CAN LLMs KEEP UP WHEN MATH BECOMES PHYSICS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Can a language model solve a physics problem as reliably as the mathematics underlying it? We introduce Math2Physics, a benchmark of 216 mathematics–physics pairs from six collections, linked by shared central derivations. An expert-guided conversion procedure expresses mathematical constraints through physical laws, observables, and measurements, producing paired tasks that probe the recovery of mathematical structure from physical descriptions. Expert-reviewed annotations characterize physical domain, authenticity, and the added interpretation steps. Across three models, eight reasoning settings, and 64 attempts per question, mathematics accuracy is 77.03% and physics accuracy is 71.58% on the common 212-pair panel: a 5.45 percentage-point gap (95% paired-bootstrap interval $[3.77, 7.14]$). The disadvantage recurs on the same problems across models, and greater reasoning effort does not consistently close it. Yet oracle coverage reaches 93.40% for physics at 64 attempts, approaching mathematics at 94.00%. This contrast reveals a gap between finding a correct physics solution among many attempts and producing one reliably in a single attempt. We further use the paired response banks to compare mathematics and physics preference training on held-out physics problems. Math2Physics provides a common foundation for studying scientific reasoning across formulations, reasoning settings, and training subjects.

1 INTRODUCTION

An equation can be stated directly in a mathematics problem or recovered from a physical law. The subsequent derivation may be shared, but the second formulation asks the solver to identify a system, interpret its observables, and translate its behavior into mathematical constraints. How reliably do language models make this transition? Competition benchmarks such as MATH measure mathematical reasoning (Hendrycks et al., 2021), while physics evaluations reveal failures in theorem application and condition analysis (Zhang et al., 2025). Comparing scores on unrelated mathematics and physics questions mixes these interpretation demands with differences in the underlying mathematics. A paired benchmark makes the connection explicit.

We introduce **Math2Physics**, which links each physics question to a mathematics problem through a shared central derivation. Its focus complements subject-specific evaluations such as OlympiadBench (He et al., 2024) and UGPhysics (Xu et al., 2025), as well as studies of alternative representations in IsoBench (Fu et al., 2024) and narrative grounding in ContextMATH (Cao et al., 2026). The organizing principle is a correspondence between mathematical constraints and a physical system: laws and measurements supply relations that the mathematics version presents directly. This allows us to compare success within problem pairs and examine how the gap varies with both the model and the physical formulation.

Math2Physics comprises 216 pairs from six mathematics collections. Thirty expert-constructed pairs from the Korean College Scholastic Ability Test (CSAT) and its mock examinations (Korea Institute for Curriculum and Evaluation, n.d.) guide a conversion procedure that produces 186 further pairs from five competition collections, followed by expert review. Physical laws, observables, and measurements connect each counterpart to its source's central derivation. In Figure 1, for example, optical reflection and travel times recover the constraints of an ellipse problem. The dataset includes expert-reviewed annotations of physical domain, authenticity, and the number of added physical interpretation steps, together with model-relative difficulty estimates. Source-linked construction records describe the mathematical correspondence and any changes to the requested quantity, making the relationship within each pair available for inspection (Section 3).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

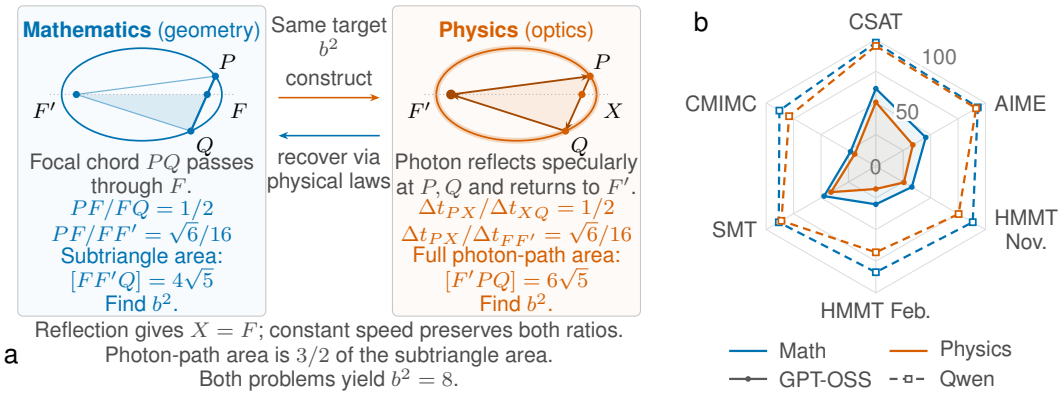


Figure 1: From mathematics to physics. a Physical laws recover the ellipse geometry and target b^2 from an optical cavity. Here $\Delta t_{FF'}$ is the direct travel time between the foci (condensed statements; full pair in Figure 8). b Accuracy by dataset for two settings: GPT-OSS-20B (GPT-OSS; OpenAI, 2025) at low reasoning effort has the largest overall math–physics gap (8.33 percentage points); Qwen3.6-35B-A3B (Qwen3.6; Qwen Team, 2026b) with thinking has higher accuracy and a 6.95-point gap. Both use flexible scoring on 212 matched pairs with 64 attempts per question. Radii span 0–100%.

Across three models and eight reasoning settings, we measure accuracy from 64 attempts per question. The common comparison contains 212 pairs with matching evaluations of the final formulations. Weighting models equally, mathematics accuracy is 77.03% and physics accuracy is 71.58%, a gap of 5.45 percentage points (pp), with a 95% paired-bootstrap interval of $[3.77, 7.14]$ (Section 4). The direction persists under stricter answer scoring and when selecting each model's highest-effort setting. Moreover, the penalty partly recurs on the same problems across models, indicating a shared pattern beyond their different overall accuracies. Higher-effort settings can improve performance without consistently narrowing the mathematics–physics gap.

Repeated sampling reveals a second aspect of this reliability gap (Section 5). Oracle coverage, which counts a problem as solved when any sampled answer is correct, reaches 94.00% for mathematics and 93.40% for physics at 64 attempts. Many physics problems therefore admit correct model-generated solutions even when individual attempts remain less reliable. The paired response banks also enable a direct comparison of training subjects: we apply preference optimization (Rafailov et al., 2023) to Qwen3.5-4B using separate mathematics and physics data, then evaluate on 60 held-out physics problems. Together, these analyses connect formulation sensitivity to opportunities for answer selection and preference learning.

Our contributions are:

- **A benchmark built around paired derivations.** We release 216 expert-reviewed mathematics–physics pairs, a reusable conversion procedure, physical-content annotations, and model response banks.
- **A systematic study of formulation-sensitive reliability.** We quantify a 5.45 pp average physics disadvantage, identify recurring problem-level gaps across models, and examine reasoning settings and physical characteristics.
- **A link from evaluation to training.** We show that repeated sampling substantially narrows the oracle-coverage gap and use the same paired structure to compare mathematics and physics preference training.

2 Related Work

Mathematics and physics benchmarks. OlympiadBench combines mathematics and physics examination problems (He et al., 2024), while SciBench covers college-level mathematics, chemistry, and physics (Wang et al., 2024). UGPhysics evaluates undergraduate physics across multiple subjects (Xu et al., 2025), and PhysReason assesses both final answers and intermediate solution steps

(Zhang et al., 2025). PRISM-Physics represents solutions as dependency graphs of formulas for process evaluation (Zhao et al., 2026). Math2Physics complements these evaluations by linking each physics question to its mathematical source, enabling comparisons within pairs that share a central derivation.

Paired tasks and problem generation. Paired representations also support comparisons across modalities and contexts. IsoBench compares visual and textual representations of the same tasks, including physics problems (Fu et al., 2024). ContextMATH rewrites mathematics problems through Scenario Grounding and Complexity Scaling; the latter turns explicit conditions into subproblems that the solver must resolve (Cao et al., 2026). Math2Physics focuses on recovering mathematical structure from physical laws, observables, and measurements, with expert-reviewed annotations of physical content. Each source–counterpart link supports a direct comparison of the reliability with which a model solves the two formulations.

Recent work on isomorphic physics generation constructs variants of existing physics problems with changes to context or numerical parameters (Liu et al., 2026; Wu & Ismaili, 2026). These preprints provide precedents for automated generation and verification within physics. Math2Physics starts from mathematics and constructs physical counterparts through expert examples, generation, and critique. This direction of construction makes the mathematical source available as a reference for both solution checking and paired evaluation.

Transfer across task domains. Human-learning research has examined transfer between isomorphic algebra and physics tasks, including arithmetic progressions and constant acceleration (Bassok & Holyoak, 1989). For LLMs, Huan et al. (2026) study transfer from mathematics post-training to other task domains. Our training comparison asks whether mathematics or physics preference data better supports held-out physics performance under direct preference optimization (Rafailov et al., 2023), matching preference-pair counts within each update regime. The paired sources make it possible to hold the underlying problem families fixed while varying the subject of the preference data.

3 Turning mathematics into physics

Math2Physics tests whether models can recover the same underlying mathematical structure when it is expressed through a physical system. Its 216 pairs (432 questions) link a mathematics problem to a physics counterpart through a shared central derivation. Physical observables, laws, and measurements determine how that derivation is reached and how its result is interpreted. We develop the construction through expert-written examples, expand it with a reusable LLM skill, and review the resulting pairs with physics experts (Figure 2).

3.1 Expert construction and mathematical correspondence

The seed collection contains 30 questions from CSAT and its associated June and September mock examinations: 15 administered in 2012–2019 and 15 in June or September 2026. This selection combines established examination material with recent questions. The Korean mathematics questions were translated before a physics expert constructed their physics counterparts. Translation retains the mathematical conditions; physical reformulation specifies a system from which the solver must recover the corresponding relations.

Three operations recur across the pairs: expressing a mathematical quantity as a physical observable, recovering an explicit mathematical relation from a physical law, and realizing a discrete sample space through measurement. Optical reflection, for example, can recover focal geometry, while a specified spin preparation can preserve the ordering and probabilities of die outcomes. A single question can combine several such operations.

The volume-to-inertia example illustrates how a physical law exposes the shared derivation. The mathematics problem asks for the volume of a solid of revolution. Its counterpart asks for the moment of inertia of a body with the same shape and radial density $\rho(s) = K/s^2$, where $K > 0$ is fixed and s is distance from the rotation axis. The material lies away from that axis. Applying the

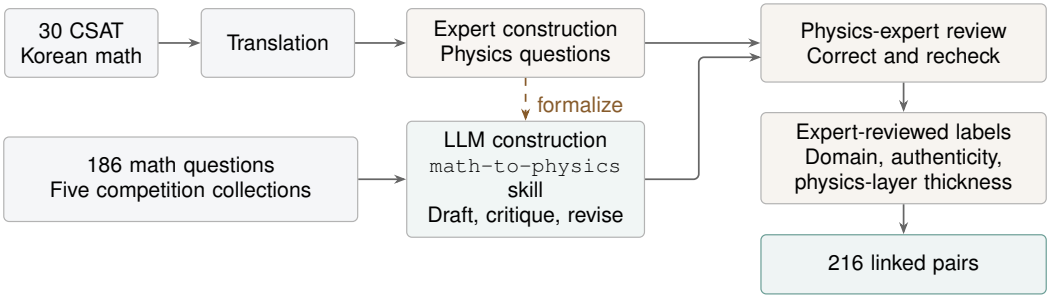


Figure 2: Construction and review of Math2Physics. Expert-built CSAT pairs inform the conversion skill used for five additional collections. LLM critique supports generation; physics experts review the resulting questions separately before annotations are finalized.

inertia law gives

$$I = \int s^2 \rho(s)\, dV = K \int dV. \tag{1}$$

The density cancels the radial weighting, leaving the source volume integral. The solver must identify the physical observable and assemble its governing integral before carrying out the shared mathematical calculation. Appendix A.3 gives the complete pair and further examples involving optics, circuits, and signal timing.

For each pair, the correspondence records the retained constraints, solution dependencies, and relation between the requested quantities. Full task equivalence additionally requires a reversible map between admissible configurations, preservation of probability weights when relevant, and an invertible relation between targets. We distinguish this stronger criterion from a shared central derivation and record changes to targets or source interpretations. Appendix A.2 formalizes the criterion, and Appendix A.1 documents source-specific qualifications.

3.2 Skill-guided expansion and expert review

We encode the expert procedure as the `math-to-physics` skill, which makes the mathematical correspondence a design constraint throughout generation. The skill first solves the source and extracts its constraints, target, and essential solution steps. It then proposes three candidate physical models, compares their suitability, and establishes an explicit correspondence before drafting the physics statement. Each construction includes a derivation of the physical answer with assumptions and units; nonstandard physical laws are supplied in the question.

The resulting statement undergoes mathematical, physical, and adversarial critique. Mathematical review checks the source constraints and solution dependencies. A physics solve from the question alone checks the stated model, while adversarial review probes alternative interpretations, boundary cases, and shortcuts. Revisions are followed by renewed checks of the affected conditions. This cycle connects generation to verification of the proposed mathematical and physical reasoning.

Physics experts then review and correct the generated questions for physical validity, clarity, and fidelity to the intended mathematics. The resulting dataset retains source identifiers, paired statements, reference answers, units, and correspondence arguments.

MathArena collects competition problems for evaluating mathematical reasoning (Balunović et al., 2025) and has expanded into a maintained evaluation platform (Dekoninck et al., 2026). We apply the skill to five of its collections: AIME 2026 (30 pairs; Mathematical Association of America, n.d.), HMMT Nov. 2025 (30; HMMT, 2025), HMMT Feb. 2026 (33; HMMT, 2026), SMT 2025 (53; Stanford Math Tournament, 2025), and CMIMC 2025 (40; Carnegie Mellon Informatics and Mathematics Competition, 2025), contributing 186 pairs in total. The abbreviated collection names in the figures refer to these editions. Appendix A details collection composition and source fidelity; generation settings appear in Appendix A.2.1.

3.3 PHYSICS ANNOTATIONS

Each physics problem receives labels for domain, authenticity, and physics-layer thickness. Domain identifies the principal topic: Classical Mechanics, Electromagnetism, Quantum Mechanics, or Thermodynamics/Statistical Mechanics. Authenticity distinguishes a physical wrapper around a puzzle (*Limited*), a coherent but engineered construction (*Moderate*), and a natural physical process or measurement task (*Good*). Idealization alone does not lower authenticity.

Thickness counts effective physical reinterpretations or added operations relative to the mathematics source: one is *Superficial*, two are *Normal*, and three or more are *Thick*. Adding units alone does not count. A direct physical mapping and its constitutive readout form one layer; operations already required by the source do not add layers. Authenticity and thickness are ordered categories with no assumed equal spacing. Additional layers need not make a construction more authentic.

