The Aftermath of DrawEduMath: Vision Language Models Underperform with Struggling Students and Misdiagnose Errors

Li Lucy[△] **Albert Zhang**[+] **Nathan Anderson**[÷] **Ryan Knight**[+] **Kyle Lo**[□△]

[△]University of Washington [+]Insource Services [÷]Worcester Polytechnic Institute

[□]Allen Institute for AI

lucy3li@cs.washington.edu kylel@allenai.org

Abstract

Effective mathematics education requires identifying and responding to students' mistakes. For AI to support pedagogical applications, models must perform well across different levels of student proficiency. Our work provides an extensive, year-long snapshot of how 11 vision-language models (VLMs) perform on DrawEduMath, a QA benchmark involving real students' handwritten, hand-drawn responses to math problems. We find that models' weaknesses concentrate on a core component of math education: student error. All evaluated VLMs underperform when describing work from students who may require more pedagogical help, and across all QA, they struggle the most on questions related to assessing student error. Thus, while VLMs may be optimized to be math problem solving experts, our results suggest that they require alternative development incentives to adequately support educational use cases.

1 Introduction

The use of vision language models (VLMs) in education has received increasing attention in both academic research (Küchemann et al., 2025; Lee et al., 2025b) and commercial AI products. Examples of the latter include Google Classroom with Gemini integration,[1] and Khan Academy's AI tutor Khanmigo,[2] powered by OpenAI models. However, the integration of these models into tutoring and classroom settings often lacks transparent, open, and realistic evaluation. With this gap in mind, we previously released DrawEduMath (Figure 1), a dataset consisting of 2,030 teacher-annotated images of real students' hand-drawn responses to K-12 math problems (Baral et al., 2025). In contrast to other multimodal math understanding or problem solving benchmarks (Alshammari et al., 2026), DrawEduMath involves noisy, naturalistic data pulled from an online learning platform (Figure 1). In the year since its release, we continuously updated the benchmark's leaderboard[3] with newer models.

Our paper offers a snapshot of how 11 VLMs have performed on DrawEduMath in the year after its release (Figure 2).[4] We surface two key findings:

F1: VLMs are worse at describing the contents of student work that contains math errors than student work without errors.

F2: VLMs still struggle the most on question types related to assessing students' correctness.

These findings suggest that VLMs underperform with students who need additional pedagogical support (**F1**), and they also fail to appropriately identify cases when support is needed (**F2**).

To investigate these patterns further, we conduct five analyses in §4-§8 targeting factors that may relate to model performance. Our experiments show that the performance gap in **F1** persists even when controlling for problem (§4) and when image noise is reduced (§5). In addition, a possible explanation for **F1** is that VLMs expect mathematically correct input images. Indeed, we find that some models' wrongly predicted answers for erroneous student work are similar to gold answers for non-erroneous student work (§6).

We also find that models do improve in assessing student correctness (**F2**) when provided gold natural language descriptions of student work (§7). However, their performance on these questions with extra textual support still lags behind their out-of-the-box performance on other question types. Finally, though models seemingly perform better

[1] https://blog.google/outreach-initiatives/education/classroom-ai-features/

[2] https://www.khanmigo.ai/

[3] https://drawedumath.org/

[4] DrawEduMath was released to accompany an initial non-archival paper in Dec 2024 (Baral et al., 2024, 2025).

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 48–62

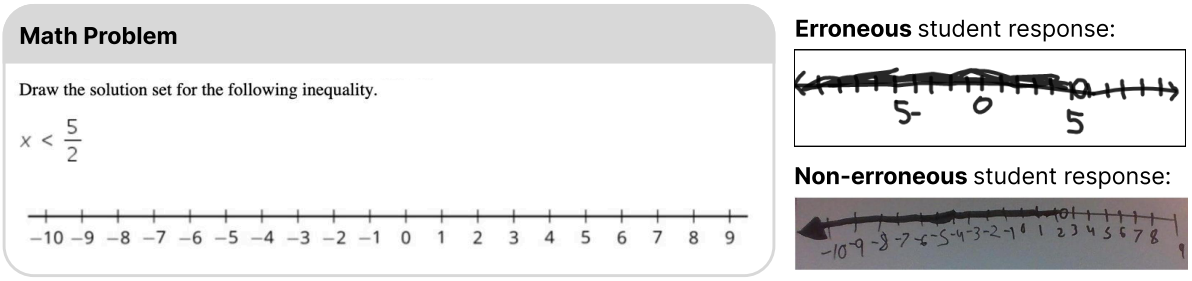


Figure 1: On the left is a math problem, where students are asked to draw $x < 5/2$ on a number line. The right side shows two example student responses that differ in correctness. DrawEduMath pairs each math problem with one student response, and prompts VLMs to answer questions about the student response.

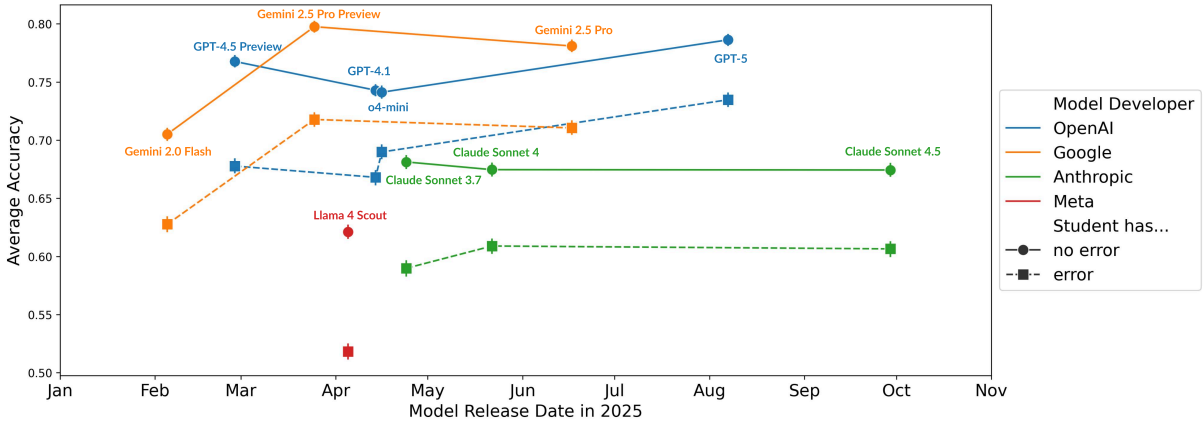


Figure 2: VLMs consistently perform worse on answering DrawEduMath benchmark questions pertaining to erroneous student responses. Performance on non-erroneous student responses (○) is labeled with specific VLMs' names; that same model's performance on erroneous student responses is directly below (□). Error bars are 95% CI.

on binary correctness questions (e.g. "*Does the student do ___ correctly?*") than open-ended ones (e.g. "*What errors does the student make in their response?*"), some VLMs' performance can sometimes be barely better than chance (§8).

