FROM REDISCOVERY TO DISCOVERY: VERIFIABLE EVALUATION ON OPEN MATHEMATICAL PROBLEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent progress in AI for mathematics has pushed frontier models from competition-style problem solving toward research-level reasoning, yet how to measure verifiable progress on genuinely open mathematical problems remains largely unresolved. Most mathematical benchmarks evaluate reasoning against known answers, while existing open-problem settings do not clearly distinguish recovering established results, making measurable progress, and resolving open questions. We introduce **OpenProblemBench**, a result-verifiable benchmark of 65 mathematical research tasks spanning **L1 Rediscovery**, **L2 Progressive Discovery**, and **L3 Decisive Discovery**. We systematically evaluate frontier reasoning models, long-horizon agents, and discovery systems under this common setting. Our results reveal markedly different capability profiles across the three levels: frontier models can already rediscover multiple recently established results; long-horizon interaction substantially improves feasible-solution coverage and task-native metrics on L2 Progressive Discovery; yet no evaluated configuration reaches L3 Decisive Discovery. To better convert additional test-time computation into mathematical progress, we introduce **MathMap**, a simple yet effective verifier-grounded harness that combines diverse exploration, external verification, and verified selection. MathMap builds a persistent Verified Discovery Tree by expanding parallel tool-using branches, externally verifying candidate solutions, and continuing from the strongest verified checkpoint. Under matched backbone and time budgets, MathMap achieves rank 1 on 14 of 20 L2 tasks among the four same-backbone discovery systems, while using substantially fewer completion tokens and solution submissions. Together, OpenProblemBench and MathMap provide an open foundation for measuring AI progress from rediscovery through progressive discovery to decisive mathematical discovery.

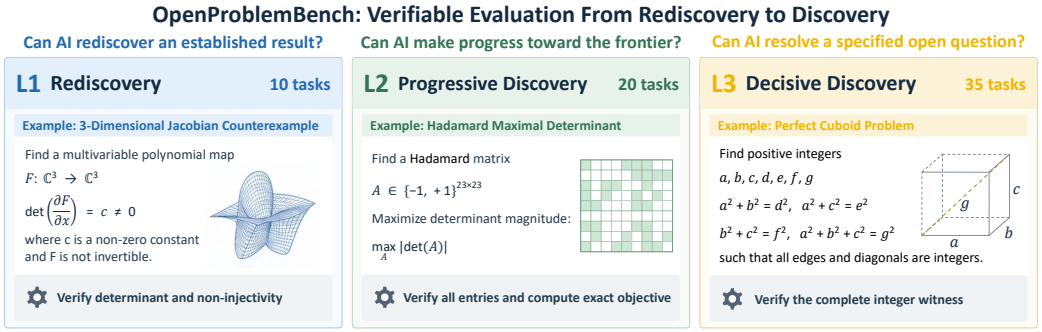


Figure 1: **OpenProblemBench is a result-verifiable benchmark spanning three increasingly consequential levels of mathematical research capability:** recovering established results (L1 Rediscovery), making verifiable progress toward the research frontier (L2 Progressive Discovery), and resolving specified open questions (L3 Decisive Discovery).

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.

1 INTRODUCTION

AI for mathematics is advancing rapidly, with frontier models and agentic systems moving beyond competition-style problem solving toward the research frontier. Terence Tao describes this shift as a transition from *proof scarcity* to *proof abundance* (Tao, 2026). Recent AI-assisted advances on Erdős's unit-distance conjecture (Alon et al., 2026), the three-dimensional Jacobian conjecture (Gao, 2026), Sendov's conjecture (Mazur, 2026), and even the forced Navier–Stokes Millennium formulation (OpenAI, 2026) already reflect this transition. As AI begins to engage with research problems whose answers are not known in advance, the evaluation problem changes as well. Progress can occur before a problem is fully resolved, and any claimed advance must be judged by what the resulting mathematical artifact actually establishes. This creates a growing gap between what AI can now attempt and what current evaluations can reliably measure, raising a central question: *how can we measure verifiable progress on open mathematical problems?*