Eight Astra runs each label all 216 physics problems. We combine equally weighted votes on each axis and adjudicate ambiguous decisions. Two physics experts review all annotations. Appendix A.4 specifies the rubric and adjudication conventions, model settings, and review criteria. Agreement and subset stability measure consistency among the model runs; they do not estimate annotation accuracy. Their results appear in Appendix B.10.

3.4 MODEL-RELATIVE DIFFICULTY

We estimate difficulty from success rates measured over 64 attempts per question under a fixed reference panel. The score characterizes the evaluated models' reliability, rather than intrinsic complexity or human difficulty. To label all 432 records with matching inputs, the reference panel uses Qwen3.6 with thinking disabled and enabled, and Qwen3.8-27B (Qwen3.8; Qwen Team, 2026c) at low, medium, and xhigh effort. Each model receives half the weight, divided equally among its settings: the two Qwen3.6 conditions have weight $1/4$ each and the three Qwen3.8 conditions have weight $1/6$ each.

For record i and condition m, let c_{im} be the number of correct answers in 64 attempts under flexible scoring (Section 4.1). Independent $\mathrm{Beta}(1/2, 1/2)$ priors with binomial likelihoods give the posterior mean panel success and difficulty

$$\widetilde{q}_i = \sum_m w_m \frac{c_{im} + 1/2}{65}, \qquad \widetilde{d}_i = 1 - \widetilde{q}_i. \tag{2}$$

We label a record *Hard* when $\widetilde{q}_i < 0.50$, *Medium* when $0.50 \le \widetilde{q}_i < 0.80$, and *Easy* when $\widetilde{q}_i \ge 0.80$. The same thresholds apply across collections and problem forms. Continuous scores, credible intervals, and category probabilities retain information lost by thresholding.

Mathematics and physics records are scored separately; a shared mathematical core does not impose equal difficulty. Because the same response bank also supports performance analysis, associations between difficulty and the mathematics–physics gap are descriptive. Appendix A.5 gives the posterior, interval assumptions, and diagnostic protocol; Appendix B.11 reports predictive checks and sensitivity to the reference panel, scoring rule, and thresholds.

4 THE MATH–PHYSICS PERFORMANCE GAP

4.1 EXPERIMENTAL SETUP

We use the paired benchmark to examine three questions: how large is the mathematics–physics gap, how does it respond to reasoning settings, and which problems exhibit it across models? We evaluate 216 pairs from six collections, with 64 attempts per question in eight conditions: GPT-OSS-20B at low, medium, and high effort; Qwen3.6-35B-A3B with thinking disabled or enabled; and Qwen3.8-27B with thinking at low, medium, and xhigh effort. Cross-model comparisons use the same 212 pairs with matching evaluations of the final formulations; the complete five-condition Qwen panel supplies the difficulty labels. Accuracy is the mean fraction of correct answers over the 64 attempts for each question. Our primary flexible scoring accepts an omitted unit when the numerical value is correct in the reference unit, while checking any supplied units. Strict scoring requires explicit

units for numerical physical quantities. Appendix B.2 specifies input coverage; Appendix B.1 gives grading and sampling settings.

The gap G is mathematics minus physics accuracy, in percentage points (pp). Each pair receives equal weight; each model receives one third of the panel weight, shared equally among its modes. Pointwise 95% confidence intervals resample whole pairs within collections, keeping both forms and all settings together. Paired tests use one observation per pair; Holm correction covers 23 performance contrasts (Appendix B.2).

4.2 Physics reduces reliability on matched problems

Physics accuracy is lower by 5.45 pp on the matched panel: 77.03% for mathematics versus 71.58% for physics, with a gap interval of $[3.77, 7.14]$ and Holm-adjusted $p = 6.64 \times 10^{-8}$. Mathematics has higher panel success on 162 pairs, lower success on 44, and equal success on six. The pattern therefore extends across individual problems as well as the aggregate score. All six collections have positive average gaps, ranging from 2.61 pp on SMT 2025 to 9.74 pp on HMMT Feb. 2026; three pass the joint correction (Appendix B.3).

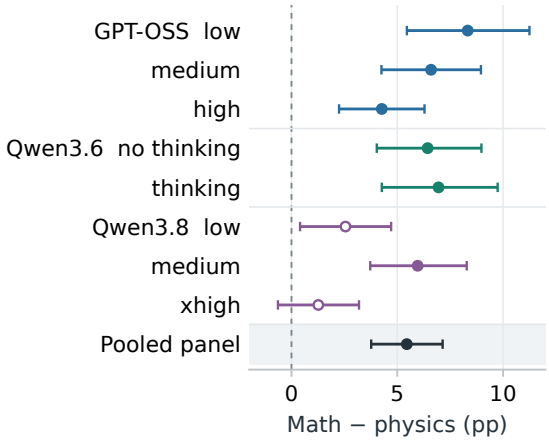


Figure 3: Paired accuracy gaps on 212 problems. Bars are pointwise 95% pair-bootstrap intervals; filled markers pass the joint 23-contrast Holm correction. Positive values favor mathematics.

Model	Mode	Math	Physics	**Gap**
GPT-OSS	low	40.15	31.82	**8.33**
	medium	70.19	63.59	**6.60**
	high	80.54	76.27	**4.27**
Qwen3.6	no thinking	70.98	64.54	**6.43**
	thinking	89.54	82.59	**6.95**
Qwen3.8	low	82.16	79.61	**2.55**
	medium	85.43	79.47	**5.96**
	xhigh	94.04	92.77	**1.27**
Pooled		77.03	71.58	**5.45**

Table 1: Accuracy (%) averaged over 64 attempts per question, and the math-minus-physics gap (pp). The pooled row weights models equally.

All eight conditions have positive aggregate gap estimates, and six pass the joint correction (Figure 3). The strongest condition, Qwen3.8 xhigh, reaches 94.04% mathematics and 92.77% physics accuracy; its 1.27 pp gap has interval $[-0.64, 3.19]$. The gap is smaller at this high level of performance. Across the panel, it persists under strict scoring (9.84 pp), the highest-effort mode of each model (4.16 pp), and the complete Qwen panel (4.90 pp; Appendix B.9). In particular, accepting omitted units already leaves a substantial gap: the physics disadvantage extends beyond unit-format compliance. Collection-level reversals coexist with the aggregate gap: Qwen3.8 xhigh has slightly higher physics accuracy on AIME 2026 and HMMT Feb. 2026 (Figure 4).

4.3 Thinking improves accuracy without consistently closing the gap

Enabling Qwen3.6 thinking raises mathematics and physics accuracy by 18.57 and 18.05 pp, respectively, leaving a gap of 6.95 pp. The change in gap is $+0.52$ pp $[-2.27, 3.28]$. Higher effort also has no uniform relationship with the gap (Figure 5). Qwen3.8 low-to-medium widens it by 3.41 pp $[1.79, 5.11]$: mathematics improves while physics accuracy is nearly unchanged. Medium-to-xhigh then narrows it by 4.69 pp $[2.36, 7.04]$. Both changes pass the joint correction ($p_{\mathrm{H}} = 0.0017$ and 0.0032). GPT-OSS low-to-high narrows the observed gap by 4.06 pp $[1.22, 6.98]$, with adjusted $p = 0.068$.

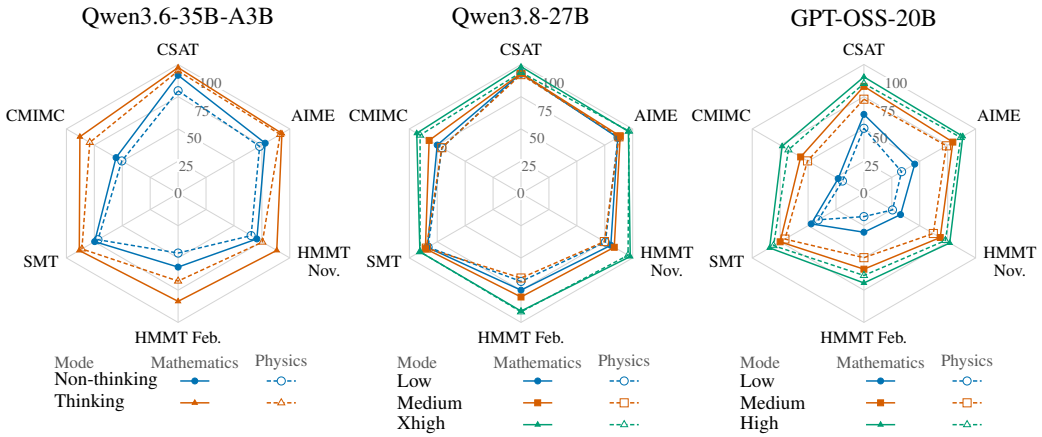


Figure 4: Model accuracy by collection on 212 pairs, with 64 attempts per question and flexible scoring. Color and marker shape identify modes. Solid lines and filled markers show mathematics; dashed lines and open markers show physics. Radial scales run from 0 to 100%.

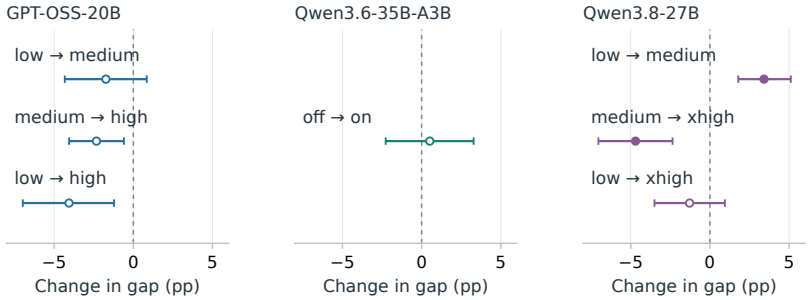


Figure 5: Change in the paired gap within each model. Negative values mean physics improves more than mathematics. Bars are pointwise 95% paired intervals; filled markers pass the joint Holm correction. Off/on refers to Qwen3.6 thinking.

Overall, these comparisons show that improvements in absolute accuracy need not translate into a smaller formulation gap. We compare each model's configured reasoning modes; their decoding parameters are given in Appendix B.1. Section 5 examines whether additional independently sampled attempts recover physics solutions that individual attempts miss.

4.4 Where the gap appears

Recurring problems across models. The paired design lets us ask whether different models struggle with the same physical reformulations. After averaging reasoning modes within each model, pair-level gaps have positive Spearman correlations of 0.286–0.488. All three model comparisons remain significant under collection-adjusted permutation tests with Holm correction ($p_{\mathrm{H}} = 6.0 \times 10^{-5}$; Appendix B.5). Of the 212 pairs, 82 favor mathematics under all three model averages, whereas 17 favor physics under all three. Thus, models with different overall capabilities share part of the same problem-level reliability pattern. This recurrence motivates examining the mathematical difficulty and physical characteristics of the affected pairs.

Problem difficulty. The largest observed penalty occurs on problems labeled Easy in mathematics: 7.16 pp, compared with 3.36 pp for Medium and −3.34 pp for Hard (Figure 6a). On the Easy subset, mathematics accuracy is 90.05% and physics accuracy is 82.89%; the disadvantage is therefore present even where the mathematical task is solved reliably. These groups describe the reference panel's behavior. Diagnostics using separate response sets are reported in Appendix B.7.

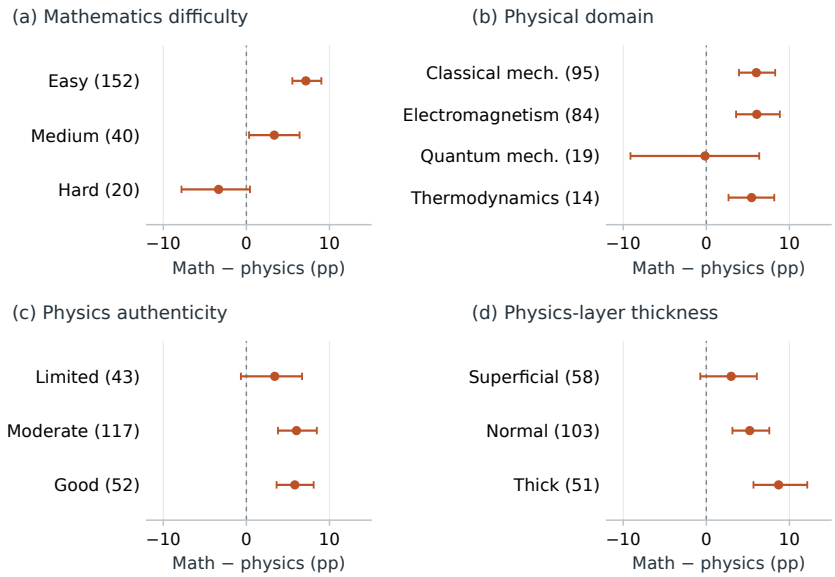


Figure 6: Paired gaps by mathematics difficulty and expert-reviewed physics labels, using the same 212 pairs and model weights. Parentheses give pair counts; bars are pointwise 95% pair-bootstrap intervals. Difficulty uses the Qwen reference panel. Thermodynamics includes Statistical Mechanics. Subgroup patterns are descriptive; joint tests and response-separation diagnostics appear in Appendix B.6 and Appendix B.7.

Domain and authenticity. Mean gaps are 6.02 pp in Classical Mechanics, 6.07 pp in Electromagnetism, and 5.45 pp in Thermodynamics/Statistical Mechanics. Quantum Mechanics has a near-zero estimate, -0.16 pp, but only 19 pairs and a wide interval $[-9.13, 6.37]$. The gap also remains positive for Good-authenticity problems (5.83 pp), as well as Moderate (6.03 pp) and Limited (3.41 pp). The disadvantage is therefore also present in constructions judged to represent natural physical processes or measurements.

Physics-layer thickness. Mean gaps rise from 2.99 pp for Superficial to 5.22 pp for Normal and 8.70 pp for Thick problems. Each model has a larger mean gap for Thick than Superficial problems. This descriptive gradient identifies physical interpretation depth as a candidate for more targeted investigation; the joint adjusted tests do not establish differences along any of the three annotation axes (Appendix B.6). The clearest common pattern is the recurrence across models: physical reformulation exposes reliability differences on identifiable problem pairs, including many whose mathematics is solved accurately.

5 Test-time scaling and reinforcement learning from physics

The paired accuracy gap raises a natural question: how often can models produce a correct physics solution when given additional attempts? We first examine oracle coverage under repeated sampling, then use the paired response banks for an initial comparison of mathematics and physics preference training.