Altogether, our in-depth error analysis of VLMs' performance on real student math responses provides a clearer picture of their current weaknesses in supporting K-12 math education. We release data and scripts for reproducing our findings.[5]

2 Background & Related Work

Multimodal math benchmarks. Mathematical content, rich with diagrams, is ripe for evaluating models' multimodal abilities. Thus, many vision-language benchmark creation efforts have targeted math (Alshammari et al., 2026; Zhang et al., 2024; Lu et al., 2024; Yan et al., 2025). Educational settings offer additional challenges, where VLMs may be assessed on their abilities to make higher-level pedagogical inferences and handle handwritten, hand-drawn content (Parsaeifard et al., 2025; Latif et al., 2025; Nath et al., 2025; Nguyen et al., 2025). For example, MathCog asks models to diagnose students' cognitive skills using binary yes/no questions and a digitally handwritten dataset (Jin et al., 2025). Within this landscape of prior work, DrawEduMath remains a significant evaluation resource, given its diversity of image and question types, its use of noisy, real student work, and its inclusion of experienced teacher annotations (Baral et al., 2025).

AI & student error. Student mistakes and misconceptions have long been a focal point in education research (Smith III et al., 1994; Radatz, 1979; Borasi, 1994; Metcalfe, 2017). With increased attention towards AI as tutors and teaching assistants, research has focused on models' abilities to identify student error (Srivatsa et al., 2025; Kochmar et al., 2025; Daheim et al., 2024), reason about patterned misconceptions (Rittle-Johnson et al., 2025; Ross and Andreas, 2025), correct errors (Mita et al.,

[5] https://github.com/lucy3/aftermath_drawedumath

2024), and provide feedback (Kaliisa et al., 2026; Botelho et al., 2023; Stahl et al., 2024). There is some research around model robustness to user error, but much of it focuses on linguistic errors in prompts (Gan et al., 2024; Chatterjee et al., 2024). There is little work on how math errors impact model performance: one example is Daheim et al. (2024), who show that language models are worse at verifying the correctness of erroneous student math than non-erroneous math. Our work reaches a similar conclusion across more QA types and with multimodal data, though our results around models' correctness & error assessments paint a less straightforward picture (§8).

Risks of AI in education. The field of AI & education could be considered a form of AI for "social good" (Cowls et al., 2021), driven by goals that counter AI's negative impacts on human skill formation and cognitive thinking (Bastani et al., 2025; Klimova and Pikhart, 2025; Shen and Tamkin, 2026; Lee et al., 2025a). However, though its end-goals may be optimistically framed, AI in education is not without risk (Blodgett and Madaio, 2021; Holstein and Doroudi, 2021). Our work is aligned with literature that investigates how AI may disparately impact different student populations (Schaller et al., 2024; Capraro et al., 2024; Hadar Shoval, 2025); our distinct approach is that we group student inputs based on demonstrated math proficiency.

3 Evaluating VLMs with DrawEduMath

3.1 Data

DrawEduMath is an English-language dataset of 2,030 images of students' handwritten responses to K-12 math problems (Baral et al., 2025). Images are provided by the online learning platform ASSISTments and contain math problems drawn from open educational resources (Heffernan and Heffernan, 2014; Feng et al., 2025). Each image includes a math problem on the left and a student's response on the right, and is accompanied by three types of data:

1. **Free-form captions** (2.0k+) from teachers describing each student response image.
2. **Synthetic QA pairs** (44.4k+), produced by Claude-3.5 Sonnet and GPT-4o reformatting facets of teachers' captions into QA, e.g. *On the left-hand side of the image, the student wrote the word syrup* → *What word did the student write on the left-hand side of the image? Syrup.*
3. **Teacher-written QA pairs** (11.6k+). Teachers wrote a set of shared questions for each math problem, followed by answers for each student response to each problem. Teachers answered two additional generic questions, *What errors does the student make in their response?* and *What strategy does the student use to solve the problem?* across all problems and student responses.

DrawEduMath includes a taxonomy of seven question types. In our analysis, we simplify this taxonomy into three types: *image creation and medium* (12.3%), *correctness & errors* (8.5%), and *content description* (79.2%). The first two match the benchmark's original taxonomy, while the third question type is an aggregation of all other question types, ranging from the student's problem solving strategy to the meaning, positioning, and frequency of drawn/written elements. We aggregate these question types in our main text, because our main findings generalize across more fine-grained types (Appendix A). The original DrawEduMath paper includes additional dataset statistics and QA examples (Baral et al., 2025).

3.2 Evaluation Setup

Scoring metric. Baral et al. (2025) used Mixtral 8x22B to judge the similarity of VLMs' generated answers to gold answers on a scale of 1 (*quite different answers*) - 4 (*basically the same*), and then binarized these ratings when computing models' accuracy. As LMs update over time, older ones like Mixtral 8x22B become deprecated in model API services. Thus, all evaluation in our work uses an updated LLM judge. We take the majority vote from three judges: Claude Sonnet 4.5, Gemini 2.5 Pro, and GPT-4o.[6] Our updated judge achieves similar correlation (Spearman $\rho = 0.808$) with the same set of human judgements as Baral et al. (2025)'s original judge ($\rho = 0.801$). We compute and report binarized model accuracies; 1-2 are counted as incorrect, and 3-4 are correct.

Models. We evaluate 11 VLMs released in 2025 on DrawEduMath. Models span four developers: Open AI (GPT-4.1, GPT-4.5 Preview, o4-mini,

[6]Though GPT-4o is an older model, it has higher individual correlation with human annotations than the two newer models. So, we included it in our set of judges.

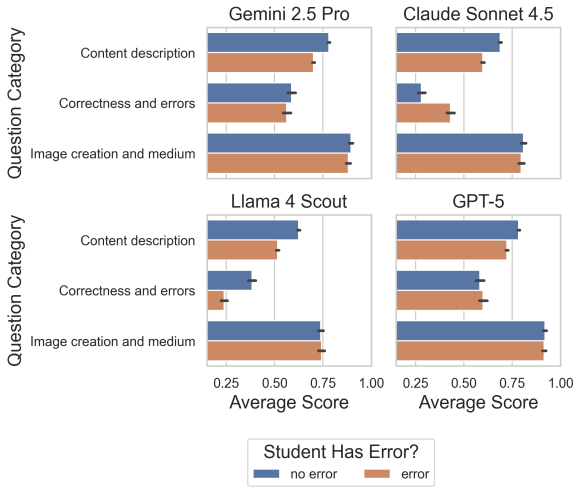


Figure 3: Content description QA consistently drives the gap in VLM performance between student responses that contain errors versus those that do not. Appendix A includes additional VLMs that expand this finding.

GPT-5), Anthropic (Claude Sonnet 3.7, Claude Sonnet 4, Claude Sonnet 4.5), Google (Gemini 2.0 Flash, Gemini 2.5 Pro, Gemini 2.5 Pro Preview,), and Meta AI (Llama 4 Scout). Though our main findings **F1** & **F2** pertain to all of these models, the analyses in §5-§8 focus on four representative models: Gemini 2.5 Pro, Claude Sonnet 4.5, GPT-5, and Llama 4 Scout.

Labeling student error. To categorize whether students' math responses contain an error or not, we use teachers' answers to the question, *What errors does the student make in their response?* We ask GPT-5-mini to interpret each open-ended answer and classify it as *yes*, as in, the teacher describes some error, or *no*, for when teachers' answers are variations of *There is no error* or *The student did not make an error* (Appendix B.1). We validate this LM annotator on a manually checked random sample of 200 examples (F1 = 0.984, Cohen's κ = 0.970).

