PWW-Bench: Probing Visual Mathematical Reasoning with Proofs Without Words

Sabrina Reguyal [*1] **Rahul Saha** [*2] **Alan Li** [*2] **Atharva Sehgal** [2]

Abstract

Human mathematics is profoundly visual at the level of intuition, discovery, and communication, yet current math vision benchmarks focus on shallow calculation-based tasks rather than deeper modes of visual reasoning. To address this critical gap, we introduce **PWW-Bench**, a benchmark for multimodal mathematical reasoning of vision-language models (VLMs) built from the *Proofs Without Words* (PWW) collection–403 single-page diagrams that illustrate proofs of mathematical theorems with little or no prose. In particular, PWW-Bench tests whether a model can *understand proof-based mathematical reasoning from a picture* by asking the model to identify the theorem, point to the geometric components that make the argument work, and recover the logical chain the figure encodes. To support VLM evaluation on PWW-Bench, we define a rubric-based grading system using an LLM judge. We then conduct pilot experiments on a balanced subset of PWW-Bench with human-calibrated rubrics, evaluating 8 frontier and open-weight VLMs. We find that all models exhibit profound visual understanding failures: even when provided the correct theorem, VLMs do not reconstruct the correct proof. Furthermore, we develop an interface for crowdsourcing rubric calibration across the entire benchmark.

1. Introduction

Mathematical intuition often operates through nonverbal, visual thought. A prominent description of this cognitive mechanism appears in Pólya's book, *How To Solve It*, in which he expounds on the utility of drawing figures for problem solving (Pólya, 1957). Indeed, visualization plays a key role in both mathematical research and education.

In the research domain, visualization is often useful as a means for intuitively manipulating mathematical concepts, even if they do not map explicitly onto geometric objects (Hadamard, 1945). Furthermore, computer-generated visuals have recently contributed to work in topology, geometry, and chaos theory by allowing mathematicians to gain high-level insights into complex, unintuitive phenomena such as everting of the 2-sphere (Mancosu, 2005). Overall, many aspects of mathematical discovery and communication involve some form of image-based representation, whether through observing structure in visualized mathematical objects or plotted data, or by communicating the ideas to an audience (Figure 1). Similarly, in the pedagogical domain, illustrations of mathematical objects and the relationships between them are often used for the purpose of communicating intuition to students (Guzman, 2002; Zimmermann & Cunningham, 1991). As such, a visual math benchmark that omits image-based reasoning and intuition misses out on an integral part of mathematical discovery and communication.

Modern visual math benchmarks such as MathVista (Lu et al., 2024), MathVerse (Zhang et al., 2024), MMMU (Yue et al., 2024), and We-Math (Qiao et al., 2024) test computation given a problem statement: the caption tells the model what to compute, and the figure provides parameters or a partial constraint. In contrast, PWW-Bench tests visual proof-based mathematical reasoning along three axes: a VLM must (i) identify which theorem the diagram is demonstrating, (ii) comprehend the structure of the proof image (e.g. panels, figure components, shading, equalities, transformations), and (iii) articulate the visual chain of reasoning that the picture encodes. Ultimately, we evaluate a crucial mechanism of proof-based reasoning used by working mathematicians which is not captured by any existing benchmark.

Contributions.

- **Comprehensive dataset.** 403 PWW problems from three Nelsen volumes (Nelsen, 1993) (Nelsen, 2000) (Nelsen, 2015). These visual proofs span six categories: geometry & algebra; trigonometry, calculus & analytic

[1]Stanford University, Stanford, California, USA [2]The University of Texas at Austin, Austin, Texas, USA. Correspondence to: Rahul Saha <rahul.saha@utexas.edu>.

The 3rd *AI for Math Workshop* at the 43rd International Conference on Machine Learning (ICML), Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

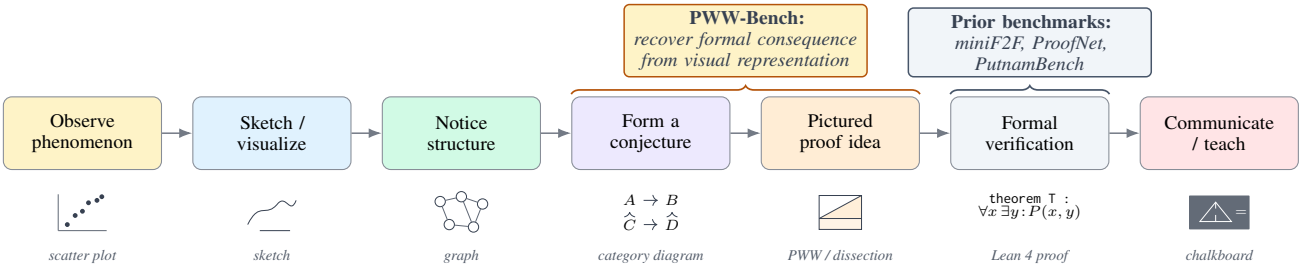

Figure 1. Where figures and intuition live in mathematical work, with representative visual types at each stage. Only *one* stage—formal verification—is fundamentally symbolic; the others are visual-heavy and use a wide variety of representations: data plots, free-form sketches, graphs, commutative diagrams, geometric dissections (PWW), and—when teaching or explaining—chalkboard diagrams. Standard math benchmarks evaluate the symbolic verification step in isolation, given a prose problem statement. PWW-Bench evaluates the harder transition the working mathematician makes constantly: extracting formal mathematical content directly from a visual representation.

geometry; inequalities; integers & integer sums; infinite series, linear algebra, & other topics. Furthermore, we identify five visual proof archetypes: dissection, lattice counting, nested figures, shaded-area equality, staged reveal. Each benchmark item comes with a cropped diagram and a canonical theorem statement.

- **Two task formulations.** The *diagram-only* formulation provides the model with only the figure; the *theorem-given* formulation provides both the figure and the canonical theorem statement. In both settings, the model is asked to state the proof. These two formulations distinguish an inability to identify the theorem from an inability to recover the theorem proof that is actually shown in the illustration.
- **Rubric-based judge.** The models are evaluated using a rubric graded on separate axes for theorem identification, visual grounding, and pictured-proof reconstruction on a scale of 0–3. This benchmark grading schema is developed in line with recent findings that LLM-as-judge agrees well with expert grading when a detailed reference rubric is supplied (Arora et al., 2025).