5.1 Physics capability under repeated sampling

Oracle pass@k measures the probability that at least one of k sampled answers is correct. It summarizes the availability of successful solutions across attempts. We estimate it within each model condition and then apply the same model weights used for accuracy. On the 212-pair evaluation panel, physics pass@64 reaches 93.40%, compared with 94.00% for mathematics (Figure 7). The mathematics–physics gap falls from 5.45 percentage points (pp) at pass@1 to 0.60 pp at pass@64. Physics gains 21.81 pp in coverage, compared with 16.97 pp for mathematics; the paired difference is 4.85 pp, with a 95% confidence interval of $[3.59, 6.10]$. At intermediate budgets, the gap is 2.65 pp

at four attempts and 1.30 pp at 16. The steady reduction shows that much of the single-attempt disadvantage concerns how reliably a correct physics solution is sampled. This creates an opportunity for methods that identify successful responses or increase their probability through training.

5.2 Learning from physics

The paired formulations allow us to ask whether learning from physics solutions transfers more effectively to new physics problems than learning from solutions to their mathematical counterparts. We apply offline direct preference optimization (DPO; Rafailov et al., 2023) to the post-trained Qwen3.5-4B model (Qwen Team, 2026a), using separate preference sets drawn from a shared pool of 60 source families, ten from each collection. Each preference favors a response with a correct final answer over an incorrect response to the same question. We compare the two training subjects under low-rank adaptation (LoRA; Hu et al., 2022) and full parameter updates. Within each update regime, the two conditions share the initialization, number of preference pairs, and training hyperparameters. Full updates use 4,632 pairs per subject at a 2,048-token cutoff, while LoRA uses 5,033 at 8,192 tokens. Thinking responses are compared only with thinking responses, and non-thinking responses with non-thinking responses; long thinking traces contribute prefixes labeled by their original outcomes. All runs train for one epoch. Evaluation uses 60 held-out physics problems, ten from each collection, with eight attempts per problem under the same thinking prompt and decoding policy. We report accuracy, which requires correct answers with explicit units for numerical physical quantities. Appendix C provides the experimental details and paired accuracy analysis.

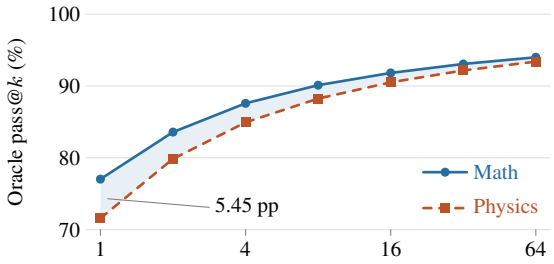


Figure 7: Oracle coverage on 212 mathematics–physics pairs under flexible scoring, with equal model weights (Equation 9). Shading marks the gap; the horizontal axis shows the number of attempts k on a logarithmic scale.

DPO data	LoRA	Full
Physics (P)	**58.33**	**57.92**
Mathematics (M)	57.08	57.08
P − M (pp)	+1.25	+0.83
95% CI (pp)	$[-3.54, 6.04]$	$[-3.54, 5.21]$

Table 2: Physics accuracy (%) after DPO: eight attempts per held-out problem ($n = 60$). Pointwise 95% paired-bootstrap CIs follow Appendix C.4.

Physics DPO yields higher observed physics accuracy than mathematics DPO in both update regimes (Table 2): the advantage is 1.25 pp with LoRA and 0.83 pp with full parameter updates. Both paired 95% confidence intervals include zero, leaving the relative benefit of the two training subjects unresolved. The experiment demonstrates how the benchmark's paired construction extends to a training comparison: mathematics and physics preference sets are drawn from a shared family pool, with equal pair counts within each update regime.

6 Conclusion: Strengths, Limitations, & Future Directions

Math2Physics turns mathematical derivations into paired tests of scientific reasoning. Its 216 expert-reviewed pairs connect mathematical sources to physical laws and observables around a shared mathematical core. On the common 212-pair panel across three models and eight settings, physics reduces single-attempt accuracy by 5.45 percentage points. Problem-level gaps partly recur across models, and greater reasoning effort does not consistently close them.

With repeated sampling, physics oracle coverage reaches 93.40% at 64 attempts, near mathematics at 94.00%. The gap between coverage and single-attempt accuracy motivates work on answer selection and post-training. These response banks support preference-training comparisons. The conversion protocol extends this framework to other mathematics collections; Lean and Physlib could further clarify the physical dependencies of each derivation.

AI USE STATEMENT

We used generative AI tools to draft and polish manuscript text and to identify relevant literature. GPT-6 Astra (OpenAI, 2026) generated 186 physics counterparts from mathematics problems and assisted with checking mathematical correspondence and physical validity (Section 3.2). It also supplied initial physics annotations and adjudicated ambiguous labels (Section 3.3). Physics experts reviewed and corrected the generated questions and annotations. The authors take responsibility for the final manuscript, including its claims, references, data, and code.

REPRODUCIBILITY STATEMENT

Appendix A documents source collections, mathematical correspondence, generation checks, annotation, and difficulty estimation. Benchmark sampling and grading appear in Appendix B.1; Appendix B specifies statistical procedures, resampling settings, and sensitivity analyses. Appendix C details preference construction, prefix handling, optimization, held-out evaluation, and paired uncertainty estimates. The preference-set configuration and family-selection records identify the training source snapshot, preprocessing revisions, and evaluation exclusions.

REFERENCES

Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. MathArena: Evaluating LLMs on uncontaminated math competitions. In *Advances in Neural Information Processing Systems*, volume 38, pp. 22851–22888, 2025. doi: 10.52202/085713-0679. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/1d27c01ebd3e3aebe226b44fc970d803-Abstract-Datasets_and_Benchmarks_Track.html.

Miriam Bassok and Keith J. Holyoak. Interdomain transfer between isomorphic topics in algebra and physics. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(1): 153–166, 1989. doi: 10.1037/0278-7393.15.1.153.

Bowen Cao, Dongdong Zhang, Yixia Li, Junpeng Liu, Shijue Huang, Chufan Shi, Hongyuan Lu, Yaokang Wu, Guanhua Chen, Wai Lam, and Furu Wei. From abstract to contextual: What LLMs still cannot do in mathematics. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/2a6fc6774932e7835349089b0287ed2c-Abstract-Conference.html.

Carnegie Mellon Informatics and Mathematics Competition. 2025 math competition. Official competition problems and solutions, 2025. URL https://cmimc.math.cmu.edu/math/past-problems/2025.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374, 2021. URL https://arxiv.org/abs/2107.03374.

Jasper Dekoninck, Nikola Jovanović, Tim Gehrunger, Kári Rögnvaldsson, Ivo Petrov, Chenhao Sun, and Martin Vechev. Beyond benchmarks: MathArena as an evaluation platform for mathematics with LLMs. arXiv preprint arXiv:2605.00674, 2026. URL https://arxiv.org/abs/2605.00674.

Deqing Fu, Ruohao Guo, Ghazal Khalighinejad, Ollie Liu, Bhuwan Dhingra, Dani Yogatama, Robin Jia, and Willie Neiswanger. IsoBench: Benchmarking multimodal foundation models on isomorphic representations. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=KZd1EErRJ1.

Samuel W. Greenhouse and Seymour Geisser. On methods in the analysis of profile data. *Psychometrika*, 24(2):95–112, 1959. doi: 10.1007/BF02289823. URL https://link.springer.com/article/10.1007/BF02289823.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3828–3850, 2024. doi: 10.18653/v1/2024.acl-long.211. URL https://aclanthology.org/2024.acl-long.211/.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html.

HMMT. Archive of November 2025. Official tournament problems and solutions, 2025. URL https://www.hmmt.org/www/archive/291.

HMMT. Archive of February 2026. Official tournament problems and solutions, 2026. URL https://www.hmmt.org/www/archive/292.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://iclr.cc/virtual/2022/poster/6319.

Maggie Ziyu Huan, Yuetai Li, Tuney Zheng, Xiaoyu Xu, Seungone Kim, Minxin Du, Radha Poovendran, Graham Neubig, and Xiang Yue. Does math reasoning improve general LLM capabilities? understanding transferability of LLM reasoning. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/65125.

Han Jiang, Sunbeom Kwon, Jinwen Luo, Ziang Xiao, and Susu Zhang. Can we trust item response theory for AI evaluation? arXiv preprint arXiv:2607.15190, 2026. URL https://arxiv.org/abs/2607.15190.

Korea Institute for Curriculum and Evaluation. College scholastic ability test (CSAT) and mock examination materials. Official examination materials, n.d. URL https://www.suneung.re.kr/boardCnts/list.do?boardID=1500234&m=0403&s=suneung. Mathematics examinations from multiple years. Mock examinations: https://www.suneung.re.kr/boardCnts/list.do?boardID=1500236&m=0403&s=suneung. Accessed September 24, 2026.

Klaus Krippendorff. Computing Krippendorff's alpha-reliability, 2011. URL https://www.asc.upenn.edu/sites/default/files/2021-03/Computing%20Krippendorff%27s%20Alpha-Reliability.pdf. Annenberg School for Communication, University of Pennsylvania; references updated 2013.

John P. Lalor, Hao Wu, and Hong Yu. Learning latent parameters without human response patterns: Item response theory with artificial crowds. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference*

on Natural Language Processing (EMNLP-IJCNLP), pp. 4249–4259. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1434. URL https://aclanthology.org/D19-1434/.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/aca97732e30bcf1303bc22ac3924fd16-Abstract-Conference.html.

Naiming Liu, Leo Murch, Charles Alliotts, Spencer Moore, Tong Wan, Shashank Sonkar, Richard Baraniuk, and Zhongzhou Chen. Scalable generation and validation of isomorphic physics problems with GenAI. arXiv preprint arXiv:2602.05114, 2026. URL https://arxiv.org/abs/2602.05114. Preprint.

Mathematical Association of America. MAA invitational competitions. Official competition website, n.d. URL https://maa.org/maa-invitational-competitions/. American Invitational Mathematics Examination (AIME). Accessed September 24, 2026.

OpenAI. gpt-oss-120b & gpt-oss-20b model card. Model card, August 2025. URL https://openai.com/index/gpt-oss-model-card/.

OpenAI. GPT-6 Astra: A new generation of intelligence. Model release announcement, 2026. URL https://openai.com/index/gpt-6-astra/.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5. Qwen3.5-4B model card: https://huggingface.co/Qwen/Qwen3.5-4B.

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all, April 2026b. URL https://qwen.ai/blog?id=qwen3.6-35b-a3b. Model card: https://huggingface.co/Qwen/Qwen3.6-35B-A3B.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026c. URL https://qwen.ai/blog?id=qwen3.8. Qwen3.8-27B model card: https://huggingface.co/Qwen/Qwen3.8-27B.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741, 2023. doi: 10.52202/075280-2338. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Stanford Math Tournament. SMT 2025. Official tournament problems and solutions, 2025. URL https://www.stanfordmathtournament.org/past-tests/SMT/2025.

The mathlib Community. The Lean mathematical library. In *Proceedings of the 9th ACM SIGPLAN International Conference on Certified Programs and Proofs (CPP 2020)*, pp. 367–381. ACM, 2020. doi: 10.1145/3372885.3373824. URL https://popl20.sigplan.org/details/CPP-2020-papers/14/The-Lean-mathematical-library.

The Physlib community. Physlib: The Lean physics library. GitHub repository, 2024. URL https://github.com/leanprover-community/physlib.

Joseph Tooby-Smith. HepLean: Digitalising high energy physics. *Computer Physics Communications*, 308:109457, 2025. doi: 10.1016/j.cpc.2024.109457. URL https://www.sciencedirect.com/science/article/pii/S0010465524003801.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R. Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. SciBench: Evaluating college-level scientific problem-solving abilities of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 50622–50649. PMLR, 2024. URL https://proceedings.mlr.press/v235/wang24z.html.

Xian Wu and Lindim Ismaili. Reliable isomorphic physics problem generation with large language models. arXiv preprint arXiv:2607.13190, 2026. URL https://arxiv.org/abs/2607.13190. Preprint.

Xin Xu, Qiyun Xu, Tong Xiao, Tianhao Chen, Yuchen Yan, Jiaxin Zhang, Shizhe Diao, Can Yang, and Yang Wang. UGPhysics: A comprehensive benchmark for undergraduate physics reasoning with large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 69849–69877. PMLR, 2025. URL https://proceedings.mlr.press/v267/xu25ai.html.

Xinyu Zhang, Yuxuan Dong, Yanrui Wu, Jiaxing Huang, Chengyou Jia, Basura Fernando, Mike Zheng Shou, Lingling Zhang, and Jun Liu. PhysReason: A comprehensive benchmark towards physics-based reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16593–16615, 2025. doi: 10.18653/v1/2025.acl-long.811. URL https://aclanthology.org/2025.acl-long.811/.

Wanjia Zhao, Qinwei Ma, Jingzhe Shi, Shirley Wu, Jiaqi Han, Yijia Xiao, Si-Yuan Chen, Xiao Luo, Ludwig Schmidt, and James Zou. PRISM-Physics: Causal DAG-based process evaluation for physics reasoning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/1a379b51d7ec80b32d8c14f40c281ce1-Abstract-Conference.html.

A Dataset methods

This appendix supplies the correspondence criteria and worked examples underlying Section 3. It also specifies the annotation conventions, agreement measures, and difficulty diagnostics used in the empirical analysis.

A.1 Collections and source fidelity

Table 3 expands the collection summary in Section 3.2 with label counts and source-release links. The CSAT seed contains 15 questions administered in 2012–2019, eight in June 2026, and seven in September 2026. Dates denote calendar years of administration, not Korean academic-year designations. The five MathArena collections use the supplied transcriptions and selections, which need not cover every competition round.

Source checking includes diagram conditions and documented corrections to transcribed statements or answer keys. Three cases require a specified source interpretation. AIME 2026 question 13 uses its explicit residue-class description despite inconsistent summation indices and an omitted zero in the opening range. CMIMC 2025 question 24 places a point strictly inside a side in the physics version; the mathematics wording does not explicitly exclude endpoints. CMIMC 2025 question 27 requires the full supporting line in physics, whereas the finite segment stated in the mathematics version admits no configuration satisfying the other conditions. For these pairs, the correspondence applies to the stated interpretation rather than the unmodified source wording.

A.2 Correspondence and verification

For strict task equivalence (Section 3.1), let $\mathcal{M}$ and $\mathcal{P}$ be the admissible mathematical and physical configurations. A map $T : \mathcal{M} \to \mathcal{P}$ must be reversible up to physically irrelevant freedoms and translate constraints in both directions. Requested quantities q_M, q_P and a declared physical scale U must satisfy

$$\frac{q_P(T(z))}{U} = h(q_M(z)), \qquad z \in \mathcal{M}, \tag{3}$$

where h is invertible on the relevant range and derived from the physical model. Counting requires preservation of distinguishability and multiplicity; probability also requires measure preservation; optimization requires ordering and attainment. We check solution dependencies separately: supplying a decisive intermediate result can preserve an answer while removing the intended reasoning.

Table 3: Dataset composition and label counts. Each pair contains one mathematics problem and one physics counterpart. Difficulty is reported separately for the two forms using the reference panel in Section 3.4; the other labels describe the physics counterparts after expert review. Links identify the source mathematics collections.

