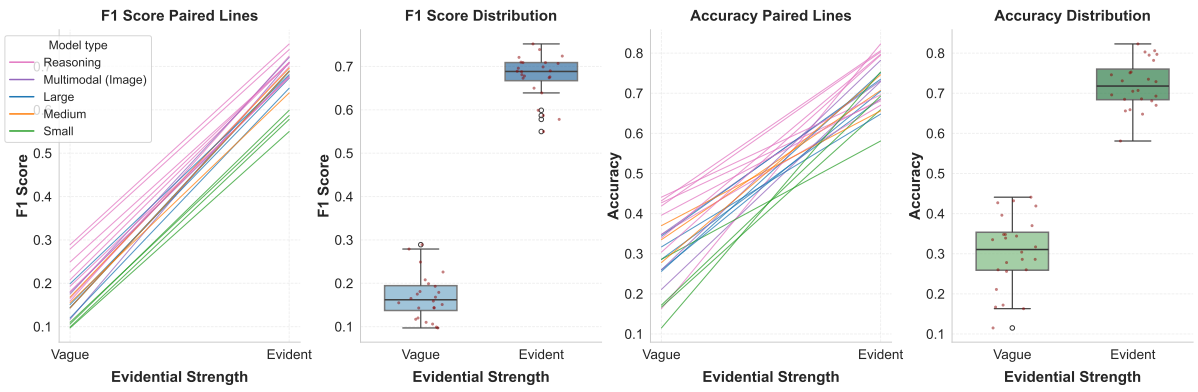


Figure 5: Effect of evidential strength on LLM diagnostic performance across all evaluated models. Paired line plots (left) show per-model changes in F1 and accuracy from vague to evident evidence conditions, with each line representing one model colored by model type. Box plots (right) summarize the overall distribution of F1 and accuracy under each condition.

Error code	Description	Count (%)
E1. Evidence Misidentification	Incorrectly identifying or relying on insufficient evidence from the student response.	**659 (23.40%)**
E2. Rubric Misinterpretation	Misunderstanding the diagnostic rubric or evaluation criteria.	**279 (9.91%)**
E3. Over-Inference	Inferring unstated reasoning beyond the information provided in the student response.	**942 (33.45%)**
E4. Consistency Issue	Inconsistencies between evidence, explanation, and final verdict.	**379 (13.46%)**
E5. Hallucination	Introducing content or reasoning not present in the student response.	**557 (19.78%)**

Table 3: Error types and their distribution identified from qualitative analysis of LLM diagnostic failures under vague evidence conditions.

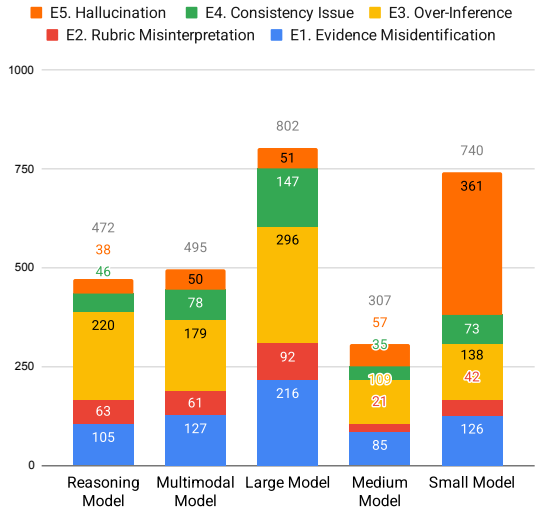


Figure 6: Error type distribution across model categories. Each bar represents the total number of diagnostic errors under vague evidence conditions, broken down by each error type.

taxonomy. Table 3 summarizes the distribution of error types across model categories. Errors related to *over-inference* were most prevalent (33.5%), followed by *evidence misidentification* (23.4%) and *hallucination* (19.8%). We further observed systematic differences across model types (Figure 6): smaller models exhibited a higher proportion of *hallucination* errors (48.78%), whereas reasoning-oriented models committed *over-inferred* more frequently (46.61%) beyond the information explicitly provided in student responses.

6 Conclusion

We introduced MATHCOG, a benchmark dataset comprising 3,036 diagnostic verdicts across 639 student responses, annotated by teachers using a TIMSS-grounded cognitive framework. Evaluating 17 LLMs, we find that diagnostic performance remains limited across all model types and cognitive skill categories, with performance degrading sharply under vague evidence conditions. Error analysis reveals systematic failure patterns that vary across model types. These findings suggest that the core challenge lies in the inherent difficulty of interpreting implicit diagnostic evidence, rather than in model capability alone. Responsible deployment of LLMs in educational assessment will require task designs that elicit more explicit student reasoning and teacher-in-the-loop workflows that support critical interpretation of model outputs.

Limitations

This work has several limitations. First, the MathCog benchmark is constructed from a specific educational dataset consisting of middle-school mathematics responses, which may limit the generalizability of our findings to other subjects, grade levels, or educational contexts. Second, the cognitive skills in this work are defined using TIMSS-based categories (*know*, *apply*, *reason*), which provide an operational taxonomy but may simplify the diversity and complexity of students' real problem-solving processes. Third, cognitive skill diagnosis is inherently subjective, and teacher annotations may reflect individual interpretations of student work. Although we mitigated this by requiring three independent teacher judgments per item and excluding topics with inter-rater agreement below 70%, residual subjectivity in the ground-truth labels may still affect evaluation reliability. Finally, while we explored multiple prompting strategies including Chain-of-Thought and few-shot prompting, our evaluation is limited to a specific task formulation; alternative formulations or more specialized prompting approaches may lead to different model behaviors and diagnostic outcomes.