  To ensure quality of the LLM judge, we develop a calibration scheme in which, for each proof, the rubric is augmented with item-specific, human-checked visual proof metadata, which includes proof intuition and visual anchors. Furthermore, we write human-graded rubrics as additional system prompt-level calibration to encourage correct reasoning and to prevent prominent failure modes, such as recovering a flawed argument with a different valid argument of its own.
- **Pilot experiments.** We test the dataset and the calibrated rubric on a 15-problem subset balanced across archetypes, evaluating 8 VLMs (3 frontier, 5 open-weight). The pilot's role is to validate the rubric and judge before scaling the rubric calibration.
- **Findings.** The dominant frontier model failure is proof-reading error rather than theorem identification. Specifically, the model writes a proof of the correct identity through a geometric step that does not hold in the figure. The theorem-given formulation fixes most theorem identification failures but does not fix proof-reading failures, so the strongest models gain little from being told the theorem.

2. Task Definition

2.1. PWW Items

A PWW item is a tuple (I, T):

- I: a single image—a cropped diagram, possibly with multiple panels arranged in a grid.
- T: a canonical theorem statement (e.g., for Viviani: "the sum of perpendicular distances from any interior point of an equilateral triangle to its three sides equals the triangle's altitude").

We define two evaluation formulations.

Diagram only. Input is I alone, with the prompt: *"Identify the theorem this proof-without-words demonstrates, then describe the proof the diagram is communicating."* Output is free-form text. All three rubric axes are scored. Maximum total: 9.

Theorem given. Input is (I, T) with the prompt: *"Here is the theorem this proof-without-words demonstrates: $\langle T \rangle$. Describe the proof the diagram is communicating."* Axis 1 (theorem identification) is not scored, since the theorem is provided. Maximum total: 6.

2.2. Rubric

We score each response on three independent axes, each in $\{0, 1, 2, 3\}$:

- **Axis 1 (theorem identification).** Does the model name the theorem (or an accepted alias) the diagram demonstrates? Scored on the diagram-only formulation only.

- **Axis 2 (visual grounding).** Does the model name the right pieces of the diagram—the specific triangles, the shaded regions, the panel labels, the summation index—without hallucination? Scored on both formulations.

- **Axis 3 (pictured-proof reconstruction).** Does the model articulate the equality chain the figure encodes, using the mechanism the figure actually uses (rather than substituting a different valid argument)? Scored on both formulations.

An important relation of cross-axes performance is that a response that scores low on Axis 2 must not score high on Axis 3, as one cannot create a valid logical chain for objects you were not correctly identified. We therefore enforce this deterministically: $\text{axis3}_{\text{final}} = \min(\text{axis3}_{\text{raw}}, \text{axis2})$.

Post-rubric grading, we collect a closed-set *failure-tag* list (theorem misid, key element miss, alternative proof, hallucinated objects, invalid grounded chain, panel order, refusal) for qualitative analysis.

2.3. Rubric calibration.

In order to effectively grade individual proof problems, the LLM judge requires specific information about the details of each problem. Thus, for each VLM output, the LLM judge receives the tuple (I, T) along with a prose proof P, which gives intuition for the illustration; anchor list A, which gives the crucial geometric visual components for the proof; and not pictured N, which gives examples of theorems or proof steps to not count towards a successful proof. As an example, the dataset entry for Viviani's theorem (Figure 2) is shown in the box below.

Example rubric entry: viviani's theorem (#012)

theorem sum of perpendicular distances from any interior point of an equilateral triangle to its three sides equals the altitude.

anchors $\triangle ABC$ equilateral; interior point P; perpendiculars PD, PF, PG; auxiliary point Q;

proof Slide the triangle $\triangle ABC$ so the interior point P lies on one of the sides. Use symmetry to equate lengths of segments.

NOT pictured (alternative proofs) area-decomposition: equate $[\triangle ABC]$ to the sum of three sub-triangle areas $\frac{1}{2}a(PD+PE+PF)$. The figure does *not* draw or shade these sub-triangles.

2.4. Judge

We use `claude-opus-4-6` as the LLM-as-judge. The judge sees the the per-axis rubric, the calibrated rubric entry for each problem, the model's response, and a system prompt that pins human-graded calibration examples. We attach the following examples of system prompt calibrations that were tested against the balanced subset of problems.

- **Calibration A (Cycloid, diagram-only, gpt-5).** Human grade: Axis 2 = 2 (not 3), Axis 3 = 1 (not 2). The model arrives at the correct identity ($3\pi R^2$) through a geometric intermediate step that is wrong: it claims the strips rearrange into a right triangle, but the rearranged region is sigmoid-shaped. The prompt instructs the judge: *do not rescue with an off-picture Cavalieri argument.*

- **Calibration B (Sums of Squares, diagram-only, gpt-5).** Human grade: Axis 2 = 1, Axis 3 = 1. The model misidentifies which panel is doing which work and hand-waves the slice-and-reflect step. Score 1 on Axis 3 is awarded specifically for this kind of right-vibe, wrong-mechanism response.

For our task setting, we implement LLM-as-judge grading in line with the work of HealthBench (Arora et al., 2025), which showed that LLM judges align closely with expert graders when supplied a detailed rubric and reference solution. However, the judge's task must be strictly easier than that of the evaluated model, and the judging task must rely on the rubric specificity rather than the inherent judge model reasoning capability.

We fulfill these requirements in our PWW-Bench implementation: rather than recovering a theorem and proof from an image, the judge receives the canonical theorem T, prose proof P, anchor list A, and not pictured N along with the standard rubric, and need only verify whether the model's response matches them. This task is effectively reading comprehension with an answer key, not open-ended reasoning.

Furthermore, the proof-specific data supplied by the calibration provides the necessary rubric specificity per problem to avoid overt reliance on judge model reasoning. A prominent failure mode we observed in rubric calibration was judge model rewarding fluent proofs that describe the correct identity but fail to identify the depicted visual proof mechanism. We demonstrate that this undesired judge behavior is removed by supplying few-shot human-graded judge examples in the system prompt, with the explicit instruction "do not rescue with an off-picture argument."