(a) Difficulty

		Mathematics			Physics		
Source dataset	Pairs	Hard	Medium	Easy	Hard	Medium	Easy
CSAT	30	1	2	27	1	4	25
AIME 2026	30	2	3	25	2	5	23
HMMT Nov. 2025	30	3	6	21	3	11	16
HMMT Feb. 2026	33	5	10	18	8	12	13
SMT 2025	53	5	7	41	5	7	41
CMIMC 2025	40	5	13	22	11	10	19
Total	216	21	41	154	30	49	137

(b) Physics annotations

	Domain				Authenticity			Layer thickness		
Source dataset	CM	EM	QM	TSM	L	M	G	S	N	T
CSAT	13	9	6	2	5	13	12	8	11	11
AIME 2026	14	15	0	1	5	20	5	2	20	8
HMMT Nov. 2025	13	15	1	1	6	16	8	10	15	5
HMMT Feb. 2026	16	13	3	1	5	19	9	10	13	10
SMT 2025	25	20	2	6	10	30	13	16	26	11
CMIMC 2025	16	14	7	3	14	21	5	15	18	7
Total	97	86	19	14	45	119	52	61	103	52

CM: Classical Mechanics; EM: Electromagnetism; QM: Quantum Mechanics; TSM: Thermodynamics/Statistical Mechanics. Authenticity: L = Limited, M = Moderate, G = Good. Layer thickness: S = Superficial, N = Normal, T = Thick.

When only an essential derivation has been established, we describe a shared-core correspondence and record any changes to the domain, constraints, or requested functional. For example, the CSAT repeated-sampling construction replaces outcomes $\{0, 1, 2\}$ by $\{-1, 0, 1\}$ and asks for a centered rather than an uncentered second moment. The circuit example below preserves the root-counting problem while reparameterizing the requested sum. The three annotation axes characterize physical content; they are not certificates of task equivalence.

A.2.1 Generation and checking protocol

We use the `math-to-physics` skill with GPT-6 Astra at `xhigh` effort. For each source, the model proposes three physical models at an introductory undergraduate level and supplies any non-standard laws. Conversions target strict task equivalence and exact answers in physical quantities, allowing stated normalization or unit conversion. Revisions have no fixed cycle cap. The source solution and its constraint map guide mathematical review; question-only physics solves and adversarial checks test the proposed statement. Relevant checks include exact arithmetic, symbolic substitution, finite enumeration, dimensional analysis, and tests of alternative interpretations or boundary cases. Numerical sampling alone does not prove global uniqueness or equivalence.

Reviewers use separate contexts where available; some contexts are reused for disjoint batches. They share model infrastructure, so separation does not establish statistically independent errors. Repairs are followed by checks of the affected conditions and review of the final question. These model-assisted checks precede the human review described in Section 3.2.

Conversion records contain source identifiers, paired statements, answers, units, correspondence arguments, and executed checks. The generation interface did not expose the exact backend model revision, sampling parameters, seed, or tokenizer/chat template. The records document the procedure but cannot reproduce the wording exactly.

A.3 WORKED EXAMPLES

Figures 8–12 expand the examples in Section 3.1 into five CSAT pairs and their derivations. Statements are condensed, use harmonized notation, and were checked against their Korean counterparts; examination identifiers give the sources. Blue and green panels show mathematics and physics, respectively.

Geometry recovered through optical reflection

Mathematics

Consider the ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

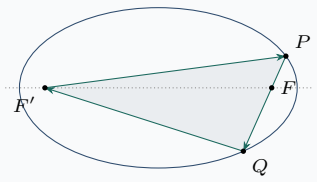
with foci $F(c, 0)$ and $F'(-c, 0)$, where $c > 0$.

Let P be a point on the ellipse in the first quadrant and Q be a point on the ellipse in the fourth quadrant. Suppose that F lies on the line segment $\overline{PQ}$ and

$$\frac{\overline{PF}}{\overline{QF}} = \frac{1}{2}, \qquad \frac{\overline{PF}}{\overline{FF'}} = \frac{\sqrt{6}}{16}.$$

If the area of triangle $FF'Q$ is $4\sqrt{5}$, find the value of b^2.

(Assume that a and b are positive.)



Common geometry and the reflected ray path.

Physics

An ideal two-dimensional elliptical optical cavity has semi-major axis a, semi-minor axis b, and foci F and F'. The distance between the two foci is $2c$, where $c > 0$.

A single photon is emitted from a point source located at F'. If the photon returns to F', it is absorbed there and its trajectory terminates. Before being absorbed, the photon undergoes two successive specular reflections from the cavity wall, first at a point P in the first quadrant and then at a point Q in the fourth quadrant.

The photon travels at a constant speed inside the cavity. Between the two reflections, its trajectory crosses the major axis. The travel time from P to this crossing point is one-half of the travel time from the crossing point to Q. The former travel time is also $\frac{\sqrt{6}}{16}$ of the time required for light to travel directly between the two foci.

The region enclosed by the complete photon trajectory from emission to absorption has area $6\sqrt{5}$. Determine b^2.

Preserved structure. Reflection gives the path $F' \to P \to F \to Q \to F'$; constant speed turns the two time ratios into the source distance ratios. Writing $[XYZ]$ for triangle area,

$$\frac{[FF'P]}{[FF'Q]} = \frac{PF}{FQ} = \frac{1}{2}, \qquad [F'PQ] = \frac{3}{2}[FF'Q] = 6\sqrt{5}.$$

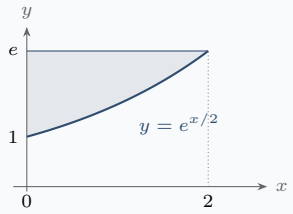
Thus the area conversion recovers the source constraint. To check the answer, set $u = PF = \sqrt{6}\,c/8$. The focal-distance sums and Stewart's theorem give $a^2 - c^2 = 4au/3$, hence $a^2 = 3c^2/2$ and $b^2 = c^2/2$. Then $[FF'Q] = \sqrt{5}\,c^2/4$ yields $c^2 = 16$ and $b^2 = 8$. Lengths and areas use a common reference unit and its square.

Figure 8: Reconstruction through physical laws: focal geometry to an optical trajectory (June 2026, geometry question 28). The focal crossing is inferred from reflection, and an area conversion recovers the source constraint. The inset shows the correspondence between the geometric and optical paths.

A volume integral becomes moment of inertia

Mathematics

Find the volume of the solid obtained by revolving about the x-axis the region bounded by the curve $y = e^{x/2}$, the y-axis, and the line $y = e$.



Region revolved about the x-axis.

Physics

An axisymmetric body has the x-axis as its axis of rotation. The body occupies the region $0 \le x \le 2\,\mathrm{m}$.

At each position x, the outer radius of the body is constant and given by $R = e\,\mathrm{m}$, while the interior region within a radius $r(x) = e^{x/(2\,\mathrm{m})}\,\mathrm{m}$ from the axis of rotation is empty.

The mass density of the body depends on the distance s from the axis of rotation and is given by

$$\rho(s) = \frac{1\,\mathrm{kg/m}}{s^2}.$$

Find the *moment of inertia* I of the body about the x-axis. Express your answer in units of $\mathrm{kg\,m^2}$.

Preserved structure. With $\xi = x/(1\,\mathrm{m})$, the common integral is

$$V_{\mathrm{math}} = \pi \int_0^2 (e^2 - e^{\xi})\, d\xi = \pi(e^2 + 1), \qquad \frac{I}{1\,\mathrm{kg\,m^2}} = V_{\mathrm{math}}.$$

Physical inference. The moment of inertia is $I = \int s^2 \rho(s)\, dV$. The specified density cancels its radial weighting. This engineered choice preserves the volume integral; it does not imply a naturally occurring mass distribution. The material lies at $s \ge 1\,\mathrm{m}$, so the density is finite throughout the body.

Figure 9: Observable substitution: volume to moment of inertia (September 2015, question 14). The integration region and numerical answer are preserved; the solver must obtain the integral from a mass distribution. The inset illustrates the source region.

A root-counting problem recovered from electrical measurements

Mathematics

Let $a > 0$ and let b be a positive integer. For $0 \le x \le 2$, the equation

$$\left(\cos(b\pi x) - \frac{1}{2}\right) \times \left(a\cos(b\pi x) + \frac{a+2}{2}\right) = 0$$

has exactly 15 distinct real solutions. Find $a + b$.

Physics

A circuit's current and voltage each combine a sinusoidal AC component with a constant DC component. Their common angular frequency is $\omega = n\pi\,\mathrm{rad/s}$, where n is a positive integer. Their AC amplitudes are I_0, V_0, and their time averages and initial values satisfy

$$\langle I \rangle = -I_0/2, \quad I(0) = I_0/2,$$
$$\langle V \rangle = \lambda V_0, \quad V(0) = (1+\lambda)V_0,$$

where $\lambda > 0$. On $0 \le t \le 2\,\mathrm{s}$, the instantaneous power $P(t) = I(t)V(t)$ vanishes at exactly 15 distinct instants. Find $n + \lambda$.

Physical inference. The averages fix the DC offsets, and the initial values place both AC components at their maxima. Therefore

$$I(t) = I_0\left(\cos(\omega t) - \frac{1}{2}\right), \qquad V(t) = V_0\big(\cos(\omega t) + \lambda\big).$$

Preserved structure. With $x = t/(1\,\mathrm{s})$, $n = b$, and $\lambda = \frac{1}{2} + \frac{1}{a}$, the power equation recovers the two source factors up to a nonzero scale. This parameter map covers $\lambda > 1/2$. The current contributes $2n$ zeros over n complete cycles; the voltage's disjoint zero set contributes 0, n, or $2n$ zeros according as $\lambda > 1$, $\lambda = 1$, or $0 < \lambda < 1$. Thus 15 zeros force $n = 5$ and $\lambda = 1$, within the shared parameter range. Mathematics asks for $a + b = 2 + 5 = 7$; physics asks for $n + \lambda = 5 + 1 = 6$.

Figure 10: Measured offsets and initial values recover a trigonometric root-counting problem (June 2026, question 14). The physics question supplies the power law, but the solver must reconstruct its factors. The parameter correspondence also explains the different reported answers.

Conic constraints reconstructed from signal arrival times

Mathematics

The hyperbola

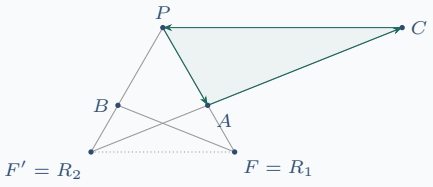
$$\frac{x^2}{16} - \frac{y^2}{b^2} = 1$$

has foci $F(c, 0)$ and $F'(-c, 0)$, with $b, c > 0$. For $P = (0, p)$, $p > 0$, its intersections with the segments PF and PF' are A and B.

An ellipse with foci A, B passes through F, P. The horizontal line through P meets the line AF' at C. If

$$\angle APF' = \angle APC,$$

find the perimeter $L = AC + CP + PA$.



Common geometry and the final relay path.

Physics

Fixed reference transceivers R_1, R_2 have a relay P on their perpendicular bisector, off the line R_1R_2. Devices A, B lie on the segments PR_1, PR_2. All signals travel at speed c, and all relays retransmit without delay.

Simultaneous reference pulses from R_1, R_2 reach A from R_2 a time $t_0 > 0$ later than from R_1; at B, the pulse from R_1 arrives t_0 later than the pulse from R_2.

Two simultaneous signals from B arrive together at A along $B \to R_1 \to A$ and $B \to P \to A$.

A plane mirror at P has its normal along PA. A laser from R_2 reflects specularly at P parallel to R_1R_2, meeting the *line* AR_2 at C.

After removing the mirror, relays return a pulse along $A \to C \to P \to A$ in time t_1. Find t_1/t_0.

Preserved structure. Travel-time differences recover the hyperbola, and equal arrival times recover the ellipse:

$$AR_2 - AR_1 = BR_1 - BR_2 = ct_0, \qquad BR_1 + R_1A = BP + PA.$$

Reflection with normal PA gives $\angle APR_2 = \angle APC$. Identifying $R_1 = F$, $R_2 = F'$ and measuring lengths in units $\ell_0 = ct_0/8$ recovers the source's focal-distance difference 8, both conic constraints, and its angle condition. Solving gives $(AC, CP, PA)/(ct_0) = (35/12, 10/3, 5/4)$, so

$$\frac{t_1}{t_0} = \frac{AC + CP + PA}{ct_0} = \frac{L}{8} = \frac{60}{8} = \frac{15}{2}.$$

The changed answer is a prescribed conversion from perimeter to a dimensionless time ratio. The nondegenerate placement of P, corresponding to the source's $p > 0$, is explicit here; at the midpoint of R_1R_2, the reflected ray would coincide with the line defining C. The inset shows the return path after mirror removal, when P acts as a relay.

Figure 11: Timing and reflection encode a hyperbola, an ellipse, and an angle constraint (September 2026, geometry question 29). Recovering the source geometry requires combining several physical observations; the final measurement reports its perimeter on a different scale.

A.4 Physics-annotation protocol

Section 3.3 defines the three annotation axes and summarizes the eight-run protocol. Here we specify the decision conventions and pre-adjudication agreement measures. Appendix B.10 reports those measures and the expert-reviewed labels.

A.4.1 Annotation conventions

Domain follows the specified physical mechanism rather than an inferred one. Authenticity concerns the physical task, so idealization alone does not lower its rating. Thickness counts a direct physical mapping and its constitutive readout as one layer; distinct preparations, constraints, or added calculations can count separately. Repeated components, unit annotations, and operations already required by the mathematics source do not add layers. The ordinal agreement measure below respects category order without assuming equal spacing.

Circumcircle geometry encoded by motion through planetary tunnels

Mathematics

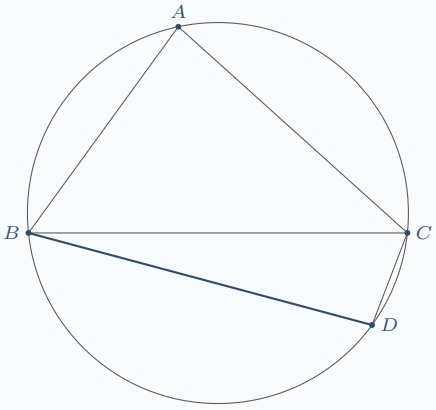
A triangle ABC satisfies

$$AB = 4, \qquad AC = 5, \qquad \cos(\angle BAC) = \frac{1}{8}.$$

A point D lies on the arc BC of the circumcircle of $\triangle ABC$ that does not contain A. If

$$\sin(\angle BCD) = \frac{\sqrt{14}}{4},$$

find BD.



The cyclic geometry of the source problem.