Ethical Impact

This work analyzes students' mathematical problem-solving responses to study whether large language models can diagnose cognitive skills from partial evidence. The dataset consists of anonymized student work and does not contain personally identifiable information. A potential risk is that LLM-based diagnostic systems may produce incorrect or overconfident interpretations of reasoning if used without appropriate verification. By systematically identifying such failure patterns, our benchmark aims to support the development and evaluation of more reliable reasoning analysis systems.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges. *arXiv preprint arXiv:2402.00157*.

John R Anderson. 1982. Acquisition of cognitive skill. *Psychological review*, 89(4):369.

Sami Baral, Li Lucy, Ryan Knight, Alice Ng, Luca Soldaini, Neil T Heffernan, and Kyle Lo. 2025. Drawedumath: Evaluating vision language models with expert-annotated students' hand-drawn math images. *arXiv preprint arXiv:2501.14877*.

Aniket Didolkar, Anirudh Goyal, Nan Rosemary Ke, Siyuan Guo, Michal Valko, Timothy Lillicrap, Danilo Jimenez Rezende, Yoshua Bengio, Michael C Mozer, and Sanjeev Arora. 2024. Metacognitive capabilities of llms: An exploration in mathematical problem solving. *Advances in Neural Information Processing Systems*, 37:19783–19812.

Meng Fang, Xiangpeng Wan, Fei Lu, Fei Xing, and Kai Zou. 2024. Mathodyssey: Benchmarking mathematical problem-solving skills in large language models using odyssey math data. *arXiv preprint arXiv:2406.18321*.

Matthew Graham, Anthony Milanowski, and Jackson Miller. 2012. Measuring and promoting inter-rater agreement of teacher and principal performance ratings. *Online Submission*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Shibo Hao, Yi Gu, Haotian Luo, Tianyang Liu, Xiyan Shao, Xinyuan Wang, Shuhua Xie, Haodi Ma, Adithya Samavedhi, Qiyue Gao, and 1 others. 2024. Llm reasoners: New evaluation, library, and analysis of step-by-step reasoning with large language models. *arXiv preprint arXiv:2404.05221*.

Scott Hellman, Alejandro Andrade, and Kyle Habermehl. 2023. Scalable and explainable automated scoring for open-ended constructed response math word problems. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 137–147.

Charles Henderson, Edit Yerushalmi, Vince H Kuo, Patricia Heller, and Kenneth Heller. 2004. Grading student problem solutions: The challenge of sending a consistent message. *American Journal of Physics*, 72(2):164–169.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4198–4205.

Hyoungwook Jin, Yoonsu Kim, Yeon Su Park, Bekzat Tilekbay, Jinho Son, and Juho Kim. 2024. Using large language models to diagnose math problem-solving skills at scale. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, pages 471–475.

Kaggle and Eedi. 2020. Eedi - mining misconceptions in mathematics dataset. Accessed: 2025-05-21.

Fong Ho Kheong. 1994. Information processing taxonomy (ipt): An alternative technique for assessing mathematical problem-solving.

Sunnie SY Kim, Jennifer Wortman Vaughan, Q Vera Liao, Tania Lombrozo, and Olga Russakovsky. 2025. Fostering appropriate reliance on large language models: The role of explanations, sources, and inconsistencies. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–19.

Jacqueline Leighton and Mark Gierl. 2007. *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge University Press.

Tianyi Liu, Julia Chatain, Laura Kobel-Keller, Gerd Kortemeyer, Thomas Willwacher, and Mrinmaya Sachan. 2024. Ai-assisted automated short answer grading of handwritten university level mathematics exams. *arXiv preprint arXiv:2408.11728*.

Samuel A Livingston. 2009. Constructed-response test questions: Why we use them; how we score them. r&d connections. number 11. *Educational Testing Service*.

Li Lucy, Tal August, Rose E Wang, Luca Soldaini, Courtney Allison, and Kyle Lo. 2024. Mathfish: Evaluating language model math reasoning via grounding in educational curricula. *arXiv preprint arXiv:2408.04226*.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2024. Towards faithful model explanation in NLP: A survey. *Computational Linguistics*, 50(2):657–723.

Haiping Ma, Changqian Wang, Siyu Song, Shangshang Yang, Limiao Zhang, and Xingyi Zhang. 2025. Large language models are zero-shot cross-domain diagnosticians in cognitive diagnosis. *Frontiers of Digital Education*, 2(2):1–14.

Jakub Macina, Nico Daheim, Ido Hakimi, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2025. Mathtutorbench: A benchmark for measuring open-ended pedagogical capabilities of llm tutors. *arXiv preprint arXiv:2502.18940*.

Briana B Morrison, Lauren E Margulieux, and Mark Guzdial. 2015. Subgoals, context, and worked examples in learning computing problem solving. In *Proceedings of the eleventh annual international conference on international computing education research*, pages 21–29.

Ina V.S. Mullis. 2017. Timss 2019 assessment frameworks: Introduction. Monitoring Trends in Mathematics and Science Achievement.