In order to develop the rubric protocol, the authors independently graded a 12-item slice (cycloid, sums-of-squares, Viviani, plus 9 others) under the same rubric and discussed

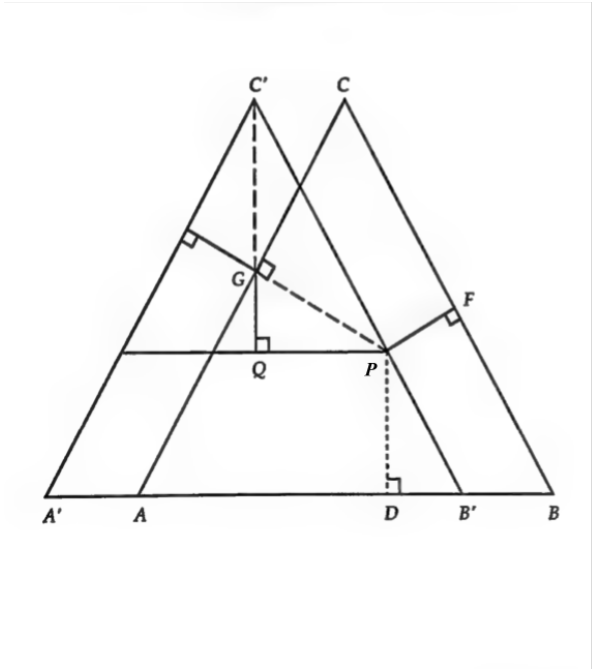

Figure 2. Viviani's theorem (PWW #012). The pictured proof slides the triangle so the interior point P lies on one of the sides. The proof then argues through matching different segment lengths using symmetry. Frontier models default instead to the *area-decomposition* alternative proof—which is correct, but not what the figure shows. This is a clean "alternative proof" failure mode (§3.2).

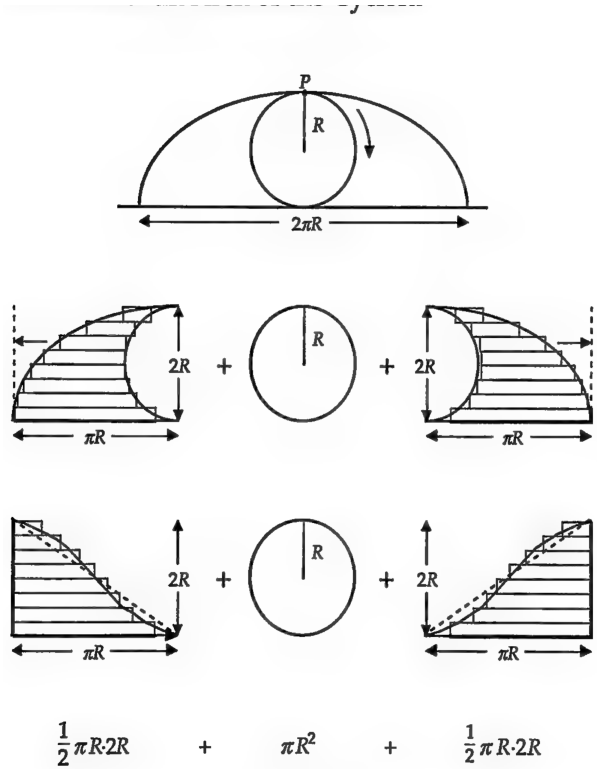

Figure 3. Area under one arch of the cycloid (PWW #039). The figure decomposes the cycloid arch into three regions and visually rearranges them into a final shape with area $3\pi R^2$. We use this as Calibration A: gpt-5 names the right identity but its rearrangement step is geometrically wrong (the rearranged region is sigmoid-shaped, not triangular).

disagreements. Initial disagreements were almost entirely on the Axis 2/Axis 3 boundary ("did this fluent paragraph *actually* use the picture?"). The most debated failure modes were developed into the calibration examples Calibration A and B.

2.5. Algorithm Walk-Through

We work through a concrete example—the cycloid problem (Figure 3)—in §2.5.1.

2.5.1. Concrete Example: Cycloid (#039)

The diagram (Figure 3) shows three panels. On the diagram-only formulation, gpt-5's response (attached in A.1) correctly identifies the theorem ("area under one arch of a cycloid is $3\pi R^2$"), correctly names the three-piece decomposition, and concludes with the right value. The *rearrangement step*, however, claims that two cycloid sub-regions slide into a right triangle of base πR and height $2R$. They do not—the rearranged shape is sigmoid-shaped (one boundary is a half-cycloid arc). The judge, anchored by Calibration A, provides the following evaluation:

- Axis 1 = 3 (theorem named correctly).
- Axis 2 = 2 (panel pieces named, but one piece's role is mis-stated).
- Axis 3 raw = 2, but the prompt explicitly forbids upgrading to 3 because the geometric rearrangement is wrong on the figure.
- After gating: $\text{axis3}_{\text{final}} = \min(2, 2) = 2$.
- Final: 7/9. Tag: `invalid_grounded_chain`.

Under the original (uncalibrated) rubric, the same response scored 9/9.

3. Experimental Evaluation

3.1. Methodology

Models. 8 VLMs: gpt-5 (`reasoning_effort=low`, 16k budget), claude-opus-4-6, claude-sonnet-4-6, llama-4-maverick, qwen3-vl-235b, pixtral-large, qwen2.5-vl-72b, and llama-3.2-11b-vision. All called via each provider's native or OpenRouter chat completions API. Single sample per (model, item, formulation). No temperature sweep.

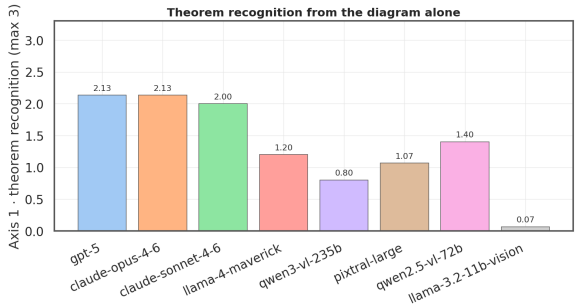

Figure 4. Theorem recognition (Axis 1) on the diagram-only formulation. No model is above 2.2/3. Frontier models hover at 2.0–2.1: "right neighborhood, wrong specific theorem." The open-weight bottom half mostly picks the wrong theorem family.

