Beyond Aggregate Accuracy: Continuous Evaluation of Residual Failures in Image-Required Educational Mathematics

Ethan Croteau ECROTEAU@WPI.EDU and **Neil Heffernan** NTH@WPI.EDU
Worcester Polytechnic Institute, Worcester, MA, USA

Abstract

Rapid model progress can make educational benchmark results stale, but responsible evaluation requires more than refreshing aggregate scores. We revisit a benchmark of 376 curriculum-authentic, image-required middle-school mathematics items using GPT-5.5 and compare its behavior with representative prior models. In the *with-images* condition, accuracy increases from about 88% for GPT-5.4 to about 93% for GPT-5.5, while the strict never-solved set decreases from 31 items to 17. In contrast, *without-images* performance remains near zero, confirming that the benchmark remains genuinely image-required. We then audit this residual set and find that its composition clarifies what the remaining failures mean: they include inaccessible-image cases, insufficient-context items, answer-format mismatches, and benchmark-quality issues rather than only generic visual-reasoning failures. We treat representation-sensitive probes as future work, using the audit to identify which residual items are appropriate probe candidates. We argue that continuous evaluation of educational AI systems should pair aggregate performance updates with residual item audits that distinguish model limitations from dataset, scoring, accessibility, and task-specification issues.

Keywords: AI in education, multimodal evaluation, image-required mathematics, benchmark design, accessibility, residual failure analysis, continuous evaluation

1. Introduction

Generative and multimodal AI models, including multimodal large language models, are increasingly being explored for educational applications such as tutoring, feedback, assessment, and content support (Heilala et al., 2025; Dong et al., 2024). In mathematics, many tasks depend on diagrams, graphs, number lines, tables, geometric figures, and other visuals that are not merely decorative but required for a correct solution. A model may appear mathematically capable while still failing to count task-relevant objects, read a scale, bind a label to the correct visual element, or preserve a spatial relation needed for reasoning.

Recent work on curriculum-authentic, image-required middle-school mathematics established this problem setting and showed that multimodal systems vary substantially in correctness, refusal behavior, and run-to-run stability under *with-images* and *without-images* conditions (Croteau and Heffernan, 2026). As stronger models become available, however, benchmark conclusions can change quickly. This creates a challenge for responsible AI evaluation in education: results need to be refreshed, but they also need to be interpreted in ways that preserve educational meaning rather than reducing evaluation to a leaderboard.

Aggregate accuracy is therefore incomplete. A large improvement in the *with-images* condition may show real progress, but it does not explain the remaining failures. Residual

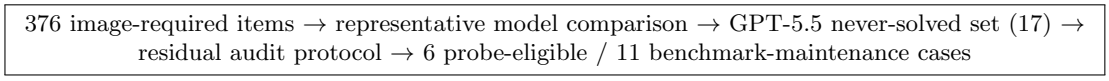

Figure 1: Continuous residual-audit workflow used in this paper.

errors may reflect visual perception, mathematical reasoning, missing problem context, answer-format mismatch, scoring assumptions, or defects in the benchmark item itself. These distinctions matter because each points to a different remedy: image validation, alternative representation, task revision, scoring adjustment, or model improvement.

This paper uses strict never-solved items as a tool for *continuous evaluation*. Rather than treating a benchmark as static, we ask how its conclusions change under a newer model and what the remaining errors reveal about the model, the benchmark, and the intended educational task. We address three questions:

1. How does GPT-5.5 change aggregate correctness and strict never-solved counts on image-required mathematics relative to representative earlier models?
2. What kinds of items remain never solved under GPT-5.5 in the *with-images* condition?
3. Which remaining items are plausible candidates for future representation-sensitive analysis using structured image descriptions or executable reconstructions?

Our results show substantial progress but also the need for careful residual-item auditing. Under GPT-5.5, the strict never-solved set is much smaller than under GPT-5.4. Yet the 17 remaining items are not a uniform class of visual failures: they include inaccessible-image, insufficient-context, answer-format, and benchmark-quality cases. The educational meaning of a residual error therefore depends on whether it reflects model behavior, image accessibility, task specification, scoring, or item quality.

The paper makes three contributions. First, it reports an updated model comparison on a 376-item image-required mathematics benchmark under both *with-images* and *without-images* conditions. Second, it presents an item-level residual-audit protocol and applies it to the 17 problems that remain never solved under GPT-5.5. Third, it uses the audit to separate six future representation-probe candidates from 11 cases that instead point to benchmark maintenance, scoring, or task-specification issues.