3.3 Main Findings

On average, models tend to perform worse when the student response contains an error (Figure 2). We find that this pattern is mostly driven by content description QA (**F1**, Figure 3). In addition, a weakness reported by Baral et al. (2025) in older VLMs persists in newer ones: questions related to students' correctness and errors are still the most difficult (**F2**). In the next few sections, we dive into five factors that we hypothesize to relate to these findings: problem effects (§4), image noise (§5), problem response defaults (§6), visual understanding bottlenecks (§7), and question open-endedness (§8).

Model	β_1
Gemini 2.0 Flash	0.0944***
Gemini 2.5 Pro Preview	0.0888***
Gemini 2.5 Pro	0.0889***
GPT-4.1	0.0865***
GPT-4.5 preview	0.0874***
o4-mini	0.0612***
GPT-5	0.0585***
Claude Sonnet 3.7	0.0922***
Claude Sonnet 4	0.0816***
Claude Sonnet 4.5	0.0842***
Llama 4 Scout	0.1013***

Table 1: Estimated effects of student correctness (β_1) on VLMs' accuracy on content description QA, where all $p < 1.0^{-12}$.

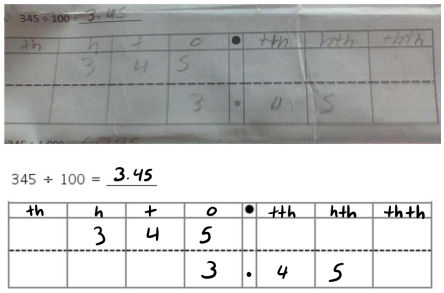
4 Models underperform on erroneous student responses even when controlling for problem

One possibility is that the model performance gap observed by **F1** is actually not affected by the presence or absence of student error, but rather by some math problems being more difficult for VLMs to understand. DrawEduMath contains 188 unique math problems targeting concepts ranging from geometry to fractions. These problems span multiple grade levels, and on average, each problem has 12.64 student responses (Baral et al., 2025). Here, we show that the effect of student error on VLMs' content description QA performance is statistically significant even when controlling for problem.

We estimate an ordinary least squares regression with problem fixed effects:

$$y_{ij} = \beta_0 + \beta_1(s_{ij}) + u_j + \epsilon_{ij}$$

In the equation above, y_{ij} is the average score a model has across content description QA for a student response i and problem j, u_j is a fixed effect for each problem, and ϵ_{ij} is the residual. If the student response is correct, s_{ij} = 1, otherwise s_{ij} = 0. Table 1 presents β_1 values across VLMs. These values show that even after controlling for problem, non-erroneous student responses significantly correspond with higher VLM performance.



345 ÷ 100 = 3.45

th	h	t	o	•	tth	hth	thth
	3	4	5				
			3	.	4	5	

Figure 4: An example of how a student response image (top) is transformed and cleaned up by our digital redrawing process (bottom). This student uses a place value chart to show how digit values change for 345 after division by 100.

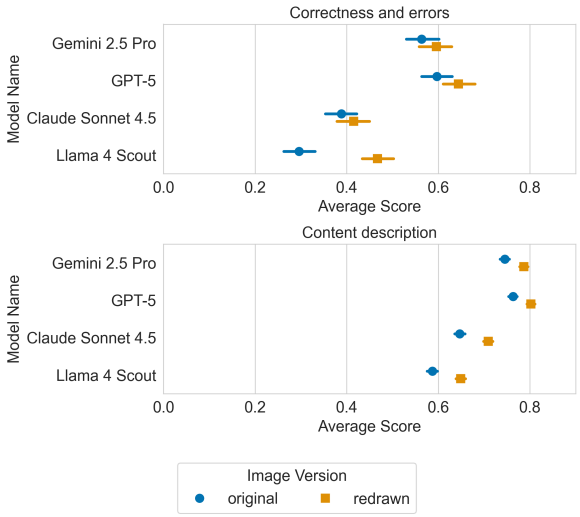


Figure 5: Models' performance for content description QA generally improves after images are redrawn. Error bars are 95% CI.

5 Models' performance gaps are not strongly impacted by image noise

Another possible explanation for **F1** is that students who make math errors may simply submit noisier images. Students on ASSISTments may submit their answers by drawing digitally, or by uploading photographs of pen & paper work, which may include smudges and blur. In this section, we ask, does the model performance gap described by **F1** remain even when students' responses are redrawn on a digital canvas in a standardized manner?

5.1 Experimental Setup

Redrawing images is a time-intensive process, requiring careful interpretation of each math problem and the intent of the original student response. So, for this experiment, we stratify sample one erroneous student response and one correct response from each problem, yielding 336 images in total. Though this sample is small, it retains statistically significant gaps in VLM performance on content description QA between erroneous and non-erroneous student response images (Table 2).

To encourage consistency, the lead author redrew all sampled student responses using a digital pen and the drawing application Procreate. This redrawing author retained students' original positioning of content, and consulted teachers' captions of images to navigate ambiguity and avoid faulty interpretation of problems and student responses. If needed, the author recreated elements such as graph paper grids and typed content in Figma. Figure 4 provides an illustrative example of how redrawing transforms students' responses.

Model	Original Images	Redrawn Images
Gemini 2.5 Pro	-0.089***	-0.096***
Claude Sonnet 4.5	-0.100***	-0.058**
GPT-5	-0.055**	-0.043**
Llama 4 Scout	-0.092***	-0.073***

Table 2: Differences in average scores on content description QA between erroneous and non-erroneous student images persist after redrawing. $^{**}p < 0.01$, $^{***}p < 0.001$.

5.2 Results

On redrawn images, VLMs' performance shift in the expected direction, where scores generally improve with less image noise (Figure 5). Though requesting students to only submit born-digital content may make their work more interpretable for AI, not all classrooms have resources and policies that make such standardization feasible. In addition, pen-and-paper work remains vital, with studies showing that this traditional mode of learning can sometimes allow students to surpass their digital-only peers (Mueller and Oppenheimer, 2014; Altamura et al., 2025; Anthony et al., 2007; Umejima et al., 2021). Thus, one implication of Figure 5 is that the integration of VLMs in education may disparately impact analog and digital learners.

Importantly, we also find that models' performance gap between erroneous and non-erroneous student images remains in redrawn images (Table 2). This result complements that of §4, by further isolating student error as a weak point in the use of VLMs in educational settings. Thus,

mitigation efforts around **F1** should focus on improving models' understanding of erroneous mathematical content, across all levels of image noise and medium types.

6 Models default to assuming error-free math solutions

Why might erroneous student images be so challenging for VLMs? During a manual examination of VLMs' QA errors, we observed that models sometimes produce plausible, though wrong, answers to benchmark questions, especially considering the context of the provided math problem (Figure 6). This observation suggests a possible explanation for **F1**: VLMs perform better on content description QA for error-free student responses, because models default to assuming error-free math solutions.