OECD. 2023. *PISA 2022 Assessment and Analytical Framework*. OECD Publishing, Paris.

OpenAI. 2024. Learning to reason with llms.

Freydoon Rahbarnia, Salehe Hamedian, and Farzad Radmehr. 2014. A study on the relationship between multiple intelligences and mathematical problem solving based on revised bloom taxonomy. *Journal of Interdisciplinary Mathematics*, 17(2):109–134.

Stephen K Reed, Carolyn C Ackinclose, and Audrey A Voss. 1990. Selecting analogous problems: Similarity versus inclusiveness. *Memory & cognition*, 18(1):83–98.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965.

Boshi Wang, Xiang Yue, and Huan Sun. 2023. Can chatgpt defend its belief in truth? evaluating llm reasoning via debate. *arXiv preprint arXiv:2305.13160*.

Daniel Weitekamp, Momin N Siddiqui, and Christopher J MacLellan. 2025. Tutorgym: A testbed for evaluating ai agents as tutors and students. *arXiv preprint arXiv:2505.01563*.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024a. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, and 1 others. 2024b. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer.

A Appendix

A.1 Prompts

The blue text represents the programmatically filled arguments, and the orange text represents LLM-generated output.

A.1.1 System Prompt

```
# **Task Description**
You are a middle school math teacher tasked
with evaluating students' mathematical thinking
skills based on their responses to math
problems. Your goal is to analyze a given
student's response and determine whether they
exhibit specific cognitive skills in solving
the problem. Your evaluation must be **strict**
and **evidence-based**, meaning that every
verdict must be backed by direct evidence from
the response. If no clear evidence exists, do
not assume correctness.

## **Evaluation Categories**
For each thinking skill in the checklist, you
must classify the student's performance into
one of the following categories:

- **Evident Yes**: The student's response
provides clear and explicit evidence that the
check item is met. A direct quote from the
response can confirm this.
- **Vague Yes**: The response suggests that the
check item might be satisfied, but no specific
part of the response directly proves it.
- **Evident No**: The response explicitly
contradicts or fails to meet the check item,
with clear evidence demonstrating the error or
omission.
- **Vague No**: The response does not appear to
satisfy the check item, but there is no direct
evidence confirming whether the student
considered it or not.

## **Input Format**
You will receive the following data:
- **Problem**: A math problem given to a
student.
- **Answer**: The correct step-by-step solution.
- **Response**: The student's response to the
problem.
- **Check Items**: A set of specific skills to
evaluate.

## **Output Format**
Return a valid JSON object structured as
follows:
```json
{
    "skills": [
        {
            "checkItem": "<Check Item's [Label]
            and the Following Question>",
            "evidence": "<Directly Quoted Part
            of Response>",
            "explanation": "<Explanation About
            Why the Evidence Supports the
            Verdict>",
            "verdict": "Evident Yes" | "Vague
            Yes" | "Evident No" | "Vague No"
        }
    ]
}
```
```

A.1.2 User Prompt

```
# **Task**
student responses to math problems, extract
direct evidence, and strictly classify thinking
skills according to the given categories.

Return a valid JSON object structured as
follows:
```json
{
    "skills": [
        {
            "checkItem": "<Check Item's [Label]
            and the Following Question>",
            "evidence": "<Directly Quoted Part
            of Response>",
            "explanation": "<Explanation About
            Why the Evidence Supports the
            Verdict>",
            "verdict": "Evident Yes" | "Vague
            Yes" | "Evident No" | "Vague No"
        }
    ]
}
```

Problem:
There is a two-digit natural number whose tens
digit is 1. If the number that changes the tens
and ones digits of this natural number is 9 less
than 5 times the first number, find the first
number.

Answer:
If the number in the ones place is x, this
natural number is $10 + x$. The number where the
tens digit and the ones digit are swapped is
$10x + 1$ because the tens digit is x and the ones
digit is 1. The changed number is 9 smaller than
5 times the first number, so
$10x + 1 = 5(10 + x) - 9$ $10x + 1 = 5x + 41$ $5x = 40$
$x = 8$ Therefore, the first number is 18.

Response:
1 x x 1 $10x + 1 = 5(10 + x) - 9$
$10x + 1 = 50 + 5x - 9$ 41 $5x = 40$ $x = 8$

Check Items:
- [recognize_0] Is the student aware that the
relationship between two numbers switched can be
expressed as
addition/subtraction/multiplication?
- [recognize_1] Has the student figured out all
the elements necessary to solve the problem
(original numbers, swapped numbers, difference or
multiple relationships, etc.)?
- [compute_0] Has the student performed algebraic
calculations correctly, such as expanding
expressions and organizing similar terms?
- [determine_0] Has the student chosen the
appropriate approach to solving the problem
(e.g., establishing a linear equation)?
- [represent_0] Are two-digit numbers expressed
correctly in the form 10x + y (or equivalent)?
- [implement_0] Does the student proceed
consistently with the process of expanding the
equation and moving terms to find x (or y)
```

```
according to the planned solution strategy?

**Output:**
{
    "skills": [
        {
            "checkItem": "[recognize_0] Is the
            student aware that the relationship
            between two numbers switched can be
            expressed as
            addition/subtraction/multiplication?",
            "evidence": "10 x+1=5(10+x)-9",
            "explanation": "The student sets up
            the equation 10x+1=5(10+x)-9, which
            shows they understand the relationship
            between the original and swapped
            numbers involves subtraction and
            multiplication.",
            "verdict": "Evident Yes"
        }, {
            "checkItem": "[recognize_1] Has the
            student figured out all the elements
            necessary to solve the problem
            (original numbers, swapped numbers,
            difference or multiple relationships,
            etc.)?",
            "evidence": "10 x+1=5(10+x)-9",
            "explanation": "The student uses the
            correct equation to represent the
            relationship between the original and
            swapped numbers, indicating they have
            identified the necessary elements.",
            "verdict": "Evident Yes"
        },
        ...
    ]
}
```