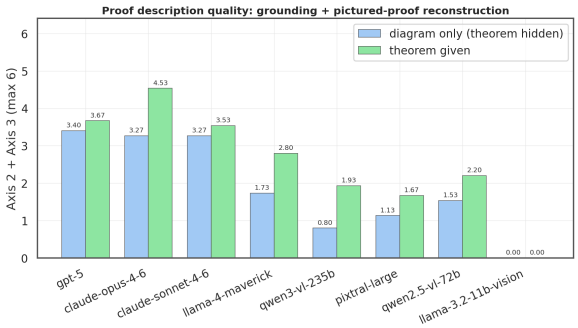

Figure 5. Proof-description quality (Axes 2+3, max 6) for diagram-only vs. theorem-given. The theorem-given formulation removes the retrieval bottleneck. Two distinct patterns appear: Opus and several mid-tier models gain a full point or more (theorem retrieval was their bottleneck); Sonnet and gpt-5 gain almost nothing (their grounding was already on the right theorem; the remaining errors are inside the proof itself).

Independent variables. (i) the formulation (diagram-only vs. theorem-given), (ii) the model.

Dependent variables. per-axis scores, total, failure tags, primary failure mode (the most consequential of {theorem, grounding, proof, refusal, none}).

Hypotheses.

- **H1.** Even when the theorem is given, models cannot reliably reconstruct the pictured proof. The bottleneck is visual grounding: understanding the geometric chain of reasoning encoded in the figure.
- **H2.** The typical frontier failure arrives at the correct identity through a geometric step that does not actually hold in the figure. An LLM-judge baseline will systematically over-reward these cases. However, this can be ameliorated by anchoring a human-developed rubric to specific pictured-proof mechanisms.
- **H3.** Failure modes are model-dependent in a structured way: weaker models fail at theorem identification, while frontier models fail at the geometric step inside the proof. Comparing the two task formulations makes this decomposition visible.

3.2. Results

Table 1 reports per-model means over the 15-problem pilot subset. On the diagram-only formulation, frontier models cluster between 5.27 (Sonnet) and 5.53 (gpt-5). The next tier, llama-4-maverick and qwen2.5-vl-72b, ties at 2.93. The 12B-class llama-3.2-11b-vision is essentially at 0. No model is above 6/9, and the calibrated rubric drops frontier scores by roughly a full point relative to a vibe-graded pass.

Theorem recognition in Diagram-only formulation. Identifying the theorem from the diagram alone is hard for all models (Figure 4). Frontier models reach ∼2/3: they say something in the right mathematical neighborhood, but rarely the exact theorem (e.g., on Viviani's theorem, the closest frontier guesses are "Pythagorean" and "geometric mean"). These results demonstrate that the diagram-only formulation captures critical visual mathematical reasoning failures and hallucinations in VLMs that are obscured by previous visual math benchmark formulations, which only focus on calculation problems.

Visual reasoning in theorem-given formulation. To further probe the ability to reconstruct visual proof steps, we tested models on the theorem given formulation. To quantify the performance difference, we study the delta on the (Axis 2 + Axis 3) sub-score from the diagram-only to the theorem-given formulation (Figure 5). In this analysis, we find that some models (opus +1.26, qwen3-vl-235b +1.13, llama-4-maverick +1.07) improved noticeably, indicating that theorem retrieval was the bottleneck. On the other hand, some models (sonnet +0.26, gpt-5 +0.27) show more limited gains, indicating that proof-reading was the bottleneck.

On the diagram-only formulation, Claude Opus often gets confused about which theorem the diagram is showing, so its proof description is for the wrong target. Once the theorem-given formulation fixes the target, Opus's proof description improves substantially. Claude Sonnet and gpt-5, by contrast, already had the right target in mind on the diagram-only formulation; what they got wrong was the geometric step inside the proof, and that error survives even when the theorem is given. But even with the theorem in hand, the strongest models do not hit ceiling performance on the benchmark, indicating that they still cannot reliably reconstruct the pictured proof.

	Diagram only				Theorem given		
Model	Axis 1	Axis 2	Axis 3	**Total / 9**	Axis 2	Axis 3	**Total / 6**
gpt-5	2.13	1.73	1.67	**5.53**	1.87	1.80	**3.67**
claude-opus-4-6	2.13	1.67	1.60	**5.40**	2.33	2.20	**4.53**
claude-sonnet-4-6	2.00	1.73	1.53	**5.27**	1.80	1.73	**3.53**
llama-4-maverick	1.20	1.07	0.67	**2.93**	1.47	1.33	**2.80**
qwen2.5-vl-72b	1.40	0.87	0.67	**2.93**	1.20	1.00	**2.20**
pixtral-large	1.07	0.67	0.47	**2.20**	0.87	0.80	**1.67**
qwen3-vl-235b	0.80	0.47	0.33	**1.60**	1.07	0.87	**1.93**
llama-3.2-11b-vision	0.07	0.00	0.00	**0.07**	0.00	0.00	**0.00**

Table 1. Per-model means on the 15-problem pilot subset under the calibrated rubric. Diagram-only max is 9 (Axes 1+2+3); theorem-given max is 6 (Axes 2+3). Frontier models cluster at 5.3–5.5 / 9 on the diagram-only formulation; the strongest open-weight tier models tie at 2.93. Opus shows the largest theorem-given gain, indicating its diagram-only bottleneck was theorem retrieval; gpt-5 and sonnet show negligible gain, indicating their residual errors are proof-reading failures (§3.2).

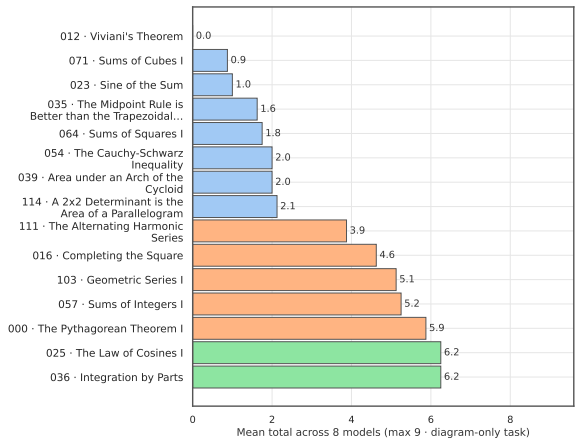

Figure 6. Per-problem difficulty on the diagram-only formulation, mean over 8 models, max 9. The eight hardest problems score under 3/9 across all models. Notably, Viviani's theorem is scored at 0 across all models.