6.1 Analysis Setup

To quantify our observation using existing DrawEduMath annotations, we filter for content description QA shared across different student response images for the same math problem. Then, for each incorrect model answer to questions pertaining to erroneous student responses, we compare the model's answer against true answers for non-erroneous student responses, and see whether the model's incorrect answer matches the majority of these true ones. We compare model answers using the ensemble LM judge from §3.2. We only consider cases where we have at least two non-erroneous student response images associated with the given question, to ensure that we have sufficient signal of correct student behavior.

6.2 Results

Across four representative VLMs, incorrect model responses for erroneous student responses sometimes do match non-erroneous student solutions, with percentages ranging from 29% of content description QA mistakes for Gemini 2.5 Pro to 35% for Claude Sonnet 4.5 (Table 3). So, a sizable portion of benchmark answers may be inferrable based on a math problem and a typical correct solution. There are many more ways a student response can be wrong than it can be correct, and so benchmark QA corresponding to correct student solutions navigate a narrower space of plausible possibilities.

Qualitatively, we observe that models especially tend to predict incorrect answers that match correct problem solutions when benchmark questions involve false presuppositions. Figure 6 illustrates an example; there, the wording of the top teacher-written question assumes that the student has drawn any array at all. Teacher-written questions in DrawEduMath are those that teachers would like VLMs to answer across all student responses to a problem, mimicking potential uses of VLMs for learning analytics. Models' susceptibility to false premises or suppositions is well-documented in prior work (e.g. Yu et al., 2023; Srikanth et al., 2024), and our work illustrates a consequence of this weakness for education-related applications.

Generally, language models are developed to be good math problem solvers. Math solving benchmarks are continuously emphasized in leaderboards and commercial LM releases (Cobbe et al., 2021; Hendrycks et al., 2021; Google DeepMind, 2025). To encourage mathematically correct outputs and hill-climb on these benchmarks, models are mostly exposed to "high quality", correct math content during training (Mahabadi et al., 2025; Paster et al.,

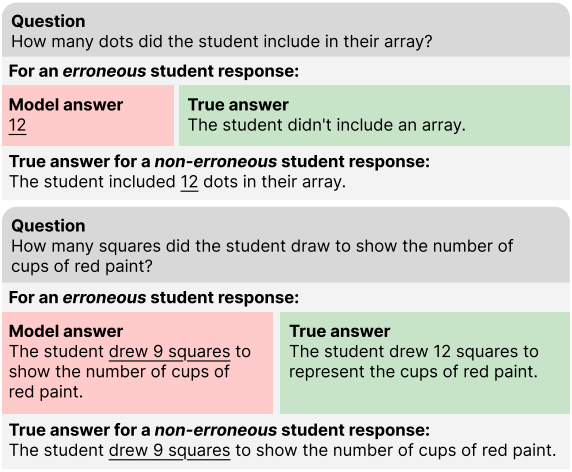


Figure 6: Illustrative examples of the phenomenon where models predict answers for erroneous student responses that match true answers for non-erroneous students.

Model	%	N
Gemini 2.5 Pro	0.2923	1,355
Claude Sonnet 4.5	0.3519	1,921
GPT-5	0.3125	1,357
Llama 4 Scout	0.3060	2,134

Table 3: The percentage of times for which an incorrect model answer for a content description question and erroneous student image matched the majority of true answers for non-erroneous student images. N is the number of incorrect model answers in the denominator of the percentage.

2024). The challenge of understanding, but not generating faulty content has received attention in other domains. For example, toxicity is another case of an understanding vs. generation tradeoff; we want models that can detect, address, and understand toxic content, without generating it (Longpre et al., 2024; Wang et al., 2025). Our findings suggest that education is another domain where the application of alternative training methods on erroneous data, such as Wang et al. (2025), could be applicable.

7 Textual support can improve models' correctness assessments to some extent

DrawEduMath QA range from low-level content description (e.g. "*How many triangles did the student draw?*") to higher-level correctness judgements. Now, we move on from examining **F1**, which focuses on content description QA, to digging deeper into **F2**, which pertains to correctness & errors QA. Earlier, we saw that the latter remain difficult even after images are digitally cleaned up (Figure 5). Perhaps, image understanding is a bottleneck for models answering these more reasoning-intensive questions. To what extent can models improve their assessment of student error when given textual descriptions of student work?

7.1 Experimental Setup

As mentioned in §3.1, DrawEduMath includes teacher-written, gold captions of students' response images. Baral et al. (2025) used these captions to synthetically generate a subset of the QA pairs in the benchmark. If a caption produces synthetic QA that fall in the correctness & error question category, we exclude that image from the current analysis to avoid input-output contamination.[7] We append to each DrawEduMath input prompt these gold captions (prompt in Appendix C.1), and re-evaluate models' performance on correctness & error QA. In addition, we also evaluate a setup where we ask models to generate their own descriptions of students' responses, and provide those captions in place of gold ones.

7.2 Results

Results are in the expected direction, in that VLM performance on correctness & errors QA improves with natural language support (Figure 7). However, this improved performance on correctness & errors QA with captions still lags behind VLMs' captionless performance in all other question categories for the same set of images. So, it is challenging for VLMs to make higher-level inferences useful for pedagogy even with gold textual support. A fully automatic two step caption-then-answer-QA process is a form of test-time scaling. We find that providing models their own generated captions of images can get models close to, but not match, their performance with teacher-written gold captions (Figure 7).

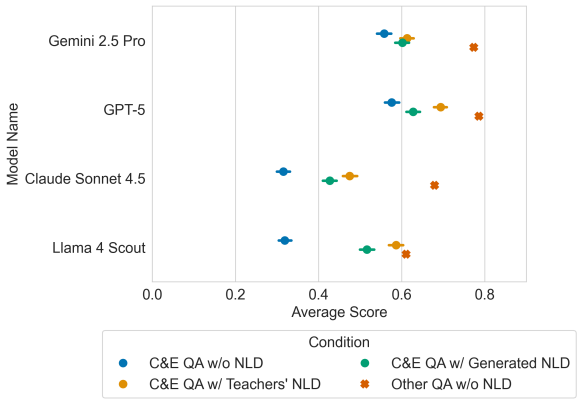


Figure 7: Model performance on correctness & error (C&E) QA, with and without natural language description (NLD) support. We evaluate with a subset of input images and captions as described in §7.1. Error bars are 95% CI.

8 Binary judgements of student correctness remain challenging

Open-ended questions are inherently more difficult than binary (e.g. yes/no) questions, as the latter is more guessable. Correctness & errors QA (**F2**) in DrawEduMath provide a natural testbed for comparing open-ended questions ("*What errors does the student make in their response?*") and binary questions that assess whether some aspect of a student's response is correct/incorrect.

8.1 Analysis Setup

Our analysis splits correctness & error QA into the following three subcategories:

- `Generic` questions (45.0%). This is the open-ended question that DrawEduMath includes for all student images: "*What errors does the student make in their response?*"

[7] We remove 262 images out of a total of 2,030. We considered editing captions rather than remove images, but correctness-related content was sometimes integrated with other caption content and would require intensive rewriting.