A.2 Data Collection

A.2.1 Instructions Given to Education Experts and Teachers

The data collection procedures were approved by the Institutional Review Board (IRB). Prior to participation, informed consent was obtained from all participants, including both the education experts and the middle school teachers. The guideline materials used in this study are publicly available on the Open Science Framework (OSF). The materials can be accessed at: https://osf.io/67kz9/overview?view_only=430e88aa45cf493db06f573dfb3eaae2. These materials document the procedures and criteria used during the data collection process.

A.2.2 Recruitment and Payment

Education Experts for Checklist Refinement: We recruited five education experts who work at the Korea Institute for Curriculum and Evaluation (KICE)[2] via cold email invitations. Given their expertise in curriculum design and assessment, each expert received 300,000 KRW (approximately USD 200) as compensation for approximately three hours of participation.

[2] https://www.kice.re.kr/main.do?s=english

Middle School Teachers for Cognitive Diagnosis: We recruited 15 middle school teachers through advertisements posted in online teacher communities in Korea. Each teacher received 150,000 KRW (approximately USD 100) as compensation for approximately three hours of participation.

A.3 MathCog Dataset Details

This section provides supplementary details of the MATHCOG dataset, including the diagnostic workflow, cognitive skill taxonomy, inter-rater reliability of teachers' diagnoses, dataset composition, and representative failure cases of model diagnoses.

A.3.1 TIMSS Cognitive Skill Framework

The MATHCOG dataset is grounded in the TIMSS cognitive framework, which categorizes mathematical thinking into the domains of *knowing*, *applying*, and *reasoning*. Table 4 provides the detailed checklist items used to operationalize each cognitive skill in our diagnostic process.

A.3.2 Problem Topics

The dataset covers a diverse set of middle school mathematics topics to ensure coverage of different cognitive processes and problem types. Table 5 lists the full set of problem topics included in MATHCOG.

A.3.3 Inter-Rater Agreement

To ensure the reliability of cognitive skill annotations, we measured agreement among teachers on diagnostic labels. Table 6 reports the inter-rater agreement statistics for the annotated dataset.

A.3.4 Dataset Distribution

We analyze the distribution of diagnostic labels across cognitive skills and evidence conditions to characterize the dataset composition. Table 7 presents detailed statistics of label distributions in MATHCOG.

A.4 Experiment Result Details

A.4.1 Few-Shot Prompting Results

Table 8 reports the diagnostic performance of five representative models under few-shot prompting, compared to their zero-shot counterparts. Few-shot prompting did not consistently improve F1 or accuracy, and in several cases led to decreased performance. However, few-shot examples did reduce tendencies toward evidence over-attribution

Domains	Cognitive Skills	Description
Knowing	Recall	Recall definitions, terminology, number properties, units of measurement, geometric properties, and notation (e.g., $a \times b = ab, a + a + a = 3a$).
	Recognize	Recognize numbers, expressions, quantities, and shapes. Recognize entities that are mathematically equivalent (e.g., equivalent familiar fractions, decimals, and percents; different orientations of simple geometric figures).
	Classify/Order	Classify numbers, expressions, quantities, and shapes by common properties.
	Compute	Carry out algorithmic procedures for $+, -, \times, \div$ or a combination of these with whole numbers, fractions, decimals, and integers. Carry out straightforward algebraic procedures.
	Retrieve	Retrieve information from graphs, tables, texts, or other sources.
	Measure	Use measuring instruments; and choose appropriate units of measurement.
Applying	Determine	Determine efficient/appropriate operations, strategies, and tools for solving problems for which there are commonly used methods of solution.
	Represent/Model	Display data in tables or graphs; create equations, inequalities, geometric figures, or diagrams that model problem situations; and generate equivalent representations for a given mathematical entity or relationship.
	Implement	Implement strategies and operations to solve problems involving familiar mathematical concepts and procedures.
Reasoning	Analyze	Determine, describe, or use relationships among numbers, expressions, quantities, and shapes.
	Integrate/Synthesize	Link different elements of knowledge, related representations, and procedures to solve problems.
	Evaluate	Evaluate alternative problem solving strategies and solutions.
	Draw conclusions	Make valid inferences on the basis of information and evidence.
	Generalize	Make statements that represent relationships in more general and more widely applicable terms.
	Justify	Provide mathematical arguments to support a strategy or solution.

Table 4: Fifteen cognitive skills and their descriptions defined in the TIMSS 2019 framework.

and false-attribution, suggesting that in-context examples may encourage more cautious, evidence-grounded responses even when overall diagnostic accuracy remains limited.

A.4.2 Model-Level Performance

Figure 7 presents model-level performance across different model families and types. Overall, performance differences across models remain modest, with no model consistently achieving high F1 scores. This suggests that cognitive skill diagnosis remains challenging regardless of model architecture or scale.