Physics

Points A, B, C, D lie on the equatorial circumference of a uniform-density spherical planet. Straight evacuated tunnels connect AB, AC, BC, BD, and CD.

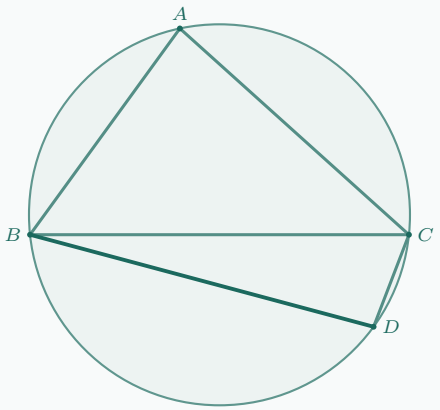
Objects released from rest in tunnels AB and AC reach

$$v_{\max}(AB) = 4\,\mathrm{km/s}, \quad v_{\max}(AC) = 5\,\mathrm{km/s}.$$

If

$$\cos(\angle BAC) = \frac{1}{8}, \qquad \sin(\angle BCD) = \frac{\sqrt{14}}{4},$$

find $v_{\max}(BD)$.



The same chords are interpreted as tunnels through the planet.

Preserved structure. The four surface points retain the cyclic geometry of the source problem, including

$$\cos\angle BAC = \frac{1}{8}, \qquad \sin\angle BCD = \frac{\sqrt{14}}{4}.$$

Thus the original circumcircle derivation still determines BD.

Physical inference. Inside a uniform-density sphere, the gravitational acceleration is linear in displacement from the center. Its component along any straight chord therefore produces simple harmonic motion with the same angular frequency ω. For a tunnel of length L, an object released from a surface endpoint has amplitude $L/2$, so

$$v_{\max} = \frac{\omega L}{2}.$$

Hence

$$\frac{v_{\max}(BD)}{v_{\max}(AB)} = \frac{BD}{AB}.$$

The preserved geometry gives $BD = 4\sqrt{2}$ and $AB = 4$. Therefore, since $v_{\max}(AB) = 4\,\mathrm{km/s}$,

$$v_{\max}(BD) = 4\sqrt{2}\,\mathrm{km/s}.$$

Figure 12: Circumcircle geometry is reframed as motion through straight tunnels inside a uniform-density spherical planet (September 2026, geometry question 12). The cyclic geometry is preserved, while the interior gravitational field converts chord lengths into maximum speeds.

A.4.2 Consensus and adjudication

Eight GPT-6 Astra runs at `xhigh` effort label all 216 physics problems in separate conversations using the same protocol. Exact model snapshots, decoding parameters, and generation seeds are unavailable.

Votes receive equal weight and are combined separately for each axis. An initial label is its unique mode, when one exists. Targeted AI review covers every decision with at most four votes for its leading category and every non-unanimous domain assignment. The reviewing assistant examines the paired mathematics and physics problems, votes, and labeling rationales. This adjudication is not blind and is not an additional sampled vote. Other decisions retain their strict-majority labels before expert review.

Two physics experts subsequently review the annotations for all 216 problems. Their final labels determine the dataset annotation strata. The agreement statistics below concern the eight model runs before adjudication and can reflect shared model biases.

A.4.3 Agreement measures

For item i, axis a, and category c, let n_{iac} denote the number of runs assigning c, and let $p_{iac} = n_{iac}/8$. Normalized entropy measures item-level disagreement:

$$H_{ia} = -\frac{\sum_c p_{iac} \log p_{iac}}{\log K_a}, \qquad 0 \log 0 = 0, \tag{4}$$

where K_a is the number of categories. Exact pairwise agreement averages over the 28 unordered run pairs and $N = 216$ problems:

$$A_a = \frac{1}{N} \sum_{i=1}^{N} \frac{\sum_c n_{iac}(n_{iac} - 1)}{8 \cdot 7}. \tag{5}$$

We also report Krippendorff's $\alpha = 1 - D_o/D_e$, where D_o and D_e are observed and expected disagreement (Krippendorff, 2011). Nominal disagreement is one for unequal categories and zero otherwise. For authenticity and thickness, ordinal disagreement uses pooled category frequencies $N_c = \sum_i n_{iac}$:

$$\delta^2(c, d) = \left(\sum_{g=c}^{d} N_g - \frac{N_c + N_d}{2} \right)^2 \quad (c < d), \qquad \delta^2(c, c) = 0. \tag{6}$$

This respects category order without treating the levels as equally spaced.

Stratified problem-bootstrap intervals describe sensitivity to item composition, conditional on the eight runs. They assume exchangeability within collections and exclude variation across new annotation panels.

Annotation intervals use 5,000 bootstrap samples of whole problems, stratified by collection, keeping all eight annotations of each problem together. The annotation bootstrap uses NumPy PCG64 with seed 20260921.

A.4.4 Subset-stability diagnostic

To measure sensitivity to panel size, we enumerate all subsets of k runs for $k = 1, \ldots, 8$, giving each subset equal weight. For items with a unique eight-run mode, each item–subset pair either recovers that mode, yields a tie, or selects a different unique mode. The full-panel reference includes the sampled runs, so agreement at $k = 8$ is one by construction. This measures consistency within the observed panel, rather than held-out annotation accuracy.

A.5 Difficulty estimation and diagnostic protocol

Section 3.4 defines the reference panel and difficulty labels. Here we give the full posterior and diagnostic protocol for that five-condition Qwen panel of 432 records. These model-relative estimates need not match human judgments (Lalor et al., 2019). The expanded diagnostic panel adds the three GPT-OSS conditions, gives each model equal total weight, and uses the 424 records with matching inputs across all eight conditions. Flexible correctness defines the reference scores; strict correctness is a sensitivity check.

A.5.1 SUCCESS PROBABILITIES AND SAMPLING PROFILES

Let c_{im} be the number of correct answers among $N = 64$ attempts for item i and condition m, with success probability p_{im} and fixed panel weight w_m. Define panel success and difficulty as

$$q_i = \sum_m w_m p_{im}, \qquad d_i = 1 - q_i, \qquad \widehat{q}_i = \sum_m w_m \frac{c_{im}}{64}. \tag{7}$$

Dataset success rates average $\widehat{q}_i$ equally over items, separately for mathematics and physics. For item-level estimates, independent $\mathrm{Beta}(1/2, 1/2)$ priors and binomial likelihoods give

$$p_{im} \mid c_{im} \sim \mathrm{Beta}\big(c_{im} + \tfrac{1}{2},\, 64 - c_{im} + \tfrac{1}{2}\big), \qquad \widetilde{q}_i = \mathbb{E}[q_i \mid c] = \sum_m w_m \frac{c_{im} + 1/2}{65}. \tag{8}$$

This recovers the posterior mean in Equation 2; the full posterior also supplies credible intervals and category probabilities. Difficulty estimates use 20,000 posterior draws for 95% credible intervals and category probabilities. Another 20,000 bootstrap replicates resample whole pairs within collections, retaining both forms and all conditions.

Mean success alone does not distinguish occasional errors from problems that remain unsolved across repeated attempts. We therefore also report the unbiased pass@k estimator of Chen et al. (2021, Section 2.1 and Appendix A):

$$\widehat{S}_D(k) = \frac{1}{|D|} \sum_{i \in D} \sum_m w_m \left[1 - \frac{\binom{64 - c_{im}}{k}}{\binom{64}{k}} \right], \qquad 1 \le k \le 64, \tag{9}$$

with a zero numerator when $64 - c_{im} < k$. All k attempts come from the same condition before averaging over the panel. Pass@k measures whether at least one answer is correct, assuming an oracle that recognizes it; it does not measure an answer selector's accuracy. Pass@64 records observed coverage at this budget, without extrapolating to further attempts.

Item-level posterior intervals condition on the response bank. Pair-bootstrap intervals measure sensitivity to problem composition while retaining both forms and all conditions. Both hold the reference panel fixed.

A.5.2 DIFFICULTY CATEGORIES

The thresholds in Section 3.4 give the following label function:

$$\mathrm{label}(i) = \begin{cases} \texttt{Hard}, & \widetilde{q}_i < 0.50, \\ \texttt{Medium}, & 0.50 \le \widetilde{q}_i < 0.80, \\ \texttt{Easy}, & \widetilde{q}_i \ge 0.80. \end{cases} \tag{10}$$

These thresholds categorize posterior mean success estimated from 64 attempts per question; they do not define clusters inferred from the data. Category proportions can differ across collections and forms.

A.5.3 CHECKING A LATENT DIFFICULTY SCALE

A Rasch model represents success as $p_{im} = \sigma(\theta_m - b_i)$, separating condition ability θ_m from item difficulty b_i. It imposes the same ordering of abilities on every item. Small model pools can also limit the accuracy of IRT estimates (Jiang et al., 2026); repeated answers improve estimation within a condition but do not add distinct abilities.

We fit a regularized Rasch model with $\sum_m \theta_m = 0$, Gaussian priors of standard deviation 3, and a binomial likelihood weighted by Mw_m, where M is the number of conditions. We compare it with condition-only, item-only, and separate item–condition predictors, using half-count smoothing for the latter three. The item-only predictor uses the panel weights. In each of 20 random splits, 32 attempts per cell are used for fitting and 32 for prediction. Scores average equally over items and use the same condition weights. This tests further attempts on observed items and conditions.

Difficulty computations use NumPy PCG64 with seed 20260922, adding offsets 1 for posterior draws, 2 for the pair bootstrap, and 100 and 101 for the five- and eight-condition predictive checks.

Random splits draw training success counts from the hypergeometric distribution, equivalent to selecting 32 of the 64 binary responses without replacement. Two additional checks use the ordered halves in both directions.

Appendix B.11.1 reports predictive performance. The latent-scale fit is diagnostic; its parameters do not determine the difficulty labels.

A.5.4 Uncertainty and interpretation

The posterior intervals assume independent, stationary Bernoulli draws within conditions and independent priors across cells. The pair bootstrap assumes exchangeability within each collection. The intervals quantify sampling uncertainty under these assumptions; they do not account for model choice, contamination, or systematic grading errors.

A.6 Formal representations for physics-layer analysis

Formal representations offer a way to inspect the dependencies behind physics-layer labels. Lean and Mathlib represent the mathematical structure, while Physlib supplies definitions of physical systems and laws. A representation can make explicit the chain from physical assumptions through intermediate relations to the source problem's mathematical core. This could support checks of which physical laws a derivation uses and which steps remain when the physical interpretation is removed.

Such checks would complement the expert judgments examined in Appendix B.10, particularly the distinction between an independent physical operation and a direct readout. They could provide a structural basis for layer counts; authenticity would remain a judgment about the physical task.

A.6.1 Preliminary Lean/Physlib representation

As a preliminary proof of concept, we examined the 30 CSAT-derived physics problems using Lean and Physlib. At present, one problem could be represented using the physical concepts available in Physlib. Many of the remaining problems require physical concepts that are not yet formalized in Physlib in a sufficiently general form.

The formalized example uses a harmonic-oscillator model available in Physlib (The Physlib community, 2024). The following excerpt illustrates how the physical potential energy and the optimization target of the reframed problem are represented:

```
import Physlib.ClassicalMechanics.HarmonicOscillator.Basic

def pairPotential (S : HarmonicOscillator) (p q : Point) : R :=
  S.potentialEnergy (separationCoordinate p q)

def objective (c : Configuration) : R :=
  U c - U1 c - U2 c

def IsAnswer (answer : R) : Prop :=
  (exists c : Configuration,
    Admissible c /\ objective c = answer) /\
  (forall c : Configuration,
    Admissible c -> objective c <= answer)
```

The excerpt separates the physical potential from the optimization target: `pairPotential` uses the harmonic-oscillator model, while `IsAnswer` specifies attainment and maximality over admissible configurations. The geometric constraints and optimization relations also require mathematical representations, motivating the combined use of Physlib and Mathlib (The mathlib Community, 2020). This separation exposes the interface between physical assumptions and the inherited mathematical core; the excerpt defines the target property rather than proving the answer.

B ADDITIONAL ANALYSIS OF THE MATH–PHYSICS PERFORMANCE GAP

B.1 EVALUATION SETTINGS AND GRADING

Table 4 specifies the sampling controls. All conditions use 64 attempts, an 81,920-token output ceiling, and default stopping behavior.

Table 4: Model sampling settings. A zero top-k disables top-k truncation. PP denotes presence penalty. Qwen3.8 uses thinking at every effort level.

Model	Modes	Temperature	Top-p	Top-k	PP
GPT-OSS-20B	low / medium / high	1.0	1.00	0	0.0
Qwen3.6-35B-A3B	non-thinking	0.7	0.80	20	1.5
	thinking	1.0	0.95	20	1.5
Qwen3.8-27B	low / medium / xhigh	1.0	0.95	20	0.0

Each user message appends the following instruction to its question:

> Please reason step by step, and put your final answer within `\boxed{}`.

The token limit is a ceiling, not a measurement of consumed compute. GPT-OSS and Qwen3.8 keep sampling parameters fixed across their effort levels. Qwen3.6 changes temperature and top-p when thinking is enabled, so its contrast includes both changes.

Correctness compares the extracted boxed answer with the reference, including equivalent expressions and unit conversions. Flexible scoring accepts an omitted unit by interpreting a numerical value in the reference unit; strict scoring requires explicit units for numerical physical quantities. Both reject incompatible supplied units and retain symbolic physical scales. For the CSAT minimum-product problem, flexible scoring also accepts a correct boxed lower bound, whereas strict scoring requires the minimum. These scoring rules apply to the same attempts.

B.2 PAIRED ANALYSIS

We compare the eight conditions in Section 4.1 on the same 212 pairs. Four CMIMC 2025 pairs do not have a common evaluation of the final paired formulations across all models: two lack GPT-OSS physics evaluations, and two differ in a problem formulation or reference answer. We exclude these four pairs from the common comparison, without imputation. Both Qwen models cover all 216 pairs, allowing a separate analysis of the full dataset.

Let $Y_{ismt} \in \{0, 1\}$ denote correctness for pair i, form $s \in \{\mathrm{M}, \mathrm{P}\}$, condition m, and attempt t. Define

$$\widehat{p}_{ism} = \frac{1}{64} \sum_{t=1}^{64} Y_{ismt}, \qquad g_{im} = \widehat{p}_{i\mathrm{M}m} - \widehat{p}_{i\mathrm{P}m}, \qquad g_i = \sum_m w_m g_{im}, \qquad G = \frac{1}{n} \sum_{i=1}^{n} g_i. \quad (11)$$

Positive gaps favor mathematics. Each pair has equal weight, and each model receives one third of the panel weight, divided equally among its settings. Each GPT-OSS and Qwen3.8 mode has weight $1/9$; each Qwen3.6 mode has weight $1/6$. The primary outcome is flexible correctness; strict correctness is a sensitivity analysis of the same attempts.