Existing benchmarks capture important parts of this question, but not the full progression of research capability. Competition-style benchmarks evaluate reasoning against answers already known to the evaluator (Sun et al., 2026; Liu et al., 2026; Luong et al., 2025), providing a controlled measure of mathematical correctness but offering no graded notion of progress once the target itself is unknown. Recent open-problem platforms move closer to research settings through executable verification and long-horizon interaction (Wang et al., 2026; Bianchi et al., 2026; Epoch AI, 2026a), yet a common evaluation structure is still needed to distinguish different mathematical consequences of research progress. In particular, open-problem evaluation should answer three complementary questions: *Can AI rediscover an established result? Can AI make progress toward the research frontier?* and, ultimately, *Can AI resolve a specified open question?* These questions correspond to distinct capabilities of reconstruction, advancement, and resolution. Each is meaningful in its own right, motivating an evaluation that measures all three within a common, result-verifiable setting rather than reducing research capability to either known-answer correctness or final breakthrough alone.

We therefore introduce **OpenProblemBench**, a result-verifiable benchmark of 65 tasks organized into three increasingly consequential levels of mathematical research capability. **L1 Rediscovery** comprises 10 tasks that ask AI to independently recover established research results, providing a controlled setting for evaluating research-level reconstruction. **L2 Progressive Discovery** comprises 20 tasks that reward increasingly strong feasible constructions under task-native objectives, making measurable progress toward the research frontier visible even when the underlying open question remains unresolved. **L3 Decisive Discovery** comprises 35 tasks in which success requires an exact accepted artifact that resolves a specified open question. These levels do not form a calibrated difficulty scale. Instead, they distinguish three different mathematical consequences: recovering what is already known, advancing an unresolved research objective, and crossing the boundary from progress to resolution. Across all three levels, each task specifies the required artifact, executable acceptance conditions, a native objective where applicable, and the precise mathematical consequence of acceptance. This shared result-verifiable interface makes these distinct forms of research capability directly auditable within a single benchmark, enabling consistent evaluation of reasoning models, long-horizon agents, and discovery systems.

Our evaluation reveals sharply different capability profiles across these three levels. On L1 Rediscovery, frontier reasoning models can already recover several established research results, including recently resolved Erdős problems, while long-horizon interaction provides moderate aggregate gains, increasing the paired pass rate from 30% to 40%. On L2 Progressive Discovery, additional test-time computation is substantially more effective: feasible-solution coverage rises from 51% to 82%, and long-horizon agents and discovery systems often produce stronger feasible constructions under task-native objectives. On L3 Decisive Discovery, however, this progress has not yet translated into resolution, with no accepted artifact in the reported evaluations resolving a specified open question. Together, these results show that current AI can already perform research-level rediscovery and increasingly convert long-horizon computation into verifiable frontier progress, while the transition from progressive to decisive discovery remains open. This gap suggests that long-horizon mathematical research requires more than additional computation alone: AI must be able to explore alternative directions, verify intermediate results, and reliably build on established progress over time.

Motivated by this observation, we introduce **MathMap**, A simple yet effective verifier-grounded harness that organizes long-horizon research through diverse exploration, external verification, and

verified selection. Starting from the current verified checkpoint, MathMap explores three independent tool-using branches, evaluates their candidates with the external task verifier, and selects the strongest verified result as the starting point for the next round. Across rounds, verified candidates and their parent relations form a persistent Verified Discovery Tree, providing an explicit map of explored directions and mathematical progress. Despite its simple *explore–verify–select* design, MathMap achieves strong empirical performance. On L1 Rediscovery, it reaches a 60% pass rate, matching GPT-5.6-Sol. On L2 Progressive Discovery, with DeepSeek-V4.1-Flash as the shared backbone and the same two-hour budget, MathMap attains or ties Rank 1 on 14 of 20 tasks among the four same-backbone discovery systems, while using substantially fewer completion tokens and solution submissions. A gap nevertheless remains to the stronger GPT-5.6-Sol + Codex configuration, leaving clear headroom toward more capable long-horizon mathematical discovery.