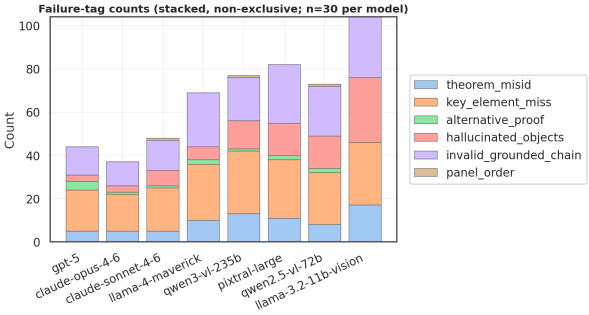

Figure 7. Failure-tag counts (stacked, non-exclusive; $n = 30$ per model). Under the calibrated rubric `invalid_grounded_chain` becomes the dominant frontier failure: 11–14 per model, up from 3–7 under the lenient rubric.

Per-problem difficulty. We find that our balanced subset of the benchmark has two problem difficulty regimes (Figure 6). Eight items score under 3/9 across all 8 models; six score above 5. The "easy" items are the canonical dissection proofs (Pythagorean, Completing the Square, Geometric Series, Parallelogram-as-determinant) where any model with strong math priors recognises the figure as iconography of a known proof. The hard items either require multi-panel temporal reasoning (Sums of Cubes, Sums of Squares) or use a specific construction the model is unlikely to have seen verbatim (Viviani, Cauchy–Schwarz).

Failure modes. After the LLM judge scored each model output, it specified tags for each failure mode it encountered (Figure 7). Under the calibrated rubric, we found that the dominant frontier failure more was `invalid_grounded_chain` (11–14/30)—"the model named the right pieces but the geometric argument is wrong". For open-weight models, `key_element_miss` (24–29/30) and `theorem_misid` (8–17/30) dominate. `hallucinated_objects` spikes for the smaller open models (llama-3.2-11b at 30/30): when they cannot read the figure, they hallucinate. Refusals are rare; the task is hard but not refused.

Visual proof archetype effects. We quantified model performance per problem archetype (Figure 8). Staged-reveal is the hardest archetype for every model. This archetype was dominated by the Viviani's theorem item, which has no canonical alternative the model can fall back on. The dissection archetype is easiest for the frontier because most pilot dissection items *are* canonical iconography (Pythagorean, Completing the Square, Geometric Series).

3.3. Discussion

The theorem-given setting does not fix proof-reading. Sonnet and gpt-5 exhibit minimal gains of $\leq 0.27/6$ when the theorem is provided. The remaining errors are proof-reading failures: the model substitutes an off-figure algebraic argument that looks right but does not follow from

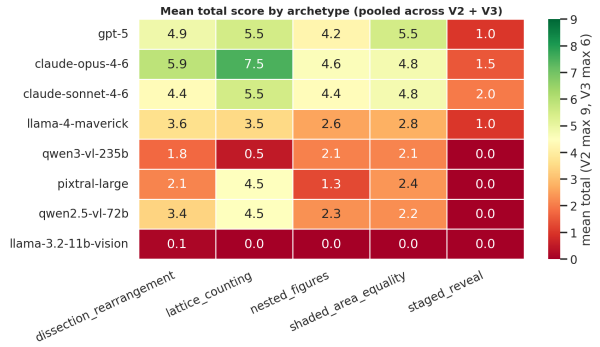

Figure 8. Mean total by archetype, pooled across both formulations. Staged-reveal (Viviani) is hardest for all models; dissection is easiest because the pilot dissection items are canonical (Pythagorean, geometric series).

the picture. Therefore, the visual-reasoning bottleneck on PWW-bench is not primarily about naming the theorem, but also about reasoning through proof steps that are implied by various geometric components of an image.

The calibrated rubric prevents an LLM judge from rewarding correct, non-illustrated proofs. Frontier scores dropped from 6–7 / 9 under a vibe-graded first pass to 5.3–5.5 / 9 under the calibrated rubric. The `invalid_grounded_chain` tag rose from 3–7 to 11–14 frontier hits, exactly the failure mode the calibration was designed to catch. Rank order between models is largely preserved, but the absolute scores now reflect proof-reading correctness rather than fluency.

Failure modes split by model capability tier. Open-weight models fail predominantly at theorem identification (Figure 7: `theorem_misid` 8–17/30, `key_element_miss` 24–29/30). Frontier models pass theorem identification more often but fail inside the proof (`invalid_grounded_chain` 11–14/30). Reading the diagram-only and theorem-given formulations together makes the decomposition visible: a model can have a strong theorem-retrieval prior and still be unable to follow the figure's argument.

PWW Items with Outlier Difficulty Every model scored 0/9 on Viviani's theorem in the diagram-only formulation. The frontier models cycled through the wrong theorem family ("Pythagorean," "Simson line," "orthocenter"). When the theorem is given, opus and gpt-5 still get partial credit only by giving the area-decomposition proof—a correct alternative argument that is not the same as the proof based on sliding line segments that the figure depicts.

4. Related Work

Math benchmarks for language models. GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021) kicked off systematic evaluation of math reasoning in language models with text-only word problems. Their problem statements supply all constraints in prose; the models do not need to read structure off of pictures. PWW deliberately removes the prose problem statement.

Visual math benchmarks. MathVista (Lu et al., 2024) mixes text + figure across 5 task families; figures are largely supplements to a textual problem statement. MathVerse (Zhang et al., 2024) addresses whether or not a model ever uses a figure by ablating caption content, while also keeping the input style of standard figure-with- caption question tasks. MMMU (Yue et al., 2024) contains math sub-tasks, again with prose problem statements. We-Math (Qiao et al., 2024) decomposes problems into sub-skills but is still text-led. MathVision (Wang et al., 2024) focuses on visual content, but its tasks emphasize computing a numerical answer given the figure and a question.

Diagrams as input. Diagram-understanding benchmarks like AI2D (Kembhavi et al., 2016) and FigureQA (Kahou et al., 2018) target chart and schematic QA, not theorem proof. Geometry-specific work (e.g., Geometry3K, (Lu et al., 2021)) tests numeric solving of geometry problems with figure + text; PWW differs in that there is no numeric answer to compute—instead, the model needs to output a proof description.