The problem pair is the resampling and testing unit. We bootstrap pairs within collections, keeping both forms and all conditions together. Unless stated otherwise, intervals are pointwise 95% percentile intervals, conditional on the six collections and the fixed model panel. Two-sided paired t-tests evaluate mean gaps and changes in gaps. We use Holm correction (Holm, 1979) jointly for 23 performance contrasts: the pooled gap, eight condition gaps, six collection gaps, seven within-model mode contrasts, and the reduction in gap from one to 64 attempts. The analyses and comparison families are retrospective. Subgroup intervals preserve collection counts within each group. The annotation and correlation families are specified in their respective subsections below.

Statistical computation. Analyses use Python 3.13.3, NumPy 2.5.2, and SciPy 1.18.1. Performance intervals use 50,000 bootstrap resamples of whole pairs within collections, retaining both

forms and all conditions. Paired-performance analyses use PCG64 with base seed 20260923 and fixed analysis-specific offsets. The gap and sampling-budget figures use collection-stratified multinomial pair-count draws with seed 20260923. Subgroup intervals preserve collection counts within each group and use seed offsets $100 + 10a + c$, where $a = 0, 1, 2, 3$ indexes difficulty, domain, authenticity, and thickness and c indexes the displayed categories from zero. Collection intervals use offsets 200–205 in the order of Table 6.

B.2.1 Scope of uncertainty

The bootstrap assumes exchangeable problems within each collection and keeps their response sets fixed. Intervals describe uncertainty from problem composition within these six collections. They do not cover new conversions, models, annotation adjudications, or a population of benchmarks.

Paired t-tests and the ANOVA reference distributions are approximate for bounded, potentially skewed pair-level scores. The pooled direction is also supported by the sign test and a tie-adjusted Wilcoxon signed-rank check ($p = 1.70 \times 10^{-16}$); the latter has a symmetry assumption and is not an exact test of the mean. The bootstrap intervals provide a complementary check without a Gaussian shape assumption. Non-rejection throughout denotes uncertainty, not equivalence. Familywise control applies to the stated comparison families, not to every exploratory question jointly.

B.3 Paired performance by collection and condition

The pooled standardized paired effect is $d_z = 0.425$, and the median pair-level gap is 3.65 pp. An exact sign test on the 206 non-tied pairs gives $p = 4.55 \times 10^{-17}$, supporting a disadvantage spread across many problems rather than a few extreme gaps. HMMT Nov. 2025, HMMT Feb. 2026, and CMIMC 2025 retain positive mean gaps under the joint performance correction (Table 6). Omitting any one collection leaves a pooled gap of 4.66–6.40 pp.

Collection-level reversals remain visible in Figure 13: Qwen3.8 xhigh favors physics on AIME 2026 and HMMT Feb. 2026 despite its positive aggregate estimate. Table 5 gives uncertainty and adjusted tests for all eight conditions.

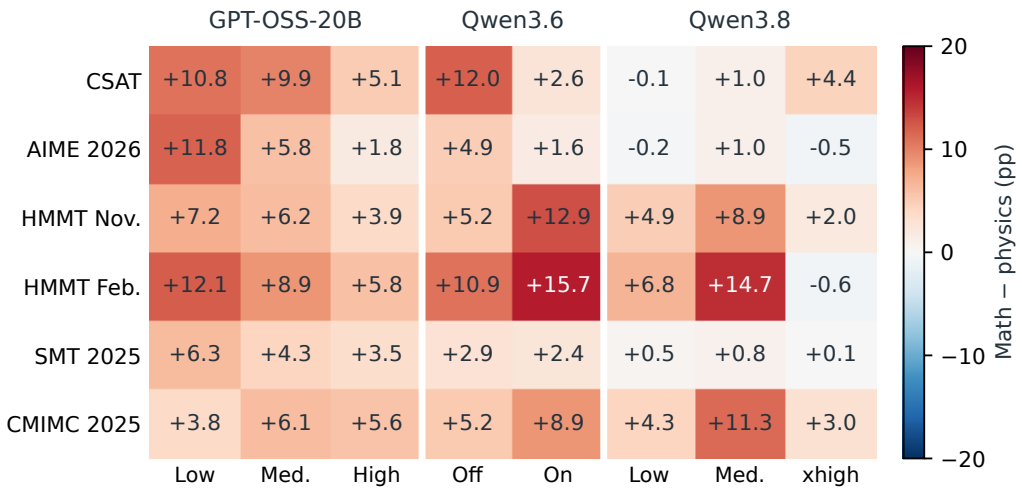


Figure 13: Mathematics-minus-physics accuracy by collection and mode (pp), with the same color scale in every cell. Positive values favor mathematics; negative values favor physics. Off/on denotes Qwen3.6 thinking. All conditions use the common 212-pair panel; cell values are descriptive means.

Table 5: Condition-level gaps (pp), pointwise 95% paired intervals, and p_{H} from the joint 23-contrast performance family.

Model	Mode	Gap	95% CI	p_{H}
GPT-OSS-20B	low	8.33	[5.45, 11.25]	1.53×10^{-6}
	medium	6.60	[4.25, 8.95]	3.00×10^{-6}
	high	4.27	[2.25, 6.29]	8.32×10^{-4}
Qwen3.6-35B-A3B	no thinking	6.43	[4.03, 8.98]	2.35×10^{-5}
	thinking	6.95	[4.27, 9.74]	5.68×10^{-5}
Qwen3.8-27B	low	2.55	[0.41, 4.71]	0.171
	medium	5.96	[3.72, 8.28]	4.28×10^{-5}
	xhigh	1.27	[-0.64, 3.19]	0.844
Pooled panel		5.45	[3.77, 7.14]	6.64×10^{-8}

Table 6: Accuracy by collection, with equal model weights. M/P are accuracies (%); gaps and pointwise 95% intervals are in pp. The correction covers all 23 performance contrasts, including these six collection comparisons.

Collection	Pairs	M	P	Gap	95% CI	p_{H}
CSAT	30	89.22	83.34	5.89	[-2.04, 13.25]	0.844
AIME 2026	30	82.41	79.14	3.27	[0.99, 5.63]	0.087
HMMT Nov. 2025	30	75.33	68.66	6.67	[3.42, 9.88]	0.0050
HMMT Feb. 2026	33	68.43	58.69	9.74	[6.44, 13.18]	5.68×10^{-5}
SMT 2025	53	79.03	76.42	2.61	[-0.56, 6.22]	0.844
CMIMC 2025	36	68.75	62.62	6.13	[2.87, 10.25]	0.031
All pairs	212	77.03	71.58	5.45	[3.77, 7.14]	6.64×10^{-8}

B.4 Reasoning modes

We compare the change in paired gaps between two settings a and b of a model:

$$\Delta G_{a \to b} = \frac{1}{n} \sum_i (g_{ib} - g_{ia}) = \Delta \mathrm{Acc}_{\mathrm{M}} - \Delta \mathrm{Acc}_{\mathrm{P}}. \tag{12}$$

A negative value means that physics improves more and the gap narrows.

Table 7 reports all seven within-model contrasts. The Qwen3.8 low-to-medium and medium-to-xhigh changes survive the joint performance correction; no GPT-OSS or Qwen3.6 contrast does. Qwen3.6 also changes its sampling parameters with thinking (Table 4), so its contrast does not isolate thinking from sampling.

Table 7: Within-model changes in the mathematics-minus-physics gap (pp). Negative changes favor physics improvement. Intervals are pointwise 95%; p_{H} uses the joint performance family.

Model	Change	ΔG	95% CI	p_{H}
GPT-OSS-20B	low → medium	-1.73	[-4.33, 0.85]	0.844
GPT-OSS-20B	medium → high	-2.33	[-4.05, -0.60]	0.087
GPT-OSS-20B	low → high	-4.06	[-6.98, -1.22]	0.068
Qwen3.6-35B-A3B	no thinking → thinking	+0.52	[-2.27, 3.28]	0.844
Qwen3.8-27B	low → medium	+3.41	[1.79, 5.11]	0.0017
Qwen3.8-27B	medium → xhigh	-4.69	[-7.04, -2.36]	0.0032
Qwen3.8-27B	low → xhigh	-1.28	[-3.50, 0.94]	0.844

B.4.1 Repeated-measures ANOVA

ANOVA is applied to the $n \times K$ matrix of pair-level gaps within one model, with problem pair as the repeated unit. This tests the form-by-mode interaction. Greenhouse–Geisser correction adjusts

Table 8: Repeated-measures ANOVA on paired gaps, with Greenhouse–Geisser degrees of freedom. Partial η^2 quantifies the within-model mode effect. Holm correction covers the three omnibus tests.

Model	F	ϵ	df_1	df_2	$\eta^2_{\mathrm{partial}}$	p_{H}
GPT-OSS	5.234	0.784	1.57	330.99	0.0242	0.021
Qwen3.6	0.126	1.000	1.00	211.00	0.0006	0.723
Qwen3.8	9.558	0.862	1.72	363.90	0.0433	6.37×10^{-4}

for departures from sphericity. These three omnibus tests form a separate supporting Holm family; the individual mode claims use the broader 23-contrast family.

For within-model repeated-measures ANOVA, let $Z_{ij} = g_{ij}$ be the gap for pair i and mode j. With n pairs and K modes,

$$F = \frac{n \sum_j (\overline{Z}_{\cdot j} - \overline{Z})^2 / (K-1)}{\sum_{ij} (Z_{ij} - \overline{Z}_{i\cdot} - \overline{Z}_{\cdot j} + \overline{Z})^2 / [(n-1)(K-1)]}. \tag{13}$$

With covariance S across mode gaps and $P = I - \mathbf{1}\mathbf{1}^{\mathsf{T}}/K$, the Greenhouse–Geisser factor $\epsilon = \mathrm{tr}(PSP)^2 / \{(K-1)\,\mathrm{tr}[(PSP)^2]\}$ multiplies both reference degrees of freedom. This adjusts for departures from sphericity (Greenhouse & Geisser, 1959). For two modes, F equals the squared paired t statistic.

B.5 Associations across models

To test whether the same problems incur a physics disadvantage across models, we correlate pair-level gaps after averaging modes within each model. The resulting Spearman correlations are 0.422 for GPT-OSS–Qwen3.6, 0.286 for GPT-OSS–Qwen3.8, and 0.488 for Qwen3.6–Qwen3.8 (Figure 14). All three remain positive after removing collection means from the rank vectors: 0.414, 0.285 and 0.462, respectively. Within-collection permutation tests reject independence of the paired rank assignments for all three comparisons (Holm-adjusted $p = 6.0 \times 10^{-5}$ in each case).

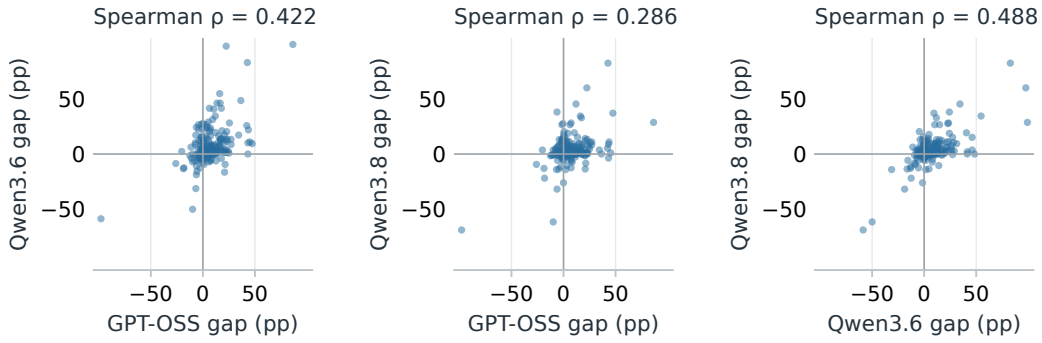


Figure 14: Pair-level mathematics-minus-physics gaps after averaging modes within each model. Each point is one of the same 212 pairs; all panels use the same axes. Positive rank associations persist after collection adjustment.

Each coordinate uses a separate model's responses. Adjusting for ranked mathematics accuracy from both models and for collection also leaves positive associations: the residual rank correlations are 0.417, 0.327 and 0.465. Omitting any single collection leaves raw rank correlations between 0.226 and 0.530. Of the 212 pairs, 82 have a positive gap for all three model averages, compared with 17 having a negative gap for all three. The remaining pairs include mixed signs and ties.

These results show that the penalty partly recurs on the same problems. They do not identify which physical operation, wording change, or solution step causes it. Differences in representation, required physical interpretation, and familiarity with the mathematical source remain possible explanations that final-answer correctness alone cannot separate.

B.6 Domain, authenticity, and physics-layer thickness

The expert-reviewed axes follow Section 3.3. Table 9 reports both accuracies as well as gaps, so a small difference can be distinguished from low success on both forms. Authenticity and thickness enter the joint regression as categorical predictors, with no assumed equal spacing between levels.

Table 9: Subgroup accuracy (%) and mathematics-minus-physics gaps (pp). M/P denote mathematics/physics; intervals are pointwise 95% pair-bootstrap intervals with fixed collection counts within each subgroup. All rows use the common panel and equal model weights. Difficulty is assigned to the mathematics form using the Qwen reference panel.

Subgroup	Pairs	M	P	Gap	95% CI
Math: Easy	152	90.05	82.89	7.16	[5.54, 9.02]
Math: Medium	40	57.44	54.08	3.36	[0.33, 6.40]
Math: Hard	20	17.30	20.63	-3.34	[-7.81, 0.43]
Classical Mechanics	95	80.32	74.30	6.02	[3.94, 8.29]
Electromagnetism	84	73.58	67.51	6.07	[3.58, 8.85]
Quantum Mechanics	19	75.56	75.71	-0.16	[-9.13, 6.37]
Thermodynamics / Stat. Mech.	14	77.39	71.94	5.45	[2.67, 8.18]
Limited	43	74.41	70.99	3.41	[-0.66, 6.70]
Moderate	117	75.64	69.62	6.03	[3.80, 8.48]
Good	52	82.32	76.49	5.83	[3.64, 8.10]
Superficial	58	76.64	73.65	2.99	[-0.72, 6.08]
Normal	103	79.19	73.97	5.22	[3.14, 7.58]
Thick	51	73.12	64.42	8.70	[5.68, 12.15]

After adjustment for collection and all three axes, Holm-adjusted p-values are 0.792 for domain, 0.965 for authenticity, and 0.206 for thickness. The adjusted Thick-minus-Superficial coefficient is 5.87 pp with pointwise interval $[0.85, 10.90]$, but the corrected thickness omnibus test does not establish an axis effect. The full model has in-sample $R^2 = 0.089$, compared with 0.036 for collection effects alone. Using all 216 Qwen pairs changes the uncorrected thickness p-value from 0.069 to 0.024, and the corrected value to 0.072. The observed thickness gradient is therefore sensitive to the panel and adjustment.