Our main contributions are:

- **A result-verifiable benchmark spanning three levels of mathematical research capability.** We introduce OpenProblemBench, comprising 65 mathematical research tasks across L1 Rediscovery, L2 Progressive Discovery, and L3 Decisive Discovery. The three levels distinguish recovering established results, making measurable progress on unresolved research objectives, and resolving specified open questions, with executable acceptance conditions, task-native objectives where applicable, and explicit mathematical consequences.
- **A systematic evaluation of AI from rediscovery to discovery.** We evaluate frontier reasoning models, long-horizon agents, and discovery systems under a common result-verifiable setting. The results reveal distinct capability profiles across the three levels: frontier models can already recover recent research results, long-horizon interaction substantially improves Progressive Discovery, while no evaluated configuration reaches L3 Decisive Discovery.
- **A simple yet effective verifier-grounded harness for long-horizon mathematical research.** We introduce MathMap, which combines diverse exploration, external verification, and verified selection through a persistent Verified Discovery Tree. With DeepSeek-V4.1-Flash as the backbone and a two-hour budget, MathMap matches strong L1 performance and attains Rank 1 on 14 of 20 L2 tasks against same-backbone discovery systems, while using substantially fewer completion tokens and solution submissions.

2 OpenProblemBench

To measure verifiable progress on open mathematical problems, we introduce **OpenProblemBench**, a result-verifiable benchmark of 65 tasks organized into three levels: **L1 Rediscovery**, **L2 Progressive Discovery**, and **L3 Decisive Discovery**. Unlike benchmarks limited to known-answer correctness or final breakthroughs, OpenProblemBench spans recovery of established results, measurable progress on unresolved problems, and resolution of open questions. Executable verification and task-native objectives keep these outcomes distinct and the boundary between progress and discovery clear.

2.1 Three Research Capability Levels: From Rediscovery to Discovery

L1 Rediscovery (10 tasks). L1 revisits recently resolved open problems, asking AI to independently rediscover the established result. Because these problems lie close to the current research frontier while remaining exactly verifiable, they provide a demanding test of research-level reconstruction. Four tasks are based on Erdős problems resolved in 2026 (e.g., Beeson et al. (2026)), while the remaining six require explicit constructions or counterexamples. Representative examples include a three-dimensional counterexample to the Jacobian conjecture, open since 1939 and refuted in July 2026 (Gao, 2026), and a Hadamard matrix of order 668, unresolved for over two decades until August 2026 (Epoch AI, 2026b). L1 therefore evaluates whether AI can independently reconstruct recent research advances under exact result-level verification.

L2 Progressive Discovery (20 tasks). L2 moves beyond rediscovering established results to ask whether AI can make verifiable progress on unresolved research objectives. By casting each task as an optimization problem over feasible constructions, L2 makes meaningful progress measurable even before the underlying open question is fully resolved. It includes six fixed-instance optimization tasks and fourteen research-oriented optimization tasks. The former cover classical settings such as the

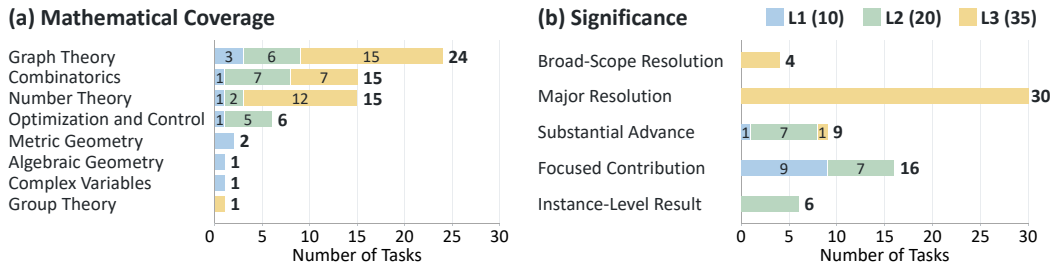


Figure 2: **Mathematical coverage and significance.** Stacked bars show the numbers of L1 Rediscovery, L2 Progressive Discovery, and L3 Decisive Discovery tasks in each category.