- `Binary` assessments of specific solution components (50.4%), e.g. "*Does the student put the decimal in the correct place in the product?*"
- `Other` questions (4.5%), which mostly pertain to the nature of a student's error, e.g. "*What incorrect product did the student calculate for 667 times 5?*"

We use GPT-5-mini as an annotator to label whether non-`generic` questions are `binary` or `other` (prompt in Appendix B.2). We validate this LM annotator on a manually labeled random sample of 200 unique questions (F1 = 0.975, Cohen's $\kappa = 0.934$). We focus on `binary` and `generic` in the main text; Appendix D includes some results involving `other`.

Do models tend to predict that students make errors when they don't, or do they tend to overlook errors instead? For `binary` QA, we use GPT-5-mini to annotate whether questions and gold answers indicate that student is correct or incorrect (prompt in Appendix B.3). For example, for the ground truth binary QA pair "*Q: Is there an error in the way that the number line has been drawn? A: Yes*", the LM annotator would output that the student is *incorrect*. We validate this LM annotator on a sample of 200 manually annotated random examples (F1 = 0.911, Cohen's $\kappa = 0.854$). In total, the LM annotator labels 59.01% of 2,274 binary QA as cases where the specified aspect of the student's response is correct, and the rest as ones where the specified aspect is incorrect.

We also examine model performance on `generic` QA, disaggregated by whether the student's response is overall incorrect or correct. In DrawEduMath, successfully answering this question for correct students requires simply stating that there is no error, while for models to score well for erroneous students, they must also faithfully describe error specifics.

8.2 Results

Figure 8 shows that performance patterns on correctness & error QA, in relation to student correctness, generally and specifically, tend to be idiosyncratic. Some models tend to overreport errors being present, with lower scores on student images with *no error*. Others struggle to detect and, in the case of `generic`, articulate errors that are present, with lower scores on student images with *error*. Model behavior patterns are not shared

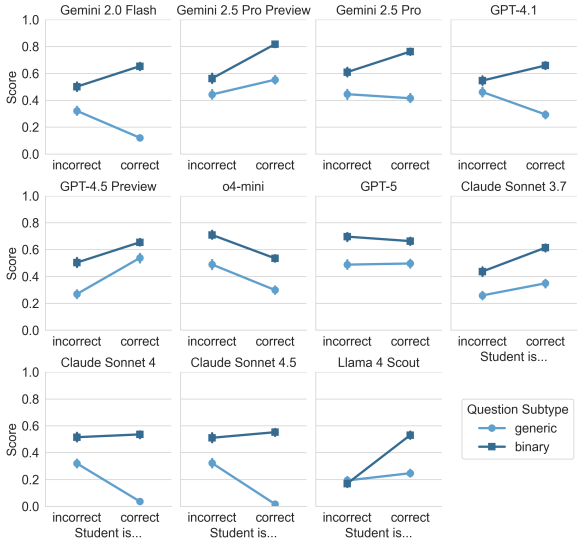


Figure 8: VLMs' performance on the two main subtypes of correctness & error QA, disaggregated by whether a student response is overall correct (`generic`) or correct based on specific aspect of their solution (`binary`). Error bars are 95% CI.

across model versions from the same family or developer. Figure 8 also indicates that some VLMs' `binary` QA scores hover closely around a random baseline of 0.5. Overall, assessing student error is incredibly challenging for VLMs, even though a substantial proportion of correctness & error QA in DrawEduMath have high by-chance floor for performance.

9 Conclusion

Despite increasing attention towards the use of multimodal AI in education, our evaluation of 11 models released in 2025 demonstrates that their application on real student data remains challenging. Our findings suggest that erroneous student work is inherently more difficult for VLMs than correct student work (**F1**). VLM training and evaluation pipelines that favor correct mathematical content are at tension with the promise of AI for education, where incorrect math requires extra emphasis and attention. We also show that QA involving assessments of student correctness are particularly tricky (**F2**), across both text and image inputs, and across open-ended and binary question forms.

Though this present paper presents a detailed error analysis of VLMs' performance on one vision-language K-12 math benchmark, our evaluation approach can be re-applied to other education-related benchmarks as well. That is, the evaluation of AI in education should be disaggregated in a man-

ner that pinpoints whether models can actually discern when a student may need pedagogical support (**F2**), and whether they equitably serve students across different levels of proficiency (**F1**). Without a careful eye on the latter, models' capabilities may be overstated, and rushed integration into classrooms may exacerbate existing academic achievement gaps.

Limitations

Scope and data representativeness. Our study focuses on a single English benchmark, which involves student response images drawn from one online learning platform, ASSISTments. Thus, our findings may not map directly onto other languages and learning contexts. Based on school-level data provided by ASSISTments, we estimate that 85% of images come from Title I schools, which are public schools in the U.S. that receive federal funding to support low-income students. ASSISTments partners with teachers and schools located across multiple location types (e.g. rural, suburban, town, city) and regions (e.g. West Coast, Midwest, East Coast, South), but self-selection is at play when it comes to which teachers, schools, and districts use the platform. DrawEduMath also contains questions that represent what was salient to Teaching Lab's teacher annotators (Baral et al., 2025); it is not comprehensive of all of the ways in which educators may interpret and support student learning. Still, our high-level evaluation approach can be re-applied to other benchmarks and contexts, because transparency around the impact of student error on model performance is relevant to nearly all education-related settings.

Data and evaluation granularity. We examine performance on DrawEduMath on aggregate with a binary scoring approach, both for students' correctness and LMs' VQA correctness. We acknowledge that the evaluation of VLMs may benefit from a more nuanced approach that captures multiple constructs beyond judgments of correct/incorrect. For example, future work should consider the form of student error present (e.g. arithmetic versus conceptual error). Future work should also consider more sophisticated evaluation frameworks that consider the context and complexity of different math problems, and the evaluative goals of different QA items (e.g. QA that pertain to VLMs' descriptive abilities, versus those that pertain to pedagogical abilities).

Data constraints. Some of our experiments and analyses navigate practical, data-related constraints. For example, our image redrawing experiment in §5 uses a small sample rather than the full dataset, since redrawing is a time-intensive process. Our other analyses rely on pre-existing teacher annotations and data present in DrawEdumath. For example, in §7, we removed some images from our analysis because their captions contained correctness & error information, because models' performance with textual support on those images would be inflated. In addition, the content description QA under consideration for the results shown in Table 3 are only questions shared across multiple student images for a problem, for which we could gather sufficient signal for what correct student response behavior should be. So, our results in that section (§6) primarily serve to illustrate one possible explanation for models' performance, and is not comprehensive of all of DrawEduMath.

Ethical Considerations

Education is a high-stakes setting for VLM use and deployment. The intermixing of AI and education involves delegating pedagogy to automated systems, impacting vulnerable underage populations, with possible life-long downstream effects related to economic mobility. Though there is optimism around AI's ability to support education (Demszky et al., 2025), there should also be caution that it does not exacerbate existing inequities or introduce new ones (Winters et al., 2020; Harvey et al., 2024). We acknowledge that our work focuses primarily on technical harms measurable from model outputs, and does not capture broader harms that may emerge via interaction of AI with students, teachers, and school systems (Harvey et al., 2025). In addition, AI research often involves a deployment-first mentality, where deployment may occur before a system has been deemed functional or necessary (Raji et al., 2022). Our work advocates for robust evaluation and auditing of AI prior to deployment (Raji et al., 2020), and accountability behind claims around model functionality and its social benefits (Kou et al., 2025; Wang et al., 2024).