A.4.3 Model-wise Confusion Patterns

Figure 8 shows model-wise confusion patterns across verdict categories.

A.4.4 Skill-Level Performance

Figure 9 shows performance and evidential attribution errors across cognitive skills. Performance varies across skills, while evidence over-attribution and false-attribution remain consistently high. This indicates that models struggle not only with skill classification but also with accurately grounding their predictions in appropriate evidence.

Consistent patterns are observed in the model-wise results (Tables 9–12), which further show that no model achieves uniformly strong performance across all skills and that attribution-related errors persist across models.

A.5 Examples of Model Errors

Figure 10 provides illustrative examples of failure cases in cognitive skill diagnosis. These examples highlight common error patterns, such as misidentifying relevant evidence, hallucinating unsupported reasoning steps, and over-inferring skills from incomplete student work.

Topics	Content Domain	Difficulty
1. Problems involving the digits of numbers	Number	1, 3
2. Solving for unknowns under special conditions on the solution	Algebra	1, 2, 3
3. Finding unknowns when two equations have the same solution	Algebra	1, 2
4. Applying linear inequalities to geometric figures	Geometry	2, 3
5. Applying linear inequalities to pricing	Algebra	1, 2
7. Using square roots to calculate lengths in geometric figures	Geometry	2, 3
8. Performing simple addition and subtraction of square roots	Number	1, 2
9. Applying multiplication formulas	Number	2, 3
10. Determining coefficients or constants to complete the square	Number	1, 2, 3
11. Solving quadratic equations by factoring	Algebra	1, 2
12. Rewriting expressions in perfect square form	Algebra	1, 2
15. Finding the quadratic function given the vertex and another point on its graph	Algebra	1, 2

Table 5: The topics and difficulty levels of problems. This information comes from the original dataset's metadata, which details the topic and difficulty level (1-3) of each problem. The isomorphic problems require the same mathematical concept to solve, but the difference in numbers makes one more tricky and complicated than the other.

Topics	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
% Agreement	95	70	96	95	89	68	74	86	80	80	100	88	40	57	78

Table 6: The inter-rater percentage absolute agreement of each topic. The percentage indicates the ratio of unanimous verdicts in each teacher group.

	Recall	Recognize	Classify/Order	Compute	Determine	Represent	Implement
Evident Yes	413	434	92	522	441	122	307
Evident No	62	56	8	125	54	33	140
Vague Yes	20	16	2	26	25	0	18
Vague No	35	28	0	20	16	3	18
Student responses	530	534	102	693	536	158	483

Table 7: Distribution of verdicts and number of diagnosed student responses in each cognitive skill. Note that the number of student responses can be larger than the sum of the four labels because some diagnostic checklists have two items for the same cognitive skill.

Few-shots	Precision	Recall	F1	Accuracy	OverAttr	FalseAttr
claude-3-5-sonnet	.398	.485	.410	.646	**.515**	.647
deepseek-V3	.404	**.493**	.404	.624	**.493**	**.497**
gemini-1.5-pro	.429	**.517**	.430	.665	**.427**	**.552**
gpt-4o	.390	**.485**	.396	.637	**.577**	**.587**
llama-3.1-405B	**.442**	**.494**	**.429**	**.677**	.476	.457

Table 8: Performance of five state-of-the-art models in few-shot experiments. Bold values denote improvements over the corresponding zero-shot results.

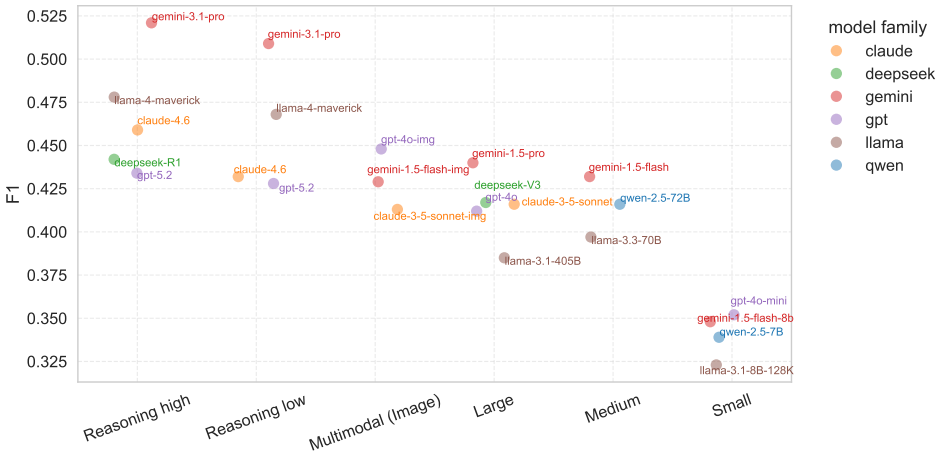


Figure 7: F1 score across by model type and family.