traveling salesman problem (Dantzig et al., 1954), while the latter include low-autocorrelation binary sequences (Packebusch & Mertens, 2016) and the Hadamard maximal-determinant problem (Orrick et al., 2003). Each feasible construction receives a task-native score, allowing L2 to quantify AI's progress and assess whether stronger constructions have the potential to advance the research frontier.

L3 Decisive Discovery (35 tasks). L3 targets the point at which measurable progress becomes a decisive resolution of a specified open question. Unlike L2, stronger intermediate constructions are not sufficient: success requires an exact artifact that satisfies all acceptance conditions and establishes the mathematical consequence defined by the task. Eight tasks require positive constructions, including a perfect cuboid (Guy, 2004), a Conway 99-graph (Cesarz & Woldar, 2025), and the missing Moore graph (Dalfó, 2019). The remaining 27 seek counterexamples to longstanding conjectures, including the Collatz conjecture (Lagarias, 1985), Hadwiger's conjecture (Seymour, 2016), and the union-closed sets conjecture (Bruhn & Schaudt, 2015). Any accepted L3 result would constitute a genuine mathematical advance within the declared task scope, allowing L3 to distinguish decisive discovery from measurable but incomplete progress.

2.2 Mathematical Coverage and Significance

OpenProblemBench spans eight primary mathematical domains based on the arXiv mathematics taxonomy. Among its 65 tasks, Graph Theory is the largest domain with 24 tasks, followed by Combinatorics and Number Theory with 15 each. As shown in Fig. 2, the distribution varies across research levels: L1 spans seven domains, L2 emphasizes combinatorial, graph, and optimization problems, while L3 concentrates on Graph Theory and Number Theory, where many open questions can be resolved through finitely verifiable constructions or counterexamples.

Beyond disciplinary coverage, we characterize each task by the mathematical significance of a successful result using five categories of decreasing consequence: **Broad-Scope Resolution**, which could reshape understanding beyond a single subfield; **Major Resolution**, which settles a recognized open problem; **Substantial Advance**, which materially advances the research frontier; **Focused Contribution**, which resolves a narrower self-contained target; and **Instance-Level Result**, which solves a fixed finite instance. Their distribution reflects the distinct roles of the three levels: 9 of 10 L1 tasks are Focused Contributions, L2 spans Instance-Level Results through Substantial Advances, while 34 of 35 L3 tasks fall into the two highest categories. Together, these distributions show how OpenProblemBench spans both mathematical breadth and consequence, from verifiable reconstruction and measurable progress to decisive results on open questions.

2.3 Verifiable Evaluation of Open Problems

Open problems are inherently difficult to evaluate because no known solution is available for direct comparison, while failure to find one provides no evidence of nonexistence. OpenProblemBench addresses this challenge through *result-level verification*, in which the submitted `OUTPUT` and any required certificate are programmatically checked against exact task-specific conditions. The verification mechanism depends on the form of evidence required by each task. As shown in Table 1, **exact-answer verification** applies to L1 rediscovery tasks with known targets; **exact finite-witness verification** validates positive constructions and counterexamples by checking all required constraints; **symbolic coverage verification** checks finite symbolic coverings that establish broader claims; and **proof-certificate verification** validates RUP/LRAT refutations of canonical propositional

encodings (Cruz-Filipe et al., 2017). Together, these mechanisms make open-ended mathematical search exactly and reproducibly evaluable through executable evidence rather than heuristic judgment.