Figure 1 summarizes how the analysis moves from aggregate model comparison to residual auditing and probe-eligibility coding.

2. Related Work

2.1. Multimodal evaluation on visually grounded mathematics

Recent multimodal benchmarks have documented progress on image-based reasoning and mathematical problem solving (Lu et al., 2024; Zhang et al., 2024). However, broad benchmark summaries can obscure item-level failures that matter in educational settings, especially when diagrams, graphs, scales, or geometric figures contain task-critical information. Prior work on curriculum-authentic middle-school mathematics has emphasized that some tasks are genuinely image-required rather than merely image-supported, making them a relevant testbed for AI in education (Croteau and Heffernan, 2026; Heffernan and Heffernan,

2014). The present paper builds on this setting by examining how residual never-solved items change after a model update.

2.2. Representation, accessibility, and responsible evaluation

Representations shape mathematical interpretation and inference (Ainsworth, 2006; Mayer, 2020). This is especially relevant for visually grounded mathematics, where task success may depend on measurement, correspondence, proportion, spatial relations, spatial decomposition, or geometric structure. Learning-science accounts of diagrams emphasize that solvers must select relevant visual information and coordinate it with symbolic and verbal representations (Larkin and Simon, 1987; Stylianou and Silver, 2004). The same issue matters for accessibility: alternative forms such as descriptions should preserve task-relevant structure rather than merely summarize appearance (W3C Accessibility Guidelines Working Group, 2024; W3C Web Accessibility Initiative, 2022). Responsible AI in education therefore requires evaluation methods that go beyond aggregate accuracy to examine reliability, inclusiveness, and the educational meaning of residual failures.

3. Benchmark and Evaluation Setting

We evaluate an image-required middle-school mathematics benchmark of 376 curriculum-authentic items for which the image is required to solve the problem correctly (Croteau and Heffernan, 2026). Each item is evaluated across three repeated attempts per model and condition, yielding 1128 runs per model-condition pair. We report both a *with-images* condition, in which the model receives the original problem text and image, and a *without-images* condition, in which the image is withheld.

The analysis has two levels. First, we compare aggregate benchmark performance across representative model generations. Second, we audit residual items for GPT-5.5. We use the term *strict never-solved item* for an item with zero correct answers across the three repeated runs for a given model and condition. This definition focuses on robust failures rather than unstable cases that are solved in some attempts but not others.

Our main comparison includes five representative systems: GPT-4o as an earlier multimodal baseline, o4-mini as a pre-GPT-5 reasoning model, GPT-5 as the first GPT-5-line model in this sequence, GPT-5.4 as the previous strongest model in our analysis, and GPT-5.5 as the newest model in this comparison. We use shortened labels for readability; the two focal model snapshots are `gpt-5.4-2026-03-05` and `gpt-5.5-2026-04-23`. The selected models are not intended as an exhaustive leaderboard, but as representative points for the continuous-evaluation argument.

4. Method

4.1. Aggregate model comparison

For each model and condition, we compute correctness across all 1128 runs, corresponding to 376 items evaluated three times each. Correctness is determined by extracting the final answer from each response and comparing it with the benchmark answer using the existing scoring pipeline. We report both run-level correctness and the number of strict never-solved

items. These metrics capture different aspects of performance: aggregate correctness reflects overall capability, while strict never-solved counts identify items that remain consistently unresolved across repeated attempts.

4.2. Strict never-solved analysis

Our residual audit focuses on the strict never-solved items for GPT-5.5 in the *with-images* condition. We compare this set with the corresponding GPT-5.4 set to identify three groups: items that remain never solved under both models, items that leave the never-solved set under GPT-5.5, and items that newly appear under GPT-5.5. This delta analysis supports the paper's continuous-evaluation framing by showing not only whether performance improves, but how the residual failure set changes.

4.3. Residual failure taxonomy

We classify each GPT-5.5 strict never-solved item into one top-level bucket and one finer-grained subcode. The taxonomy, summarized in Table 3, distinguishes inaccessible-image cases, visual misinterpretations, answer-format mismatches, missing-context items, and apparent problem-quality issues. Coding draws on attempt summaries, extracted answers, missing-information fields, image characteristics, and manual review. Image characterization includes transparency-related statistics, which help distinguish genuine visual misinterpretation from cases where the source image is effectively unusable.

Operationally, the residual audit follows four steps: inspect whether the visual input is accessible under the evaluation pipeline; compare repeated attempts for stable missing information, wrong visual bindings, or answer-format mismatches; check whether the source item or scoring key supplies enough information to justify the expected answer; and assign both a failure bucket and an appropriate next action. We therefore present the taxonomy as a reusable audit protocol for small residual sets, not as a fully validated measurement instrument.