Model	Recall	Recognize	Classify/Order	Compute	Determine	Represent	Implement
GPT-5.2-Low	0.362	0.425	0.297	0.431	0.373	0.404	0.427
GPT-5.2-High	0.346	0.420	0.302	0.411	0.353	0.456	0.448
GPT-4o-Img	0.443	0.471	0.335	0.399	0.410	0.408	0.413
GPT-4o	0.400	0.368	0.333	0.392	0.357	0.395	0.440
GPT-4o-mini	0.361	0.339	0.178	0.344	0.297	0.392	0.322
Claude-Sonnet-4.6-Low	0.419	0.366	0.298	0.413	0.415	0.438	0.412
Claude-Sonnet-4.6-High	0.460	0.390	0.338	0.436	0.437	0.377	0.452
Claude-3.5-Sonnet-Img	0.364	0.374	0.328	0.431	0.407	0.393	0.412
Claude-3.5-Sonnet	0.358	0.382	0.325	0.424	0.438	0.473	0.396
Gemini-3.1-Low	0.438	0.525	0.309	0.452	0.493	0.407	0.531
Gemini-3.1-High	0.474	0.561	0.224	0.449	0.481	0.394	0.598
Gemini-1.5-Flash-Img	0.362	0.415	0.226	0.387	0.405	0.417	0.462
Gemini-1.5-pro	0.395	0.429	0.319	0.367	0.392	0.416	0.468
Gemini-1.5-Flash	0.398	0.422	0.266	0.378	0.414	0.395	0.451
Gemini-1.5-Flash-8b	0.276	0.330	0.185	0.337	0.334	0.456	0.349
Deepseek-R1	0.469	0.483	0.357	0.376	0.384	0.416	0.395
Deepseek-V3	0.402	0.439	0.224	0.372	0.390	0.392	0.380
Llama-Maverick-Low	0.438	0.481	0.232	0.407	0.418	0.384	0.410
Llama-Maverick-High	0.430	0.482	0.232	0.409	0.461	0.394	0.460
Llama-3.1-405B	0.344	0.342	0.231	0.373	0.385	0.349	0.409
Llama-3.3-70B	0.337	0.399	0.312	0.376	0.409	0.469	0.393
Llama-3.1-8B-128K	0.322	0.266	0.334	0.316	0.271	0.285	0.291
Qwen-2.5-72B	0.387	0.462	0.296	0.369	0.371	0.394	0.394
Qwen-2.5-7B	0.321	0.337	0.231	0.343	0.338	0.463	0.308

Table 9: F1 by model (rows) and skill (columns).

Figure 8: Model-wise confusion patterns in cognitive skill diagnosis across verdict categories (Weak/Strong $\times$ Yes/No and Vague/Evident $\times$ Yes/No). Each heatmap shows the normalized distribution of predicted labels for a given model. Darker colors indicate higher proportions.

Model	Recall	Recognize	Classify/Order	Compute	Determine	Represent/Model	Implement
GPT-5.2-Low	0.502	0.697	0.784	0.691	0.705	0.816	0.644
GPT-5.2-High	0.549	0.717	0.814	0.698	0.690	0.829	0.648
GPT-4o-Img	0.700	0.730	0.833	0.759	0.754	0.861	0.712
GPT-4o	0.589	0.610	0.804	0.716	0.683	0.829	0.675
GPT-4o-mini	0.543	0.622	0.412	0.622	0.700	0.829	0.600
Claude-Sonnet-4.6-Low	0.568	0.596	0.775	0.713	0.646	0.715	0.642
Claude-Sonnet-4.6-High	0.574	0.635	0.735	0.722	0.681	0.728	0.644
Claude-3.5-Sonnet-Img	0.564	0.710	0.902	0.737	0.715	0.823	0.627
Claude-3.5-Sonnet	0.534	0.674	0.804	0.688	0.692	0.810	0.602
Gemini-3.1-Low	0.691	0.803	0.794	0.762	0.825	0.880	0.760
Gemini-3.1-High	0.696	0.830	0.775	0.740	0.819	0.867	0.793
Gemini-1.5-Flash-Img	0.615	0.670	0.784	0.722	0.754	0.810	0.696
Gemini-1.5-pro	0.666	0.719	0.824	0.677	0.731	0.880	0.665
Gemini-1.5-Flash	0.583	0.669	0.725	0.696	0.720	0.829	0.667
Gemini-1.5-Flash-8b	0.366	0.566	0.343	0.610	0.659	0.797	0.545
Deepseek-R1	0.743	0.805	0.873	0.753	0.804	0.873	0.712
Deepseek-V3	0.649	0.730	0.775	0.734	0.720	0.848	0.673
Llama-Maverick-Low	0.689	0.773	0.833	0.769	0.810	0.842	0.700
Llama-Maverick-High	0.692	0.787	0.833	0.766	0.819	0.861	0.702
Llama-3.1-405B	0.487	0.566	0.843	0.680	0.681	0.703	0.621
Llama-3.3-70B	0.458	0.584	0.765	0.685	0.728	0.816	0.621
Llama-3.1-8B-128K	0.592	0.678	0.873	0.658	0.707	0.791	0.542
Qwen-2.5-72B	0.619	0.766	0.824	0.723	0.726	0.854	0.648
Qwen-2.5-7B	0.634	0.700	0.637	0.729	0.787	0.829	0.636

Table 10: Accuracy by model (rows) and skill (columns)