Acknowledgments

We are grateful for valuable feedback from Douglas Jaffe, who encouraged us to dig further into the impact of student error on model performance. We are also grateful for data-related support from Sami

Baral, Neil Heffernan, and Cristina Heffernan. Our work is funded by the Gates Foundation.

References

Shaden Alshammari, Kevin Wen, Abrar Zainal, Mark Hamilton, Navid Safaei, Sultan Albarakati, William T. Freeman, and Antonio Torralba. 2026. MathNet: A global multimodal benchmark for mathematical reasoning and retrieval. In *The Fourteenth International Conference on Learning Representations*.

Lidia Altamura, Cristina Vargas, and Ladislao Salmerón. 2025. Do new forms of reading pay off? A meta-analysis on the relationship between leisure digital reading habits and text comprehension. *Review of Educational Research*, 95(1):53–88.

Lisa Anthony, Jie Yang, and Kenneth R. Koedinger. 2007. Benefits of handwritten input for students learning algebra equation solving. In *Proceedings of the 2007 Conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work*, page 521–523, NLD. IOS Press.

Sami Baral, Li Lucy, Ryan Knight, Alice Ng, Luca Soldaini, Neil Heffernan, and Kyle Lo. 2024. Drawedumath: An expert-annotated dataset of students' math images. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*.

Sami Baral, Li Lucy, Ryan Knight, Alice Ng, Luca Soldaini, Neil Heffernan, and Kyle Lo. 2025. DrawEduMath: Evaluating vision language models with expert-annotated students' hand-drawn math images. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6902–6920, Albuquerque, New Mexico. Association for Computational Linguistics.

Hamsa Bastani, Osbert Bastani, Alp Sungu, Haosen Ge, Özge Kabakcı, and Rei Mariman. 2025. Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122(26):e2422633122.

Su Lin Blodgett and Michael Madaio. 2021. Risks of AI foundation models in education. *arXiv preprint arXiv:2110.10024*.

Raffaella Borasi. 1994. Capitalizing on errors as "springboards for inquiry": A teaching experiment. *Journal for Research in Mathematics Education*, 25(2):166–208.

Anthony Botelho, Sami Baral, John A. Erickson, Priyanka Benachamardi, and Neil T. Heffernan. 2023. Leveraging natural language processing to support automated assessment and feedback for student open responses in mathematics. *Journal of Computer Assisted Learning*, 39(3):823–840.

Valerio Capraro, Austin Lentsch, Daron Acemoglu, Selin Akgun, Aisel Akhmedova, Ennio Bilancini, Jean-François Bonnefon, Pablo Brañas-Garza, Luigi Butera, Karen M Douglas, and 1 others. 2024. The impact of generative artificial intelligence on socioeconomic inequalities and policy making. *PNAS nexus*, 3(6):pgae191.

Anwoy Chatterjee, H S V N S Kowndinya Renduchintala, Sumit Bhatia, and Tanmoy Chakraborty. 2024. POSIX: A prompt sensitivity index for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14550–14565, Miami, Florida, USA. Association for Computational Linguistics.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Josh Cowls, Andreas Tsamados, Mariarosaria Taddeo, and Luciano Floridi. 2021. A definition, benchmark and database of AI for social good initiatives. *Nature Machine Intelligence*, 3(2):111–115.

Nico Daheim, Jakub Macina, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2024. Stepwise verification and remediation of student reasoning errors with large language model tutors. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8386–8411, Miami, Florida, USA. Association for Computational Linguistics.

Dorottya Demszky, Jing Liu, Heather C. Hill, Shyamoli Sanghi, and Ariel Chung. 2025. Automated feedback improves teachers' questioning quality in brick-and-mortar classrooms: Opportunities for further enhancement. *Computers & Education*, 227:105183.

Mingyu Feng, Linlin Li, Chunwei Huang, Natalie Brezack, and Kim Luttgen. 2025. Empowering teachers with technology: A national study on a formative assessment platform. In *International Conference on Artificial Intelligence in Education*, pages 119–126. Springer.

Esther Gan, Yiran Zhao, Liying Cheng, Mao Yancan, Anirudh Goyal, Kenji Kawaguchi, Min-Yen Kan, and Michael Shieh. 2024. Reasoning robustness of LLMs to adversarial typographical errors. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10449–10459, Miami, Florida, USA. Association for Computational Linguistics.

Google DeepMind. 2025. Gemini 3 Pro model card.

Dorit Hadar Shoval. 2025. Artificial intelligence in higher education: Bridging or widening the gap for diverse student populations? *Education Sciences*, 15(5).

Emma Harvey, Allison Koenecke, and Rene F Kizilcec. 2024. Towards an educator-centered method for measuring bias in large language model-based chatbot tutors. In *AI for Education: Bridging Innovation and Responsibility at the 38th AAAI Annual Conference on AI*.

Emma Harvey, Allison Koenecke, and Rene F. Kizilcec. 2025. "Don't forget the teachers": Towards an educator-centered understanding of harms from large language models in education. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Neil T Heffernan and Cristina Lindquist Heffernan. 2014. The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, 24(4):470–497.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Kenneth Holstein and Shayan Doroudi. 2021. Equity and artificial intelligence in education: Will "AIEd" amplify or alleviate inequities in education? *Preprint*, arXiv:2104.12920.

Hyoungwook Jin, Yoonsu Kim, Dongyun Jung, Seungju Kim, Kiyoon Choi, Jinho Son, and Juho Kim. 2025. Investigating large language models in diagnosing students' cognitive skills in math problem-solving. *arXiv preprint arXiv:2504.00843*.

Rogers Kaliisa, Kamila Misiejuk, Sonsoles López-Pernas, and Mohammed Saqr. 2026. How does artificial intelligence compare to human feedback? a meta-analysis of performance, feedback perception, and learning dispositions. *Educational Psychology*, 46(1):80–111.

Blanka Klimova and Marcel Pikhart. 2025. Exploring the effects of artificial intelligence on student and academic well-being in higher education: A mini-review. *Frontiers in Psychology*, 16:1498132.

Ekaterina Kochmar, Kaushal Maurya, Kseniia Petukhova, KV Aditya Srivatsa, Anaïs Tack, and Justin Vasselli. 2025. Findings of the BEA 2025 shared task on pedagogical ability assessment of AI-powered tutors. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 1011–1033, Vienna, Austria. Association for Computational Linguistics.

Tianqi Kou, Dana Calacci, and Cindy Lin. 2025. Dead zone of accountability: Why social claims in machine learning research should be articulated and defended. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pages 1501–1512.

Stefan Küchemann, Karina E Avila, Yavuz Dinc, Chiara Hortmann, Natalia Revenga, Verena Ruf, Niklas Stausberg, Steffen Steinert, Frank Fischer, Martin Fischer, and 1 others. 2025. On opportunities and challenges of large multimodal foundation models in education. *npj Science of Learning*, 10(1):11.

Ehsan Latif, Zirak Khan, and Xiaoming Zhai. 2025. SketchMind: A multi-agent cognitive framework for assessing student-drawn scientific sketches. *arXiv preprint arXiv:2507.22904*.