4.4. Probe eligibility coding

Finally, we code whether each residual item is a plausible target for future representation-sensitive probing. An item is marked probe-eligible when changing the image representation could plausibly clarify whether the original failure was tied to image encoding. Under the current taxonomy, this includes *Unseeable* and *Seen-but-misinterpreted* items. Items classified as *Missing context*, *Problem issue*, or *Seen and understood* are excluded because alternative image representation is unlikely to address the apparent failure mechanism.

We do not report representation-sensitive probe results in this paper. Instead, we use probe eligibility to identify the appropriate scope for future work. Structured image descriptions and executable reconstructions have been drafted for the six eligible residual items, making the next evaluation stage concrete without treating it as a completed experiment.

5. Results

5.1. Aggregate gains remain image-dependent

Table 1 reports run-level correctness and strict never-solved counts for five representative models. In the *with-images* condition, performance increases substantially across the model sequence: GPT-4o answers 37.7% of runs correctly, while GPT-5.5 answers 93.1% correctly. The strict never-solved count falls from 197 items to 17 items over the same comparison.

The *without-images* condition remains near zero across all reported models, ranging from 0.9% to 2.7% correct. Thus, the strong *with-images* performance of GPT-5.5 does not indicate that the benchmark has become text-solvable; the visual channel remains essential.

Table 1: Representative model results on the 376-item image-required benchmark. Accuracy percentages are rounded to one decimal place and computed over 1128 runs, corresponding to three attempts per item. Never-solved counts are item-level counts out of 376.

	With images		**Without images**	
Model	**Acc. %**	**Never-solved**	**Acc. %**	**Never-solved**
GPT-4o	37.7	197	0.9	371
o4-mini	58.2	117	2.7	365
GPT-5	64.9	95	2.1	366
GPT-5.4	87.9	31	1.7	368
GPT-5.5	93.1	17	2.0	368

5.2. The strict never-solved set shrinks and changes

The strict never-solved set decreases from 31 items under GPT-5.4 to 17 items under GPT-5.5 in the *with-images* condition, a reduction of 14 items (45.2%). As a share of the benchmark, this corresponds to a decrease from 8.2% to 4.5% of items.

Table 2 shows that 15 items remain never solved under both models, 16 leave the strict never-solved set under GPT-5.5, and two newly appear. This supports the continuous-evaluation framing: model updates can materially alter not only aggregate performance but also which items require residual analysis.

Table 2: Delta comparison between GPT-5.4 and GPT-5.5 strict never-solved sets in the *with-images* condition.

Comparison	**Count**
Never solved by GPT-5.4	31
Never solved by GPT-5.5	17
Intersection	15
Only never solved by GPT-5.4	16
Only never solved by GPT-5.5	2
Net reduction from GPT-5.4 to GPT-5.5	14

5.3. The GPT-5.5 residual items are heterogeneous

The remaining set is small, but its composition is informative. Table 3 shows that the largest categories are *Missing context* and *Problem issue*, with five items each. Four items are *Unseeable*, all involving transparent-image issues. Only two items are classified as *Seen-but-misinterpreted*, and one is classified as *Seen and understood* because the model appears to understand the task but misses the expected answer format. Because the denominator is only 17 items, counts are the primary evidence; percentages and confidence intervals are included only to make the uncertainty visible.

Table 3: Item-level taxonomy of the 17 strict never-solved items for GPT-5.5 in the *with-images* condition.

Bucket	**Count**	**%**	**95% CI**	**Representative subcodes or issues**
Unseeable	4	23.5	9.6–47.3	Transparent image; partial alpha; visually inaccessible figure.
Seen-but-misinterpreted	2	11.8	3.3–34.3	Visual misinterpretation; snap-cube volume.
Seen and understood	1	5.9	1.0–27.0	Apparent task comprehension with answer-format mismatch.
Missing context	5	29.4	13.3–53.1	Equation; show reasoning; geometric relationship; unspecified dimension.
Problem issue	5	29.4	13.3–53.1	Incorrect answer; mislabeled figure; inconsistent item information.

Wilson binomial intervals are shown to emphasize imprecision in this small residual set; they are not used for hypothesis testing.