LLM-as-judge. Zheng et al. (2023) introduced LLM-as-judge for chat-quality evaluation. More recent rubric-grounded benchmarks such as HealthBench (Arora et al., 2025) demonstrate that LLM judges agree well with expert graders when given a detailed rubric and reference solution—the regime we operate in. Our contribution is the use of human-graded counter-examples pinned in the system prompt, which reduced instances of grading a correct-sounding answer with an incorrect or non-illustrated proof.

5. Future Work

Calibrated rubric sample size. Our pilot study of human-calibrated rubrics with proof data covers 15 problems. A scaled run on 100–150 problems will allow for reporting on model performance on archetype means with confidence intervals. To support this effort, we develop a preliminary version of our online interface for crowdsourcing rubric calibration, which is described in Appendix A.2.

Few-shot calibration breadth. The current calibration examples may over-specialise on cycloid- and sums-of-

squares-style failures. We will expand to 8-10 examples spanning each archetype, with the human-grading step done jointly by the two annotators.

Image-output as a follow-up probe. Recent VLMs (Gemini 2.5 Flash Image, Gemini 3 Pro Image, GPT-5 Image) emit images as well as text. A natural inverse task is reconstructing the diagram from the caption: given P and T, generate a candidate PWW image and ask whether it (i) is a valid proof and (ii) matches the original schema. This probes whether the model has a mental picture of the proof, not just verbal knowledge of it.

Cross-judge robustness. We use a single LLM judge (`claude-opus-4-6`). The next step is a cross-judge ablation: regrade the same responses with sonnet and gpt-5 as judges and report inter-judge agreement and rank-correlation across judges, to confirm that the calibrated rubric is what drives the scores rather than any judge-specific quirk.

Hidden-labels ablation. Creating a labels-hidden formulation that blurs all in-figure labels (variable names, "$n = 4$," etc.) would further isolate to what extent the model depends on reading the text inside a picture.

Multi-sample / temperature. Single-sample evaluation understates the ceiling for stochastic models. Multi-sample with majority-vote on Axis 1 and best-of-k on Axis 3 is a reasonable extension.

Reasoning-effort and budget. gpt-5 was run at `reasoning_effort=low` with a 16k token budget for cost. A high-effort sweep on the 8 hardest problems would clarify whether more thinking time improves model ability to read and describe the correct proof depicted in the illustration.

6. Conclusion

We present PWW-Bench, a multimodal mathematical reasoning benchmark that tests the visual understanding of current VLMs. Visual representations are a core part of mathematical research and education. Despite this, current benchmarks only evaluate VLMs on simple, calculation- and symbolic manipulation-based visual tasks. In contrast, PWW-Bench evaluates a VLM's ability to deconstruct a diagram, understand its components, and reason through the relationships between these components. We do a pilot study of 8 VLMs on this benchmark, and categorize various failure modes. Importantly, frontier models often identify what theorem to prove and how to prove it, but produce a proof unrelated to the diagram. This result shows a strong tendency of VLMs to be biased by their language priors rather than true visual understanding. We believe this benchmark will be useful in the design and evaluation of models that can do true multimodal mathematical reasoning.

Impact Statement

This paper establishes informal proof-based reasoning through visual mathematical proofs as a distinct task for multi-modal models and presents a benchmark for this task. We thereby hope to advance the state of multi-modal reasoning both as it pertains to AI for math, as well as broader multi-modal AI applications. The immediate societal and ethical impacts we foresee are improvement in multi-modal models for math education and research.

7. Acknowledgements

We would like to thank the anonymous reviewers for their insightful feedback, which helped improve the quality of this manuscript.

References

Arora, R. K. et al. HealthBench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025.

Cobbe, K., Kosaraju, V., Bavarian, M., et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Guzman, M. d. The role of visualization in the teaching and learning of mathematical analysis. 2002.

Hadamard, J. *An Essay on the Psychology of Invention in the Mathematical Field*. Princeton University Press, Princeton, NJ, 1945.

Hendrycks, D., Burns, C., Kadavath, S., et al. Measuring mathematical problem solving with the MATH dataset. In *NeurIPS Datasets and Benchmarks*, 2021.

Kahou, S. E., Michalski, V., Atkinson, A., et al. FigureQA: An annotated figure dataset for visual reasoning. In *ICLR Workshop*, 2018.

Kembhavi, A., Salvato, M., Kolve, E., et al. A diagram is worth a dozen images. In *ECCV*, 2016.

Lu, P., Gong, R., Jiang, S., et al. Inter-GPS: Interpretable geometry problem solving with formal language and symbolic reasoning. In *ACL*, 2021.

Lu, P., Bansal, H., Xia, T., et al. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *ICLR*, 2024.

Mancosu, P. Visualization in logic and mathematics. In *Visualization, explanation and reasoning styles in mathematics*, pp. 13–30. Springer, 2005.

Nelsen, R. B. *Proofs Without Words: Exercises in Visual Thinking*. Classroom Resource Materials. Mathematical Association of America, 1993.

Nelsen, R. B. *Proofs Without Words II: More Exercises in Visual Thinking*. Classroom Resource Materials. Mathematical Association of America, 2000.

Nelsen, R. B. *Proofs Without Words III: Further Exercises in Visual Thinking*. Classroom Resource Materials. Mathematical Association of America, 2015.

Pólya, G. *How to Solve It: A New Aspect of Mathematical Method*. Princeton University Press, Princeton, NJ, 2nd edition, 1957. ISBN 9780691080970.

Qiao, R., Tan, Q., Dong, G., et al. We-Math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*, 2024.

Wang, K., Pan, J., Shi, W., et al. Measuring multimodal mathematical reasoning with MATH-Vision dataset. *arXiv preprint arXiv:2402.14804*, 2024.

Yue, X., Ni, Y., Zhang, K., et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *CVPR*, 2024.

Zhang, R., Jiang, D., Zhang, Y., et al. MathVerse: Does your multi-modal LLM truly see the diagrams in visual math problems? In *ECCV*, 2024.

Zheng, L., Chiang, W.-L., Sheng, Y., et al. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *NeurIPS Datasets and Benchmarks*, 2023.

Zimmermann, W. and Cunningham, S. Editor's introduction: What is mathematical visualization. *Visualization in teaching and learning mathematics*, 1(8):1–7, 1991.

A. Appendix

A.1. gpt-5 Cycloid Response (PWW #039)

Goal. The figure shows that the area under one arch of the cycloid generated by a circle of radius R equals $3\pi R^2$ (three times the area of the generating circle).