Hao-Ping (Hank) Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. 2025a. The impact of generative ai on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Jimin Lee, Steven-Shine Chen, and Paul Pu Liang. 2025b. Interactive Sketchpad: A multimodal tutoring system for collaborative, visual problem-solving. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25, New York, NY, USA. Association for Computing Machinery.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. 2024. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3245–3276, Mexico City, Mexico. Association for Computational Linguistics.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*.

Rabeeh Karimi Mahabadi, Sanjeev Satheesh, Shrimai Prabhumoye, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2025. Nemotron-CC-Math: A 133 billion-token-scale high quality math pretraining dataset. *Preprint*, arXiv:2508.15096.

Janet Metcalfe. 2017. Learning from errors. *Annual Review of Psychology*, 68:465–489.

Masato Mita, Keisuke Sakaguchi, Masato Hagiwara, Tomoya Mizumoto, Jun Suzuki, and Kentaro Inui. 2024. Towards automated document revision: Grammatical error correction, fluency edits, and beyond. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications*

(BEA 2024), pages 251–265, Mexico City, Mexico. Association for Computational Linguistics.

Pam A. Mueller and Daniel M. Oppenheimer. 2014. The pen is mightier than the keyboard: Advantages of longhand over laptop note taking. *Psychological Science*, 25(6):1159–1168. PMID: 24760141.

Oikantik Nath, Hanani Bathina, Mohammed Safi Ur Rahman Khan, and Mitesh M Khapra. 2025. Can vision-language models evaluate handwritten math? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14784–14814, Vienna, Austria. Association for Computational Linguistics.

Thu Phuong Nguyen, Duc M. Nguyen, Hyotaek Jeon, Hyunwook Lee, Hyunmin Song, Sungahn Ko, and Taehwan Kim. 2025. VEHME: A vision-language model for evaluating handwritten mathematics expressions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31793–31813, Suzhou, China. Association for Computational Linguistics.

Behnam Parsaeifard, Martin Hlosta, and Per Bergamin. 2025. Automated grading of students' handwritten graphs: A comparison of meta-learning and vision-large language models. *arXiv preprint arXiv:2507.03056*.

Keiran Paster, Marco Dos Santos, Zhangir Azerbayev, and Jimmy Ba. 2024. OpenWebMath: An open dataset of high-quality mathematical web text. In *The Twelfth International Conference on Learning Representations*.

Hendrik Radatz. 1979. Error analysis in mathematics education. *Journal for Research in mathematics Education*, 10(3):163–172.

Inioluwa Deborah Raji, I Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. 2022. The fallacy of AI functionality. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 959–972.

Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 33–44, New York, NY, USA. Association for Computing Machinery.

Bethany Rittle-Johnson, Rebecca Adler, Kelley Durkin, L Burleigh, Jules King, and Scott Crossley. 2025. Detecting math misconceptions: An AI benchmark dataset. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Works in Progress*, pages 20–24, Wyndham Grand Pittsburgh, Downtown, Pittsburgh, Pennsylvania, United States. National Council on Measurement in Education (NCME).

Alexis Ross and Jacob Andreas. 2025. Learning to make MISTAKEs: Modeling incorrect student thinking and key errors. *Preprint*, arXiv:2510.11502.

Nils-Jonathan Schaller, Yuning Ding, Andrea Horbach, Jennifer Meyer, and Thorben Jansen. 2024. Fairness in automated essay scoring: A comparative analysis of algorithms on German learner essays from secondary education. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 210–221, Mexico City, Mexico. Association for Computational Linguistics.

Judy Hanwen Shen and Alex Tamkin. 2026. How AI impacts skill formation. *arXiv preprint arXiv:2601.20245*.

John P Smith III, Andrea A. diSessa, and Jeremy Roschelle. 1994. Misconceptions reconceived: A constructivist analysis of knowledge in transition. *Journal of the Learning Sciences*, 3(2):115–163.

Neha Srikanth, Rupak Sarkar, Heran Mane, Elizabeth Aparicio, Quynh Nguyen, Rachel Rudinger, and Jordan Boyd-Graber. 2024. Pregnant questions: The importance of pragmatic awareness in maternal health question answering. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7253–7268, Mexico City, Mexico. Association for Computational Linguistics.

Kv Aditya Srivatsa, Kaushal Kumar Maurya, and Ekaterina Kochmar. 2025. LLMs cannot spot math errors, even when allowed to peek into the solution. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10914–10928, Suzhou, China. Association for Computational Linguistics.

Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. 2024. Exploring LLM prompting strategies for joint essay scoring and feedback generation. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 283–298, Mexico City, Mexico. Association for Computational Linguistics.

Keita Umejima, Takuya Ibaraki, Takahiro Yamazaki, and Kuniyoshi L Sakai. 2021. Paper notebooks vs. mobile devices: Brain activation differences during memory retrieval. *Frontiers in Behavioral Neuroscience*, 15:634158.

Angelina Wang, Sayash Kapoor, Solon Barocas, and Arvind Narayanan. 2024. Against predictive optimization: On the legitimacy of decision-making algorithms that optimize predictive accuracy. *ACM J. Responsib. Comput.*, 1(1).

Ryan Yixiang Wang, Matthew Finlayson, Luca Soldaini, Swabha Swayamdipta, and Robin Jia. 2025. Teaching models to understand (but not generate) high-risk data. In *Second Conference on Language Modeling*.

Niall Winters, Rebecca Eynon, Anne Geniets, James Robson, and Ken Kahn. 2020. Can we avoid digital structural violence in future learning systems? *Learning, Media and Technology*, 45(1):17–30.

Yibo Yan, Jiamin Su, Jianxiang He, Fangteng Fu, Xu Zheng, Yuanhuiyi Lyu, Kun Wang, Shen Wang, Qingsong Wen, and Xuming Hu. 2025. A survey of mathematical reasoning in the era of multimodal large language model: Benchmark, method & challenges. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11798–11827, Vienna, Austria. Association for Computational Linguistics.

Xinyan Yu, Sewon Min, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. CREPE: Open-domain question answering with false presuppositions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10457–10480, Toronto, Canada. Association for Computational Linguistics.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, Peng Gao, and Hongsheng Li. 2024. MATHVERSE: Does your multi-modal LLM truly see the diagrams in visual math problems? In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part VIII*, page 169–186, Berlin, Heidelberg. Springer-Verlag.

A Model Performance by Question Type

In the main text, Figure 3 shows that the gap in VLM performance between erroneous and non-erroneous student responses is primarily driven by content description QA. Figure 10 expands upon that finding, by disaggregating content description QA's overall pattern into finer-grained question categories and showing results for all 11 VLMs. Across these expanded plots, we see that the performance gap between erroneous and non-erroneous student responses persists across finer-grained content description QA categories.

B Language Model-Assisted Data Annotation

For each LM-assisted labeling task, we iteratively developed prompts that yield solid performance on small samples, before validating our final prompts on larger samples. The main text details the performance of each prompt on the intended task.