These categories change how the remaining failures should be interpreted. The transparent image cases are not ordinary geometry or arithmetic errors; the visual input is effectively inaccessible under the evaluation pipeline. The missing context cases appear better understood as insufficient context items than as image-required reasoning failures. The problem issue cases identify fixable benchmark or source item problems, such as apparent answer key, labeling, or image-text inconsistencies. As the never-solved set shrinks, the residual audit increasingly functions as a benchmark quality and accessibility audit rather than only as a model-error analysis.

5.4. Only six residual items are plausible future probe targets

Table 4 summarizes the probe-eligibility decision. Six of the 17 residual items are plausible targets for future representation-sensitive probing: the four *Unseeable* items and the two *Seen-but-misinterpreted* items. The other 11 are excluded because their apparent failure mechanisms involve missing context, answer format, or item quality rather than alternative image representation.

Table 4: Probe eligibility for the 17 GPT-5.5 residual items. The probe itself is deferred to future work.

Status	Count	Included buckets / item types
Probe-eligible	6	Unseeable; seen-but-misinterpreted.
Not probe-eligible	11	Missing context; problem issue; seen and understood.

This probe-eligibility decision helps prevent over-attributing all residual errors to image representation or visual reasoning. Treating every never-solved item as a visual representation failure would conflate inaccessible images, genuine visual misinterpretation, missing information, answer format mismatch, and benchmark quality problems. The smaller probe-eligible set gives a cleaner target for future work using structured image text or executable reconstructions.

5.5. Illustrative residual cases

Three cases illustrate the taxonomy's interpretive role. In one prism-volume item, GPT-5.5 declares the problem unsolvable in all three runs and asks for base-area information or measurements; image analysis shows that the source image is overwhelmingly transparent, supporting the *Unseeable* classification. Pedagogically, the item depends on coordinating a three-dimensional diagram with volume structure, but the accessible visual information needed for that coordination is not available to the model. In the wire picture-frame item, the model consistently treats all visible scallops as semicircles and returns the same incorrect perimeter, making it a clearer visual-misinterpretation case tied to counting and decomposing repeated curved components. In a surface-area item, the model consistently returns 72 square centimeters while the benchmark answer is 60; manual review suggests a possible mismatch among the visible figure, labels, and expected answer, so the item should be audited before the result is treated as straightforward model weakness.

6. Discussion

The small GPT-5.5 residual set shifts how persistent failure can be interpreted. Rather than revealing a large remaining class of generic visual-reasoning failures, the residual items increasingly function as a benchmark audit. Four involve transparent or effectively unseeable images, five appear to lack necessary context, and five expose apparent problem-quality issues such as incorrect, mislabeled, or inconsistent item information. Continuous evaluation is therefore valuable not only because model performance changes, but also because stronger models can make benchmark defects and task-specification problems more visible.

This distinction matters for educational AI evaluation. If a model fails because an image is effectively transparent, the appropriate response may be accessibility-focused preprocessing, image validation, or alternative representation. If a problem omits necessary information, the issue is task specification rather than model reasoning alone. If a mathematically meaningful response does not match the expected answer format, the scoring pipeline may

need review. If the source item is mislabeled or internally inconsistent, the benchmark itself should be corrected before the result is treated as evidence of model weakness.

The *without-images* results further clarify the interpretation. Across the reported models, *without-images* accuracy remains near zero, indicating that strong *with-images* performance by GPT-5.5 does not make the benchmark text-solvable. The visual channel remains essential, which makes residual item auditing especially important.

The deferred representation-probe artifacts fit naturally into this picture. Structured image descriptions and executable reconstructions are promising for the six residual items where image representation is plausibly central to the failure, but they are not appropriate for missing-context, answer-format, or problem-quality cases. Thus, the taxonomy is not only descriptive; it helps determine what kind of follow-up analysis is methodologically appropriate.

For responsible evaluation, the practical implication is that benchmark maintenance should pair aggregate score updates with residual item audits. As models improve, the remaining failure set may become smaller but more concentrated in edge cases, accessibility failures, scoring assumptions, and item-quality problems. Treating those cases carefully is important before using benchmark results to justify educational deployment or to draw broad conclusions about model competence.

7. Limitations and Future Work

This study has several limitations. First, the residual taxonomy involves analyst judgment. Although the categories are grounded in attempt summaries, image characteristics, and item review, the coding should be treated as an initial audit rather than a fully validated scheme. We do not report inter-rater reliability for this 17-item residual set; larger follow-up audits should use multiple coders and report agreement statistics such as Cohen's kappa or Krippendorff's alpha (Cohen, 1960; Krippendorff, 2011). Second, the benchmark items are drawn from public curriculum materials and the evaluated models are proprietary and rapidly changing, so contamination or memorization cannot be ruled out. Third, the analysis is restricted to one educational domain, one image-required mathematics benchmark, and one proprietary model family; it does not establish generalizability to open-weight models, other educational levels, or other subject areas.