How the picture works.

The arch and its scale. The top panel shows one cycloid arch produced by the rolling circle of radius R. The base of the arch has length $2\pi R$ (the distance the circle's center translates in one turn). P is the tracing point on the rim that generates the arch; the highest point of the arch occurs when P is at the top of the circle.

Split the arch into three panels. Draw the vertical line through the circle's center at the top of the panel; this slices the arch into a left cap and a right cap, each over a base of length πR and of height $2R$. The middle "panel" is reserved for the rolling circle itself (radius R).

Rearranging by horizontal slices (Cavalieri/shear). In the two lower rows each cap is cut into many thin horizontal strips (shown as "steps"). A horizontal strip may be slid left–right without changing area; doing the same slide to every strip preserves the area of the whole cap (Cavalieri's principle).

First shear: making the circle appear (middle row). For each height y ($0 \le y \le 2R$) there is a point of the cycloid reached when the wheel has turned through some angle t. If each strip is shifted horizontally by the same amount as the center of the wheel has translated at that instant (a distance Rt), the cycloid boundary of that strip becomes the corresponding point on the circle of radius R centered at height R. In symbols, the cycloid is

$$x = R(t - \sin t), \qquad y = R(1 - \cos t),$$

and subtracting the translational piece Rt gives

$$(x - Rt,\ y) = (-R\sin t,\ R(1 - \cos t)),$$

which lies on the circle $(x - Rt)^2 + (y - R)^2 = R^2$. After this shear the "wavy" tops of the two caps fit around a circle of radius R in the center panel, so

$$\text{area(arch)} = \text{area(left cap)} + \pi R^2 + \text{area(right cap)}.$$

Second shear: straightening to triangles (bottom row). Now slide the horizontal strips inside each cap by an amount that increases linearly from 0 to πR as the wheel turns from $t = 0$ to $t = \pi$ (i.e., across a half-arch). This "straightens" the dashed boundary in each cap into a straight segment, yielding a right triangle. Each resulting triangle has base πR and height $2R$, as labeled on the panels.

Areas read off the bottom line.

$$\begin{aligned} \text{Left triangle:} &\quad \tfrac{1}{2}(\pi R)(2R) = \pi R^2, \\ \text{Middle circle:} &\quad \pi R^2, \\ \text{Right triangle:} &\quad \tfrac{1}{2}(\pi R)(2R) = \pi R^2. \end{aligned}$$

Adding the three labeled areas gives

$$\text{area under one cycloid arch} = \pi R^2 + \pi R^2 + \pi R^2 = 3\pi R^2.$$

Thus the "proof without words" shows—by cutting into horizontal strips and performing area-preserving shears—that the cycloid arch's area equals three times the area of its generating circle.

Figure 9. **GPT-5's generated response to Proof Without Words #039 (the cycloid arch).** Asked to verbalize a wordless visual proof, the model reconstructs the argument that one cycloid arch encloses an area of $3\pi R^2$: it identifies the two area-preserving shears in the figure and reads off the result as two right triangles of area πR^2 flanking the generating circle.

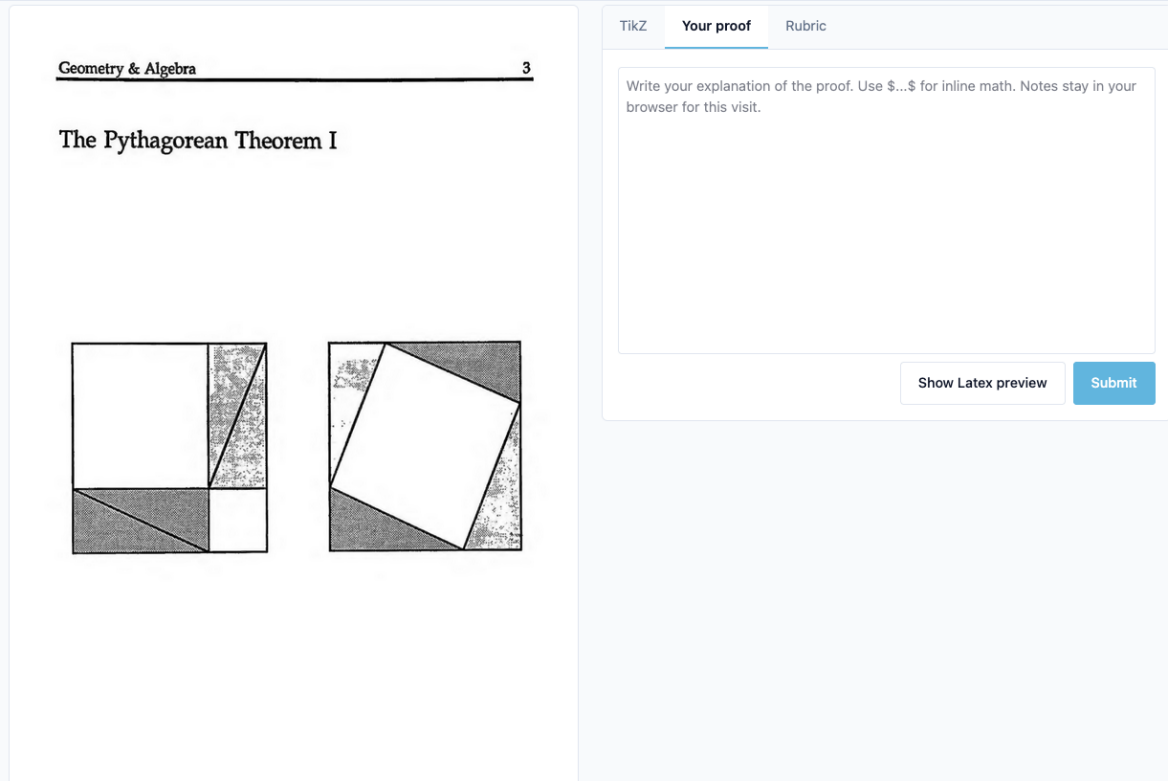

Figure 10. Submission form for a proof P of a PWW-Item.