B.1 Student Error

One of our main findings, **F1**, pertains to how models perform differently between student responses that contain errors versus those that do not. To determine whether a student response contains an error or not, we rely on teachers' free-form descriptions of student error. Since teachers' written responses may span a variety of phrasings, we use GPT-5-mini to decisively label whether the teacher indicates that the student response contains an error. Here, `ans` is the teacher's answer to the question, *What errors does the student make in their response?* We use the following prompt:

```
When asked about what errors a student
makes in their response to a math problem,
a teacher writes, '{ans}'. Based on the
teacher's feedback, does the student make
any error? Respond 'yes' or 'no'.
```

B.2 Finer-grained Correctness & Error Questions

In §8.1, we discuss how correctness & error QA span both binary assessments of specific student errors and more-open-ended questions. We use GPT-5-mini as an annotator to label whether each Correctness & Error `question` is `binary` or `other`. We use the following prompt:

```
Is the following question a binary
question that asks whether a student does
something correctly or not?
Question: '{question}'
Decide whether the question above is a
binary question that judges a student's
correctness. Your response should start
with 'Yes' or 'No':
```

B.3 Binary Student Correctness

In §8.2, we're interested in investigating whether models tend to under- or over-report student error on binary correctness & error questions. We use GPT-5-mini to annotate whether `binary` QA's questions and gold answers indicate that student is correct or incorrect. We use the following prompt:

```
Teacher A is examining a student's
solution to a math problem. Teacher B asks
Teacher A, '{question}'
Teacher A says, '{answer}'.
Does this exchange indicate that the
student's solution has an error? Respond
"yes" or "no":
```

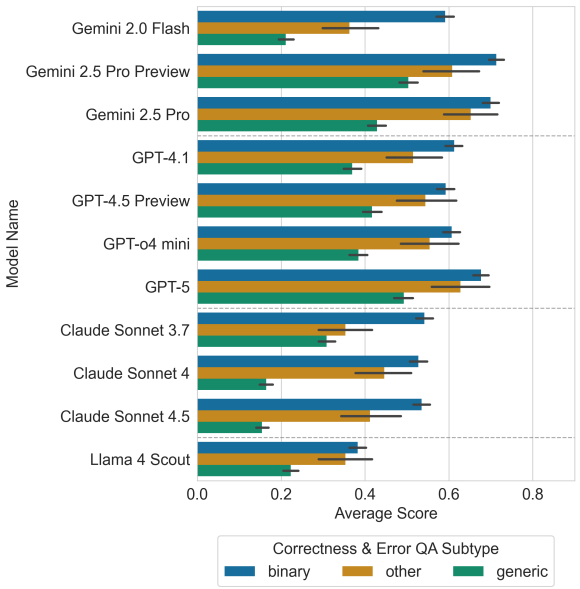


Figure 9: VLMs' performance across different subtypes of correctness & error questions, as defined in §8.1.

C Natural Language Description Experiments

In §7, we investigate whether the inclusion of textual descriptions of students' work can support VLMs' abilities to make higher-level inferences around students' correctness and errors.

C.1 Prompts

Our prompt that adds in natural language descriptions/captions is intuitive and simple:

```
Description of image:
{caption}

Answer the following question: {question}
```

To generate descriptions or captions using language models, we use the following prompt: `Describe the Student Response on the right side of the image in one paragraph.`

D Additional Correctness & Error Results

The relative performance ranking of `binary`, `other`, and `generic` correctness & error QA is consistent across all 11 VLMs (Figure 9).

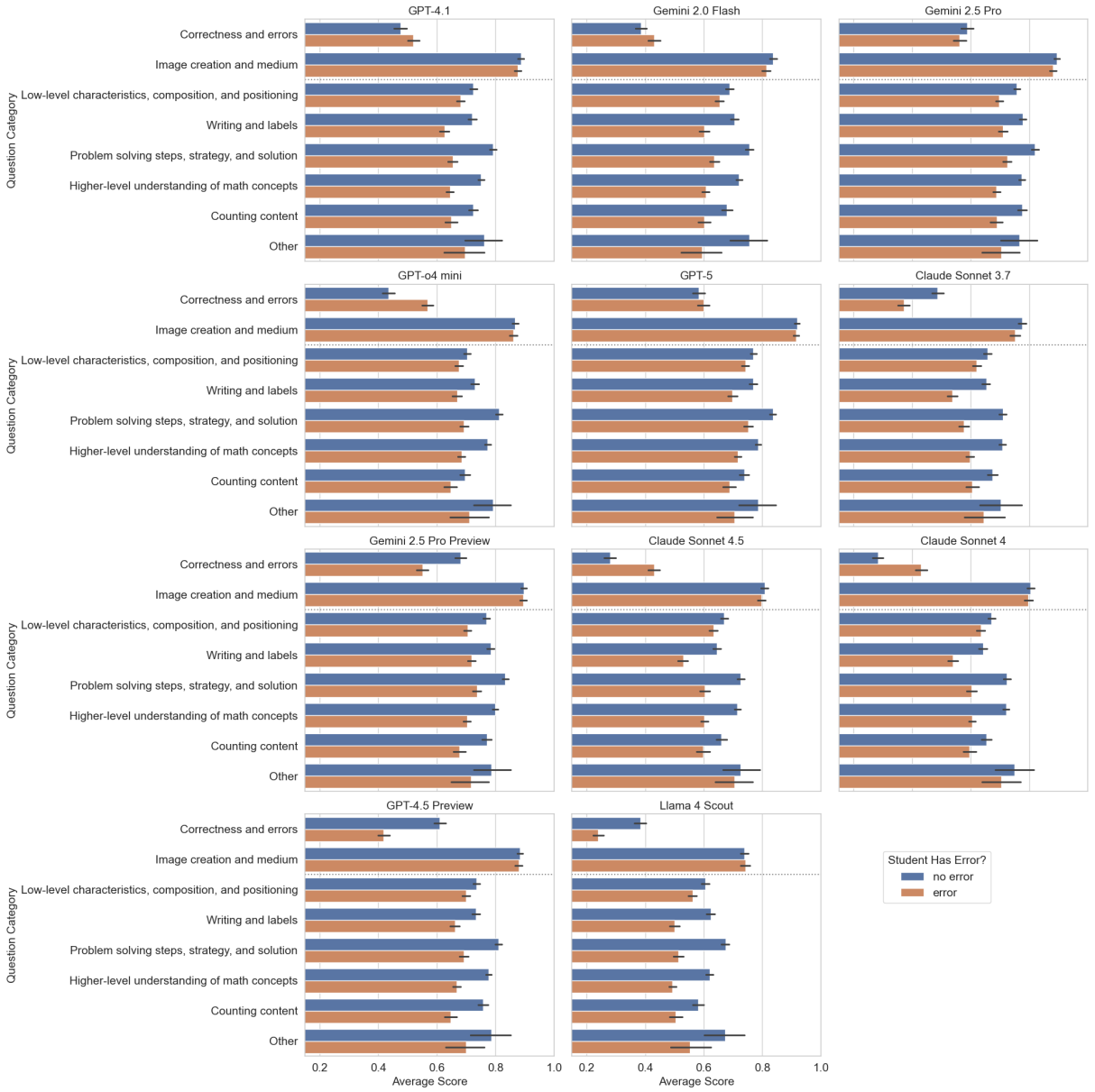


Figure 10: An expanded version of Figure 3, showing which question categories contribute to the gap in VLM performance between student responses that contain errors versus those that do not. Questions below the dotted line in each subplot are content description QA.