B.6.1 Joint model and omnibus sensitivity tests

The annotation regression is

$$g_i = \alpha + \gamma_{\text{collection}(i)} + \boldsymbol{\beta}_D^{\mathsf{T}}\mathbf{D}_i + \boldsymbol{\beta}_A^{\mathsf{T}}\mathbf{A}_i + \boldsymbol{\beta}_T^{\mathsf{T}}\mathbf{T}_i + \varepsilon_i, \tag{14}$$

with dummy-coded axes, 13 coefficients, and 199 residual degrees of freedom. We use joint HC3 tests. For design matrix X, residuals e_i, and leverage h_{ii},

$$\widehat{V}_{\text{HC3}} = (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\,\text{diag}\left(\frac{e_i^2}{(1-h_{ii})^2}\right)X(X^{\mathsf{T}}X)^{-1}. \tag{15}$$

For an axis with r dummy coefficients, the joint Wald statistic is $F = \widehat{\beta}^{\mathsf{T}}\widehat{V}^{-1}\widehat{\beta}/r$, using the corresponding coefficient and covariance blocks and an approximate $F_{r,199}$ reference distribution.

As sensitivity checks, ordinary one-way ANOVA compares unadjusted group means under a common-variance model, Welch ANOVA allows unequal variances, and Kruskal–Wallis compares rank distributions. Kruskal–Wallis is not generally a test of medians unless distribution shapes are comparable. Each test uses one gap per problem pair; applying it to all attempts or treating different modes of a pair as independent observations would inflate the sample size. The joint HC3 model is the primary analysis; the other tests check its sensitivity to modeling assumptions.

B.7 Difficulty rankings and correlations with the gap

Difficulty uses the five-condition Qwen reference panel and the thresholds in Section 3.4. Correlations use the common 212 pairs; category transitions use all 216 pairs.

Table 10: Omnibus annotation tests on the common 212 pairs. The first three columns are unadjusted sensitivity p-values; HC3 tests include collection effects and all three axes. Only the last column applies Holm correction, over the three joint HC3 axis tests.

Axis	ANOVA	Welch ANOVA	Kruskal–Wallis	HC3	HC3 p_{H}
Domain	0.260	0.651	0.148	0.396	0.792
Authenticity	0.506	0.589	0.854	0.965	0.965
Thickness	0.065	0.085	0.088	0.069	0.206

Mathematics and physics success rates have strongly related rankings on the common 212 pairs: Spearman $\rho = 0.761$ $[0.680, 0.828]$. Across all 216 pairs, 182 retain their difficulty category, 29 move to a harder category in physics, and five move to an easier category (Figure 15a). Most pairs retain their category; changes more often increase difficulty.

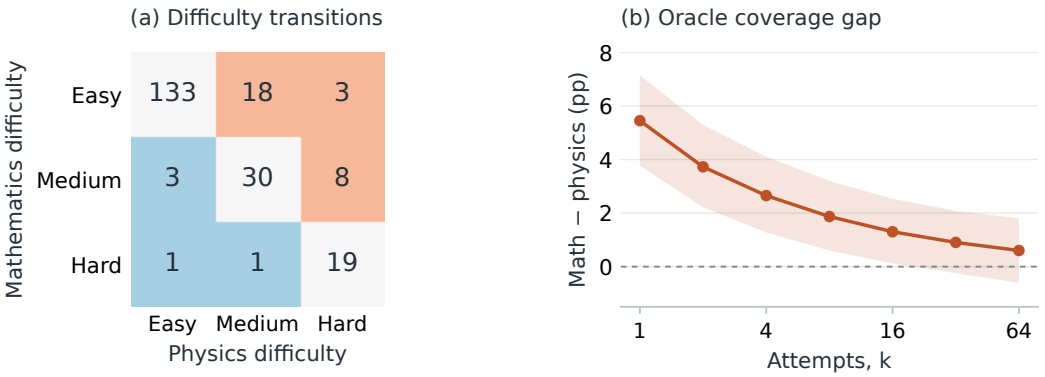


Figure 15: (a) Difficulty transitions for all 216 pairs under the Qwen reference panel: 29 become harder in physics, five easier, and 182 retain their category. (b) Oracle coverage gap on the common 212 pairs, with pointwise 95% paired intervals. Positive values favor mathematics.

The eight-condition mean gap is 7.16 pp among the 152 common pairs labeled Easy on mathematics, 3.36 pp among the 40 Medium pairs, and -3.34 pp among the 20 Hard pairs. These conditional means are descriptive: the grouping variable and the gap share mathematics outcomes. More generally, writing $\widehat{q}_{is} = \sum_m w_m \widehat{p}_{ism}$ for a fixed panel,

$$g_i = \widehat{q}_{i\mathrm{M}} - \widehat{q}_{i\mathrm{P}} = \widehat{d}_{i\mathrm{P}} - \widehat{d}_{i\mathrm{M}}, \qquad \widehat{d}_{is} = 1 - \widehat{q}_{is}. \tag{16}$$

A correlation between difficulty and the gap can therefore arise through this arithmetic relationship, together with ceiling and floor constraints.

We check this dependence in two ways (Table 11). First, mathematics difficulty estimated from attempts 1–32 is correlated with a gap estimated from attempts 33–64, and then the halves are reversed. Neither association passes correction. Second, mathematics difficulty from one Qwen model predicts the gap of the other. Qwen3.8 difficulty has a modest negative association with the Qwen3.6 gap (collection-adjusted rank correlation -0.198, adjusted $p = 0.026$), whereas the reverse comparison is near zero (-0.014, adjusted $p = 0.840$). Separating response sets removes shared sampling noise but does not remove the algebraic dependence of a true gap on true mathematics success. These correlations do not support a general or causal claim that harder mathematics incurs a larger physics penalty.

B.7.1 Rank associations and response separation

Ranks use average ranks for ties, with numerical scores rounded to 12 decimal places before ranking. Collection adjustment subtracts the within-collection mean of each rank vector before computing their correlation. For a two-sided permutation test, one rank vector is shuffled within collection

Table 11: Difficulty associations and checks separating response sets. ρ is raw Spearman correlation; ρ_{C} removes collection effects from ranks. Intervals concern ρ; permutation p_{H} concerns ρ_{C} and adjusts all seven rows. Row 1 compares success rates; remaining rows compare the named difficulty predictor with a paired gap.

Comparison or predictor	ρ	95% CI	ρ_{C}	p_{H}
Math vs. physics success	0.761	[0.680, 0.828]	0.742	1.40×10^{-4}
Mathematics difficulty	-0.081	[-0.211, 0.055]	-0.142	0.174
Physics difficulty (coupled)	0.445	[0.315, 0.566]	0.402	1.40×10^{-4}
Math difficulty 1–32; gap 33–64	-0.067	[-0.199, 0.069]	-0.135	0.174
Math difficulty 33–64; gap 1–32	-0.075	[-0.206, 0.061]	-0.120	0.174
Qwen3.6 math difficulty; Qwen3.8 gap	0.056	[-0.082, 0.197]	-0.014	0.840
Qwen3.8 math difficulty; Qwen3.6 gap	-0.162	[-0.289, -0.025]	-0.198	0.026

49,999 times, with $p = (1 + \#\{|T^*| \geq |T|\})/50{,}000$. This tests exchangeability of the paired rank assignments within each collection. Raw Spearman intervals use 20,000 stratified pair-bootstrap replicates and rerank in every replicate; they concern the raw correlation, whereas the permutation p-values concern collection-adjusted association.

The three cross-model gap comparisons constitute one Holm family. The seven comparisons in Table 11 constitute another. Difficulty correlations use the five-condition Qwen panel on the common 212 pairs, except where a single model is named. The split-response comparisons keep all 32 attempts of a half together when estimating each success rate. Difficulty thresholds and correlations remain descriptive of this evaluated panel.

B.8 Repeated sampling reduces the gap in oracle coverage

Pass@k uses Equation 9, calculated within each condition before applying the eight-condition panel weights. Attempts from different models are not pooled into one solver. The estimator is unbiased under independent sampling at fixed success probability.

The panel gap decreases from 5.45 pp at one attempt to 2.65 pp at four, 1.30 pp at 16, and 0.60 pp at 64 (Figure 15b). At 64 attempts, oracle coverage is 94.00% for mathematics and 93.40% for physics; the gap interval is $[-0.60, 1.81]$ pp. The paired reduction from one to 64 attempts is 4.85 pp $[3.59, 6.10]$ (adjusted $p = 9.14 \times 10^{-11}$). The gap narrows at a larger sampling budget. This does not establish equivalence: pass@64 uses an oracle, its interval is not an equivalence bound, and the result does not extend beyond 64 attempts.

B.9 Sensitivity to scoring, weighting, and coverage

The average gap remains positive under strict correctness, equal condition weights, selection of each model's highest-effort mode, and restriction to or expansion of the Qwen panel (Table 12). Strict scoring gives a larger gap, 9.84 pp, but the flexible-score gap is already positive. Excluding CSAT leaves gaps of 5.38 pp under flexible and 7.64 pp under strict scoring. The difference between scoring criteria therefore affects the magnitude, but does not account for the direction of the result.

B.10 Annotation reliability and final labels

The diagnostics use the eight runs before adjudication and the definitions in Appendix A.4.

B.10.1 Agreement across runs

Domain agreement is high, at 98.18%, while authenticity and thickness have 70.78% and 64.62% agreement (Table 13). Most graded disagreements are between adjacent categories: Limited–Good pairs account for only 1.70% of disagreeing authenticity pairs, and Superficial–Thick pairs for 1.82% of disagreeing thickness pairs. Ordinal α is correspondingly higher than nominal α, although uncertainty at category boundaries remains. Only 16 of 216 problems have a unanimous three-axis annotation.

Table 12: Sensitivity of the mean mathematics-minus-physics gap. All rows use paired outcomes. Intervals are pointwise 95% pair-bootstrap intervals in pp.

Analysis	Pairs	Gap	95% CI
Primary: equal model weights	212	5.45	[3.77, 7.14]
Strict correctness	212	9.84	[7.33, 12.42]
Equal condition weights	212	5.29	[3.64, 6.96]
Highest-effort mode per model	212	4.16	[2.37, 6.00]
Qwen panel, all pairs	216	4.90	[3.12, 6.76]
Qwen panel, common pairs	212	4.98	[3.18, 6.85]
Flexible, excluding CSAT	182	5.38	[3.91, 6.95]
Strict, excluding CSAT	182	7.64	[5.47, 10.04]
Equal collection weights	212	5.72	[3.96, 7.45]

Table 13: Agreement among the eight AI runs before adjudication. Unanimous and tied counts are out of 216 problems on each axis. $\overline{H}$ is mean normalized item entropy; brackets give conditional 95% bootstrap intervals. A tie means that multiple categories share the largest vote count.

Axis	A (%)	Unanimous	Tied	$\overline{H}$	α_{nominal}	α_{ordinal}
Domain	98.18	206	1	0.017	0.971 [0.951, 0.988]	—
Authenticity	70.78	66	15	0.355	0.502 [0.444, 0.558]	0.649 [0.592, 0.699]
Layer thickness	64.62	46	26	0.422	0.451 [0.402, 0.497]	0.660 [0.612, 0.701]

Category use varies across runs. Good authenticity ranges from 37 to 87 problems per run (17.1–40.3%), while Superficial thickness ranges from 17 to 114 (7.9–52.8%). One source of thickness disagreement is whether an apparatus and its readout count as separate layers or as one physical mapping. Removing one run at a time leaves ordinal α between 0.630 and 0.658 for authenticity and between 0.637 and 0.691 for thickness; the disagreement is therefore not confined to a single run.

B.10.2 Consensus and subset stability

Initial consensus uses equally weighted votes on each axis. A strict majority of at least five votes exists for 601 of 648 decisions; five more have a unique plurality. The remaining 42 decisions are tied four to four: one domain, 15 authenticity, and 26 thickness labels. For the ordered axes, selecting the best-supported category within the ordinal median set gives the same candidate set as the mode on every problem. We therefore use per-axis modes, with ambiguous cases resolved through review.

The unique eight-run mode is available for 215 domain, 201 authenticity, and 190 thickness labels; tied full-panel modes are excluded from the subset-stability diagnostic.

With one run, the unique mode matches the full-panel reference for 83.5% of authenticity and 79.8% of thickness item–subset pairs. With five runs, these fractions rise to 95.8% and 94.3%. Even subset sizes produce more ties, so stability need not increase monotonically. The reference includes the sampled runs, and agreement at $k = 8$ is one by construction. This measures consistency within the observed panel, not annotation accuracy.

B.10.3 Adjudication outcomes and expert review

AI adjudication covers 56 decisions across 53 problems: 10 domain, 18 authenticity, and 28 thickness decisions. It resolves all 42 ties and changes four unique modal choices. The other 592 decisions retain their strict-majority labels.

Two physics experts then review the annotations for all 216 problems. Their review changes 66 of the 648 labels (10.2%) across 60 problems: 40 authenticity labels, 25 thickness labels, and one domain label. Table 14 reports the final annotations.

Table 14: Final label counts after review by two physics experts. CM, EM, QM, and TSM denote Classical Mechanics, Electromagnetism, Quantum Mechanics, and Thermodynamics/Statistical Mechanics; authenticity columns are Limited, Moderate, Good, and thickness columns are Superficial, Normal, Thick.

		Domain				Authenticity			Thickness		
Collection	N	CM	EM	QM	TSM	L	M	G	S	N	T
AIME 2026	30	14	15	0	1	5	20	5	2	20	8
HMMT Nov. 2025	30	13	15	1	1	6	16	8	10	15	5
HMMT Feb. 2026	33	16	13	3	1	5	19	9	10	13	10
SMT 2025	53	25	20	2	6	10	30	13	16	26	11
CMIMC 2025	40	16	14	7	3	14	21	5	15	18	7
CSAT	30	13	9	6	2	5	13	12	8	11	11
Total	216	97	86	19	14	45	119	52	61	103	52

Table 15: Held-out prediction, averaged over 20 splits of the response bank. Lower log loss and Brier score are better. The panels cover 432 and 424 records, respectively; splits are not independent experiments.

	Five conditions		Eight conditions	
Predictor	Log loss	Brier	Log loss	Brier
Condition only	0.4603	0.1460	0.5062	0.1656
Item only	0.3022	0.0937	0.3955	0.1268
Rasch	0.2589	0.0775	0.2899	0.0878
Item–condition probabilities	0.2276	0.0679	0.2524	0.0766

The authenticity changes distinguish physical reasoning from artificial answer encoding. The AIME 2026 impulse controller, which implements a custom binary operation through odd/even packet rules, receives Limited authenticity. In contrast, the HMMT Nov. 2025 wire loop in a time-varying magnetic field receives Good because the induced emf is a direct Faraday-law measurement.