Model	Recall	Recognize	Classify/Order	Compute	Determine	Represent/Model	Implement
GPT-5.2-Low	0.218	0.341	0.500	0.587	0.366	1.000	0.556
GPT-5.2-High	0.236	0.250	0.500	0.500	0.220	0.667	0.417
GPT-4o-Img	0.436	0.477	1.000	0.848	0.732	1.000	0.833
GPT-4o	0.364	0.614	1.000	0.826	0.756	1.000	0.722
GPT-4o-mini	0.455	0.682	1.000	0.848	1.000	1.000	0.944
Claude-Sonnet-4.6-Low	0.255	0.477	1.000	0.761	0.317	0.667	0.694
Claude-Sonnet-4.6-High	0.164	0.500	0.000	0.674	0.220	0.667	0.611
Claude-3.5-Sonnet-Img	0.527	0.773	1.000	0.717	0.439	1.000	0.639
Claude-3.5-Sonnet	0.455	0.614	1.000	0.674	0.341	0.667	0.667
Gemini-3.1-Low	0.255	0.273	0.500	0.739	0.561	1.000	0.611
Gemini-3.1-High	0.255	0.295	0.500	0.674	0.537	1.000	0.528
Gemini-1.5-Flash-Img	0.236	0.295	0.500	0.652	0.561	0.667	0.694
Gemini-1.5-pro	0.309	0.409	1.000	0.804	0.512	1.000	0.556
Gemini-1.5-Flash	0.236	0.409	1.000	0.696	0.585	1.000	0.611
Gemini-1.5-Flash-8b	0.218	0.386	0.000	0.804	0.659	0.667	0.750
Deepseek-R1	0.564	0.727	1.000	0.739	0.927	1.000	0.944
Deepseek-V3	0.327	0.545	0.500	0.804	0.659	1.000	0.806
Llama-Maverick-Low	0.291	0.364	1.000	0.717	0.610	0.667	0.861
Llama-Maverick-High	0.309	0.364	1.000	0.739	0.683	1.000	0.722
Llama-3.1-405B	0.200	0.568	1.000	0.587	0.366	1.000	0.639
Llama-3.3-70B	0.218	0.295	1.000	0.630	0.488	0.333	0.611
Llama-3.1-8B-128K	0.273	0.795	1.000	0.891	0.878	1.000	0.778
Qwen-2.5-72B	0.236	0.432	1.000	0.848	0.659	1.000	0.722
Qwen-2.5-7B	0.527	0.614	1.000	0.913	0.878	0.667	0.944

Table 11: Over-Attribution by model (rows) and skill (columns)

Model	Recall	Recognize	Classify/Order	Compute	Determine	Represent/Model	Implement
GPT-5.2-Low	0.190	0.281	0.429	0.682	0.371	0.800	0.664
GPT-5.2-High	0.277	0.286	0.333	0.633	0.293	1.000	0.621
GPT-4o-Img	0.509	0.560	0.875	0.818	0.750	0.778	0.742
GPT-4o	0.470	0.567	0.500	0.822	0.677	0.636	0.769
GPT-4o-mini	0.640	0.820	0.904	0.787	0.854	0.895	0.821
Claude-Sonnet-4.6-Low	0.368	0.404	0.500	0.766	0.529	0.500	0.656
Claude-Sonnet-4.6-High	0.438	0.438	0.889	0.844	0.346	0.579	0.607
Claude-3.5-Sonnet-Img	0.425	0.526	1.000	0.741	0.427	0.737	0.594
Claude-3.5-Sonnet	0.424	0.414	0.364	0.743	0.429	0.833	0.717
Gemini-3.1-Low	0.349	0.531	0.308	0.752	0.633	1.000	0.800
Gemini-3.1-High	0.358	0.542	0.222	0.752	0.583	1.000	0.845
Gemini-1.5-Flash-Img	0.191	0.274	0.500	0.726	0.541	0.522	0.688
Gemini-1.5-pro	0.301	0.500	0.556	0.811	0.547	0.765	0.817
Gemini-1.5-Flash	0.201	0.370	0.375	0.675	0.486	0.667	0.628
Gemini-1.5-Flash-8b	0.218	0.378	0.604	0.813	0.562	0.769	0.703
Deepseek-R1	0.737	0.897	0.727	0.846	0.865	1.000	0.948
Deepseek-V3	0.368	0.447	0.556	0.805	0.476	0.875	0.728
Llama-Maverick-Low	0.472	0.483	0.556	0.806	0.588	0.682	0.656
Llama-Maverick-High	0.427	0.467	0.556	0.774	0.640	0.842	0.688
Llama-3.1-405B	0.091	0.237	0.800	0.451	0.306	0.333	0.531
Llama-3.3-70B	0.182	0.394	0.333	0.551	0.471	0.542	0.519
Llama-3.1-8B-128K	0.345	0.563	0.727	0.750	0.625	0.967	0.690
Qwen-2.5-72B	0.175	0.493	0.571	0.619	0.557	0.722	0.680
Qwen-2.5-7B	0.674	0.702	0.889	0.867	0.824	0.926	0.883

Table 12: False-Attribution by model (rows) and skill (columns)