A.2. Rubric Calibration Online Interface

In order to facilitate the creation of calibrated rubrics across a scaled up set of PWW items, we develop an online interface for crowdsourcing human feedback on problems. After navigating through a menu that organizes all problems by category, the user is presented with a single PWW-Item and provided with fields for submitting a prose proof P (Fig. 10); a proposed rubric that contains information on anchor list A, not-pictured N, and other proof-specific details (Fig. 11); and a tikz sketch of the PWW item for creation of higher-quality images (Fig. 12).

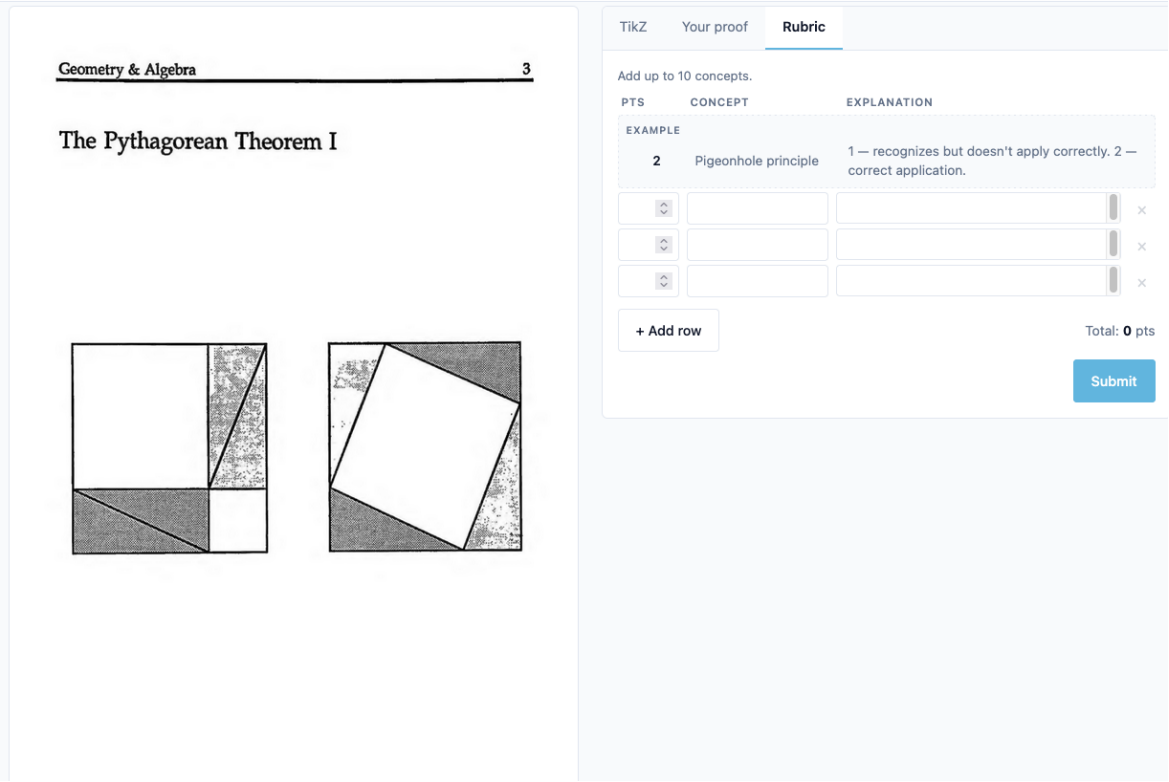

Figure 11. Submission form for a rubric specifying anchor list A, not-pictured N, and other proof-specific details.

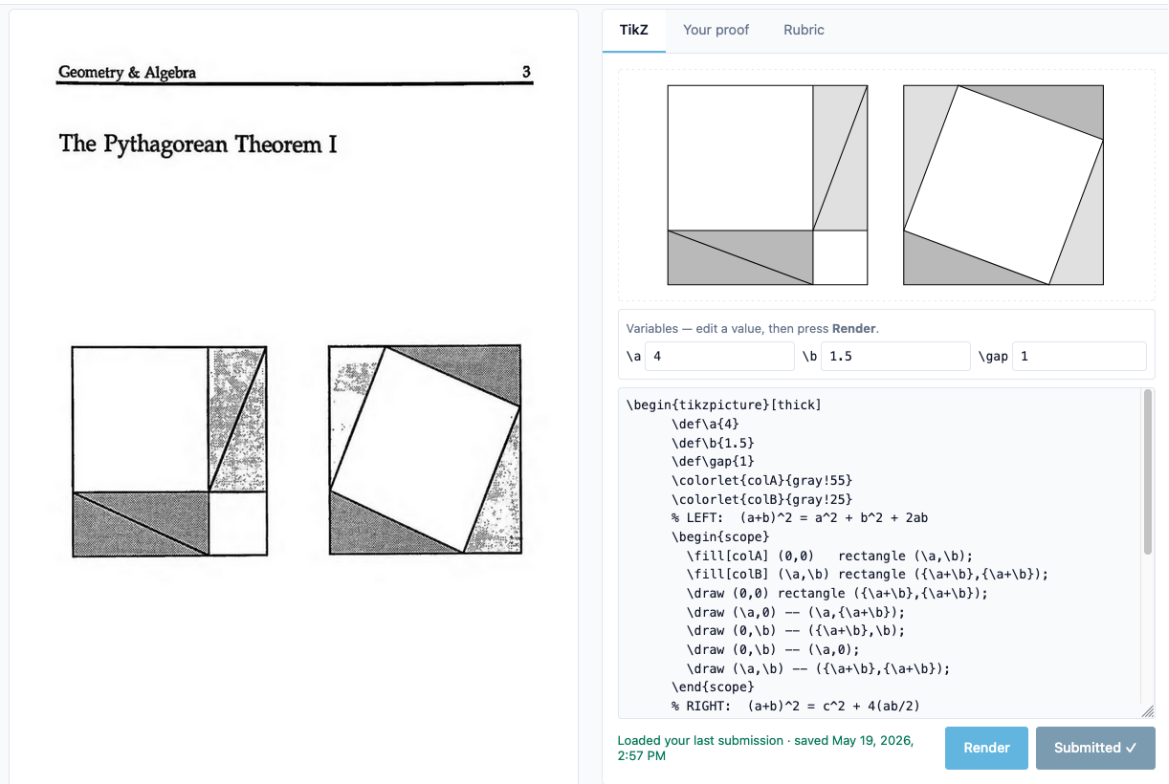

Figure 12. Submission form for a tikz sketch of the PWW item with built-in viewer.