Thickness changes concern the number of distinct physical operations. The CMIMC 2025 race problem changes from Thick to Superficial because renaming racers as carts supplies a direct reinterpretation. The HMMT Feb. 2026 AC-network problem changes from Normal to Thick: it combines cascaded complex gains, coherent path summation, and a null phasor measurement. The single domain change is the CSAT Coulomb-orbit problem, from Electromagnetism to Classical Mechanics, because its central task is orbital motion under a specified force.

B.11 Difficulty validation and sensitivity

These diagnostics use the five-condition Qwen reference panel on all 216 pairs unless the expanded eight-condition panel is named. Appendix A.5 defines the estimators and predictive checks.

B.11.1 Held-out prediction

Separate item–condition estimates outperform Rasch in every random split of both panels (Table 15). The ordering also holds when training on the first 32 attempts and testing on the remaining 32, and when reversing those halves. We therefore retain condition-specific probabilities for the difficulty analysis. This comparison does not exclude more flexible IRT models.

B.11.2 Difficulty labels and their stability

The dataset contains 291 Easy, 90 Medium, and 51 Hard records (Table 16); 21 mathematics records and 30 physics records are Hard. For 63 records, a 95% credible interval crosses a category boundary. Continuous estimates, intervals, and category probabilities accompany the labels to retain this uncertainty.

Table 16: Difficulty labels under the five-condition Qwen reference panel and Equation 10.

	Mathematics			Physics		
Collection	Easy	Medium	Hard	Easy	Medium	Hard
CSAT	27	2	1	25	4	1
AIME 2026	25	3	2	23	5	2
HMMT Nov. 2025	21	6	3	16	11	3
HMMT Feb. 2026	18	10	5	13	12	8
SMT 2025	41	7	5	41	7	5
CMIMC 2025	22	13	5	19	10	11
Total	154	41	21	137	49	30

Labels are more sensitive to the reference panel than to modest smoothing changes. A uniform $\mathrm{Beta}(1,1)$ prior changes 3/432 labels; equal weighting of conditions changes 12/432. Using only the highest-effort setting of each model changes 103/432, while the expanded panel changes 90/424 on matching inputs. Despite these category changes, the expanded-panel scores have Spearman correlation 0.949 with the reference scores. Giving the six collections equal weight yields a physics-minus-mathematics difference of -5.72 points $[-7.44, -3.97]$ under the expanded panel.

Moving the thresholds to 0.45/0.75 or 0.55/0.85 changes 34 or 41 labels, respectively. Strict scoring changes 20 labels; its largest effect on a dataset mean is CSAT physics, whose success falls from 90.2% to 66.9%. Labels therefore depend on the reference panel and scoring rule; continuous scores retain differences within categories.

C Test-time scaling and reinforcement learning details

C.1 Oracle coverage across model conditions

Figure 16 shows how repeated sampling reduces the mathematics–physics gap across the evaluation conditions. At 64 attempts, six of the eight conditions reach equal aggregate coverage on the two forms. GPT-OSS low and Qwen3.6 non-thinking retain gaps of 1.89 and 2.36 pp, respectively.

Coverage is computed within each problem and condition using Equation 9, then averaged over the 212 common pairs. Each model receives equal weight, divided among its modes: $1/9$ for each GPT-OSS and Qwen3.8 mode, and $1/6$ for each Qwen3.6 mode. Pointwise 95% confidence intervals resample whole problem pairs within collections, retaining both forms and all conditions together.

C.2 Preference data and training

Response sources and problem selection. Mathematics and physics training use the same pool of 60 problem families, sampled with seed 42, ten from each collection. We use English questions and combine GPT-6 Astra solutions with responses from the eight open-weight model conditions in Section 4.1. The preference sets were fixed from the September 20, 2026 rollout snapshot, before all source collections were complete. Subsequent rollouts do not enter these training runs. Unscored responses are excluded rather than treated as incorrect. Text screening excludes empty responses, unresolved chat-control markers, and malformed thinking boundaries; source exports lack finish reasons, so this screen does not establish completion from generation metadata. To retain consistent problem versions, we exclude GPT-OSS CMIMC mathematics questions 7 and 23 and physics question 23. The released selection records identify the sampled families and compatible sources. The preference-set configuration records the source snapshot, tokenizer and framework revisions, and seed; the family-selection records define evaluation exclusions.

Preference construction. Preference pairs favor a flexible-correct response over an incorrect response to the same question. Flexible scoring permits omitted units; we do not construct a separate preference class for formatting errors. Thinking responses are paired only with thinking responses, and non-thinking responses only with non-thinking responses. Within each collection, subject, and response group, we rank sources by flexible accuracy on the ten selected training families, placing Astra first in the non-thinking group. The chosen source must rank at least as highly as the rejected

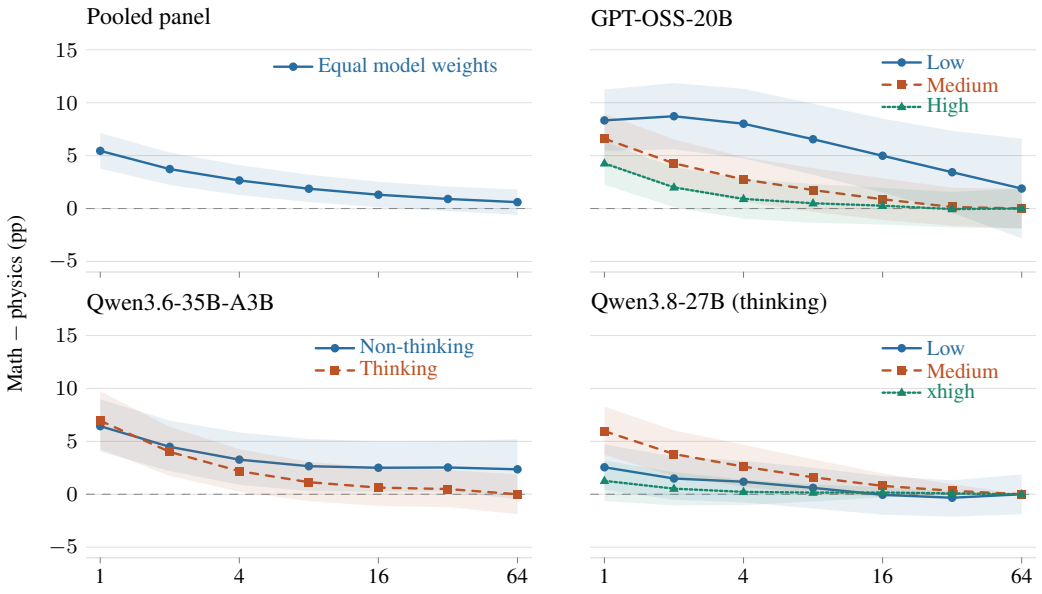


Figure 16: Mathematics minus physics oracle coverage on 212 paired problems, pooled with equal model weights and by model and mode. All panels share the same vertical scale; horizontal axes show the number of attempts k on a logarithmic scale. Shaded bands indicate pointwise 95% confidence intervals from a stratified pair bootstrap; dashed horizontal lines mark zero gap.

source. Source rank restricts eligible comparisons; it does not replace the correctness label or certify intermediate reasoning.

We deduplicate whitespace-equivalent retained responses, keeping the highest-ranked source, and exclude retained texts with conflicting labels. Pair selection first maximizes the number of distinct responses covered, then minimizes repeated appearances, subject to a maximum of two uses per response across either role. We balance the total number of pairs between mathematics and physics separately at each cutoff. Families without eligible pairs are not replaced: 57 mathematics and 59 physics families contribute examples.

The two subjects share the initial family pool and total pair counts, but differ in contributing families, response mixtures, and token exposure (Table 17).

Response representation. Full parameter updates use 4,632 preference pairs per subject with a 2,048-token cutoff. LoRA uses 5,033 pairs per subject with an 8,192-token cutoff. The cutoff applies to the combined prompt and response in each branch of a preference pair. Prompts are preserved in full; thinking responses may be shortened to a prefix, while non-thinking responses must fit within the cutoff. Available Qwen thinking text is retained; GPT-OSS contributes only its final channel, which can itself contain a step-by-step solution. All responses use the same Qwen3.5 thinking-enabled chat template (`qwen3_5`); responses without thinking text receive an empty thinking block. Training prompts use the boxed-answer instruction in Appendix C.3. For truncated responses, tokenization preserves the exact leading token sequence, without inserting an artificial closing thinking tag or end-of-sequence token into the scored prefix. Preference labels are inherited from the original final answers, which may lie beyond the cutoff; the retained prefixes are not independently graded.

Optimization. Each run starts independently from the post-trained Qwen3.5-4B model, without additional supervised fine-tuning. The reference policy is the same initial model and remains fixed; LoRA obtains this reference by disabling the newly trained adapter. We use the sigmoid DPO objective with $\beta = 0.1$, no label smoothing, and no auxiliary supervised loss. Response log-probabilities are summed over scored tokens, with prompt tokens masked. Training uses one epoch, effective batch size 16, seed 42, and a cosine learning-rate schedule with 10% warmup. Peak learning rates are 5×10^{-7} for full updates and 5×10^{-6} for LoRA. AdamW uses zero weight decay and gradient clipping at norm 1.0; full updates use its fused implementation. The LoRA rank is 8, with $\alpha = 16$

Table 17: Composition of the mixed preference sets. Truncated counts pairs with at least one shortened response. Tokens sum the serialized prompt and retained response over both branches, including repeated responses; they measure sequence exposure rather than training FLOPs.

Update	Subject	Thinking	Non-thinking	Truncated	Tokens (M)
Full	Mathematics	1,017	3,615	995	11.08
	Physics	1,202	3,430	1,146	11.94
LoRA	Mathematics	1,017	4,016	915	27.77
	Physics	1,193	3,840	1,069	30.57

and zero dropout, applied to the attention, gated-delta, and feed-forward linear projections. The released adapter configurations specify the target modules.

Held-out families. The held-out set contains a further ten families per collection, sampled after excluding all 60 training families across both forms. All translations and alternative versions of a selected family remain excluded, including families that yield no training pairs. The same 60 test families are used for all four trained checkpoints. This split evaluates transfer to new problems within the six source collections.

C.3 Evaluation protocol

The physics comparison uses 60 held-out physics questions with eight attempts per question for each of four trained checkpoints: mathematics and physics DPO, each with LoRA or full parameter updates. This gives 1,920 scored physics responses. All four conditions use the same question text, references, thinking chat template, decoding parameters, and generation seeds.

Each question is supplied in a single user message with the instruction:

> Please reason step by step, and put your final answer within `\boxed{}`.

The shared template opens an assistant thinking block. Sampling uses temperature 0.7, top-p 0.8, top-k 20, and presence penalty 1.5, with an 81,920-token completion limit and a 131,072-token context. All conditions stop at the same end-of-sequence token (`<|endoftext|>`, token 248044). The generation seed is $8i+r-1$, where the physics questions have fixed indices $i = 60, \ldots, 119$ and $r = 1, \ldots, 8$ indexes attempts. Inference uses vLLM 0.28.0 in BF16 on A100 and H200 GPUs, with 32 concurrent requests per endpoint. LoRA weights are merged into the model before evaluation.

Grading compares the last complete boxed answer with the reference, requires explicit units for numerical physical quantities, checks supplied units, and retains symbolic physical scales. All responses, including those reaching the token limit, are graded from the saved answer; missing or malformed boxes receive zero.

C.4 Paired accuracy comparison

Let $Y_{idr}^{(u)}$ indicate correctness on physics problem i and attempt r, where $d \in \{\mathrm{P}, \mathrm{M}\}$ denotes the training subject and $u \in \{\mathrm{LoRA}, \mathrm{full}\}$ denotes the update regime. Accuracy and the difference between training subjects are

$$\widehat{a}_d^{(u)} = \frac{1}{60}\sum_{i=1}^{60}\frac{1}{8}\sum_{r=1}^{8} Y_{idr}^{(u)}, \qquad \widehat{\Delta}^{(u)} = 100\left(\widehat{a}_{\mathrm{P}}^{(u)} - \widehat{a}_{\mathrm{M}}^{(u)}\right). \tag{17}$$

Here $\widehat{\Delta}^{(u)}$ is expressed in percentage points. Physics DPO produces 280 correct responses out of 480 with LoRA and 278 with full updates; mathematics DPO produces 274 in each regime. These counts give the differences in Table 2.

Table 2 reports uncertainty for the two direct comparisons. We first average the eight attempts within each problem, then compute differences on matched problems. Pointwise 95% confidence intervals use 20,000 stratified bootstrap replicates, resampling ten problem families with replacement within each collection and retaining all four conditions and their attempts together. Resampling uses seed

42; accuracy differences are computed before rounding. The intervals condition on the evaluated checkpoints and response samples.

D Potential applications

D.1 Using Math2Physics

Math2Physics can support physics post-training through supervised fine-tuning on checked solutions and preference learning from correct and incorrect responses. Its paired structure enables comparisons of mathematics and physics training with matched source problems and data budgets. Our DPO experiment provides an initial example, although it does not establish an advantage for physics training (Section 5). All formulations and responses derived from one source problem should remain in the same data split to prevent overlap between training and evaluation.

The dataset also offers problems for studying autoformalization and verified solving in Lean 4 using Physlib,[1] whose predecessor HepLean formalizes physics definitions, calculations, and proofs (Tooby-Smith, 2025). A formalized subset could pair each natural-language question with explicit physical assumptions, a theorem stating the requested result, and a machine-checked proof. Proving the reduction to the shared mathematical core would allow separate evaluation of physical modeling and proof completion. This requires adding definitions where library coverage is insufficient and checking that the formal statement faithfully represents the original problem; proof checking alone does not establish that correspondence or the suitability of the physical assumptions.

The released response banks could also support training and evaluating answer verifiers. Selecting among sampled solutions, as studied by Lightman et al. (2024), directly addresses the difference between single-attempt accuracy and oracle coverage observed in Section 5. Final-answer labels support outcome supervision; supervising intermediate reasoning would require additional step-level annotations. The physics labels also permit comparisons across domains, authenticity levels, and physics-layer thickness.

D.2 Generating new physics datasets

The `math-to-physics` skill can be reused with further mathematics collections or adapted to physics sources by extracting their governing assumptions and derivations before constructing new physical settings. The latter extends the use of LLMs to generate isomorphic physics problem banks (Liu et al., 2026). Each variant should retain source attribution and undergo renewed solution and physical-validity checks; shared structure does not guarantee equal difficulty. We recommend review by physics experts, as in this work, to assess the assumptions, mathematical correspondence, and answers before releasing a new dataset.

[1] `https://physlib.io/`