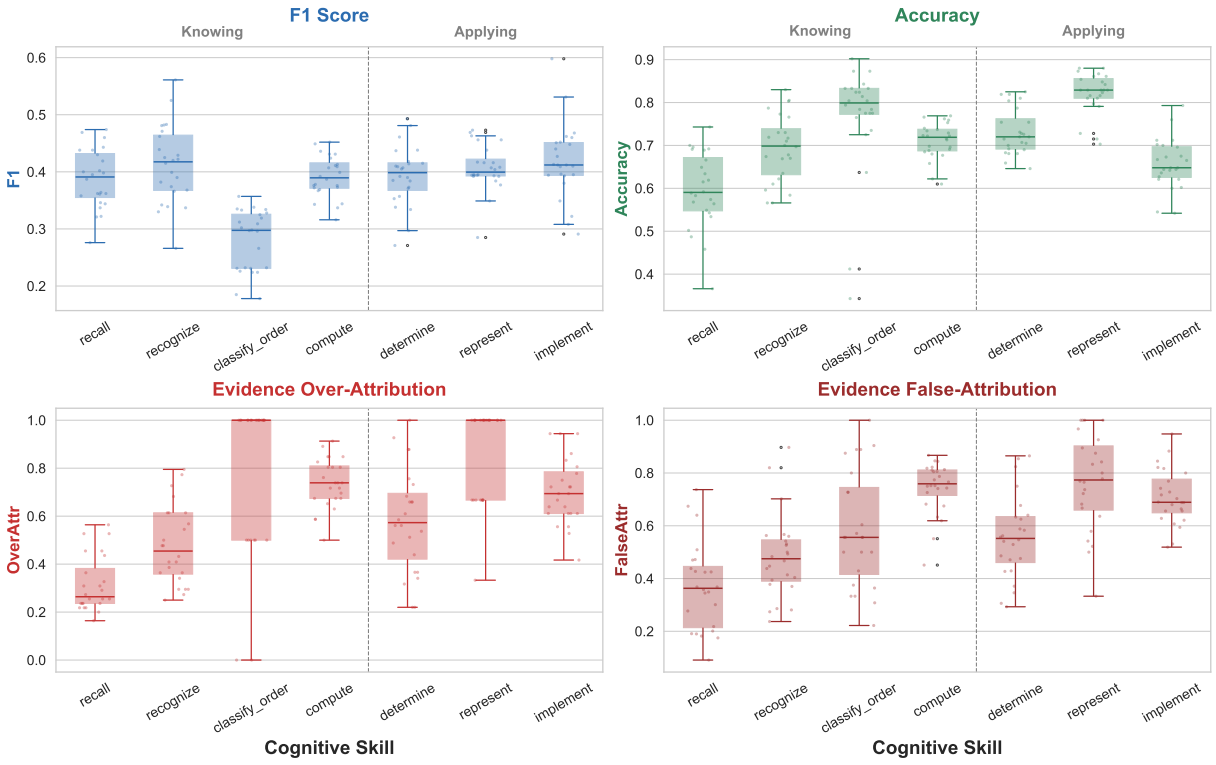
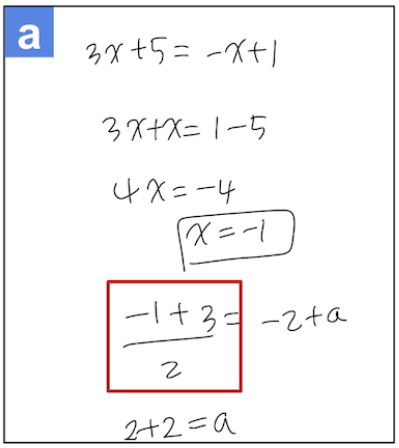


Figure 9: Performance and evidential attribution errors across cognitive skills, measured by F1 score, accuracy, evidence over-attribution, and false-attribution.

[Compute] After substituting the value of x into the second equation, were the four arithmetic operations performed correctly in calculating the value of a or m?

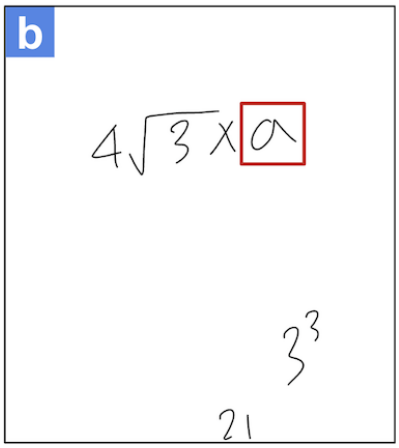


Teachers:
Evident No

Claude-3-5-Sonnet:
Evident No **X**

Claude-3-5-Sonnet-img:
Evident No **O**

[Compute] When calculating the prime factorization of a constant, was it performed accurately without arithmetic errors?

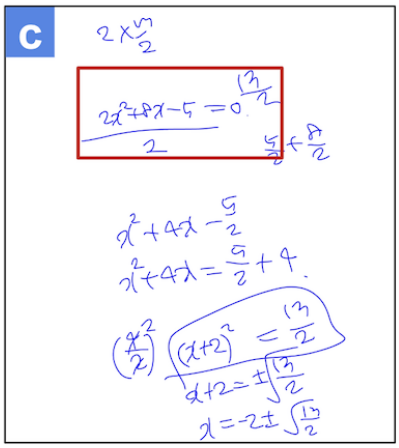


Teachers:
Evident No

DeepSeek-R1:
Vague No **X**

DeepSeek-V3:
Vague No **X**

[Recognize] Does the student notice that, to simplify the equation, both sides need to be divided by the coefficient of the highest order term, if necessary?



Teachers:
Evident Yes

Gemini-1.5-Pro:
Evident No **X**

GPT-4o:
Evident No **X**

Figure 10: Illustrative examples of diagnosis check items and student responses that LLMs failed to diagnose correctly. Evidence for human judgment is marked with a red box.