This work does not report a full representation-sensitive probe. Although we drafted structured image descriptions and executable reconstructions for the six probe-eligible GPT-5.5 residual items, evaluating those artifacts would require additional model runs, scoring, and analysis. We therefore treat them as concrete future work rather than evidence in the present paper.

Future work should evaluate the six probe-eligible items under controlled representation variants, such as the original image, structured image text, executable reconstruction, reconstructed PNG, and possibly SVG-as-text. This would help distinguish visual-encoding failures from failures that persist after task-relevant structure has been externalized. A second direction is benchmark cleaning: insufficient-context, answer-format, and problem-quality cases should be reviewed before residual errors are attributed to model limitations.

8. Conclusion

Rapid model progress can materially change conclusions on image-required educational mathematics. In this evaluation, GPT-5.5 substantially improves *with-images* performance and reduces the strict never-solved set from 31 items to 17, while *without-images* performance remains near zero. The remaining items are few but informative: they include inaccessible-image cases, insufficient-context items, answer-format mismatches, and benchmark-quality issues rather than only generic visual-reasoning failures. Responsible evaluation of educational AI should therefore pair aggregate performance updates with residual item audits that clarify what kind of failure remains and what kind of intervention is warranted.

Acknowledgments

We would like to thank NSF (e.g., 2118725, 2118904, 1950683, 1917808, 1931523, 1940236, 1917713, 1903304, 1822830, 1759229, 1724889, 1636782, & 1535428), IES (e.g., R305N210049, R305D210031, R305A170137, R305A170243, R305A180401, & R305A120125), GAANN (e.g., P200A180088 & P200A150306), EIR (U411B190024 & S411B210024), ONR (N00014-18-1-2768), NIH (R44GM146483), and Schmidt Futures. None of the opinions expressed here are those of the funders.

References

Shaaron Ainsworth. Deft: A conceptual framework for considering learning with multiple representations. *Learning and Instruction*, 16(3):183–198, 2006.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

Ethan Croteau and Neil Heffernan. Seeing is solving: MLLMs, reasoning, and refusal in visual math. *Journal of Educational Data Mining*, 18(1):244–285, 2026. doi: 10.5281/zenodo.19420820.

Bingyu Dong, Jie Bai, Tao Xu, and Yun Zhou. Large language models in education: A systematic review. In *2024 6th International Conference on Computer Science and Technologies in Education (CSTE)*, pages 131–134, 2024. doi: 10.1109/CSTE62025.2024.00031.

Neil T Heffernan and Cristina Lindquist Heffernan. 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, 2014. doi: 10.1007/s40593-014-0024-x.

Ville Heilala, Roberto Araya, and Raija Hämäläinen. Beyond text-to-text: An overview of multimodal and generative artificial intelligence for education using topic modeling. In *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing*, pages 54–63, 2025.

Klaus Krippendorff. Computing Krippendorff's Alpha-Reliability. Working Paper 43, University of Pennsylvania, Annenberg School for Communication, Philadelphia, PA, January 2011. URL https://repository.upenn.edu/handle/20.500.14332/2089. Postprint version.

Jill H Larkin and Herbert A Simon. Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, 11(1):65–100, 1987.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*, Vienna, Austria, 2024. OpenReview.net. URL https://openreview.net/forum?id=KUNzEQMWU7.

Richard E. Mayer. *Multimedia Learning*. Cambridge University Press, Cambridge, UK, 3rd edition, 2020.

Despina A Stylianou and Edward A Silver. The role of visual representations in advanced mathematical problem solving: An examination of expert-novice similarities and differences. *Mathematical thinking and learning*, 6(4):353–387, 2004. doi: 10.1207/s15327833mtl0604_1.

W3C Accessibility Guidelines Working Group. Web content accessibility guidelines (wcag) 2.2. W3C Recommendation, December 2024. URL https://www.w3.org/TR/2024/REC-WCAG22-20241212/. Latest version: https://www.w3.org/TR/WCAG22/.

W3C Web Accessibility Initiative. Images tutorial: Complex images. https://www.w3.org/WAI/tutorials/images/complex/, 2022. Updated 17 January 2022. Accessed 27 Apr 2026.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. MATHVERSE: Does your multi-modal LLM truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186, Cham, 2024. Springer Nature Switzerland.