Table 1: **Verification mechanisms across task levels.** Each task is assigned a primary result-level verification method according to the evidence required for acceptance.

Verification method	L1	L2	L3	Total
Exact-answer verification	4	0	0	4
Exact finite-witness verification	6	20	29	55
Symbolic coverage verification	0	0	4	4
Proof-certificate verification	0	0	2	2
All tasks	**10**	**20**	**35**	**65**

3 MathMap Harness: Explore, Verify, and Select

Long-horizon mathematical research requires more than additional attempts: it must explore alternative directions, verify intermediate results, and build on established progress. Motivated by these needs, we introduce **MathMap**, a simple yet effective verifier-grounded harness that combines **diverse exploration**, **external verification**, and **verified selection**. As shown in Fig. 3, each round expands the current verified checkpoint into three independent branches, whose candidates are evaluated by the external verifier. The best verified result is then selected as the next checkpoint, concentrating computation on the strongest verified state rather than continuing every branch. Across rounds, the verified candidates and their parent relations form a **Verified Discovery Tree**, providing a map of explored mathematical directions and verified progress, hence the name **MathMap**.

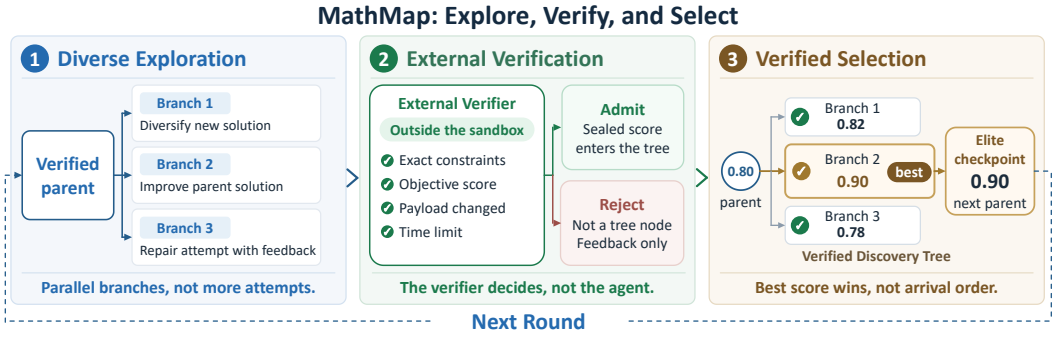


Figure 3: **MathMap organizes long-horizon mathematical research through diverse exploration, external verification, and verified selection.** Three parallel branches explore candidates, and the best verified result becomes the next checkpoint in a Verified Discovery Tree.

Diverse Exploration. From the current verified checkpoint, each round launches three independent tool-using branches. Each branch may search for a new construction, repair a failed candidate using verifier feedback, or explore an alternative direction.

External Verification. MathMap relies on the external verifier rather than agent self-evaluation. Each candidate p is checked for feasibility and, when applicable, assigned the task-native objective $f_V(p)$. Only valid, changed, and in-budget candidates, with finite task-native scores where applicable, are admitted to the Verified Discovery Tree.

Verified Selection. After each round, MathMap selects the best verified candidate together with the current checkpoint: $v_k = \mathrm{best}\,(\{v_{k-1}\} \cup A_k)$, where A_k denotes the verified candidates from round k and the comparison follows the task objective. For L1 and L3, the search terminates once an accepted result is verified; for L2, selection follows the task-native objective as defined above. The current checkpoint is retained unless a verified candidate improves it. Only the selected checkpoint and verified score are carried forward, and the next three branches are expanded from this state. This single-checkpoint continuation concentrates computation on the strongest verified progress rather than extending all branches, keeping MathMap simple, transparent, and cost-aware.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTING

Our experiments address two questions: how much long-horizon interaction improves over direct reasoning, and how discovery systems perform under the same backbone and time budget. Each task–configuration pair is evaluated once, with full implementation details in Appendix A.

Reasoning models vs. long-horizon agents. We first compare six reasoning-only models with five matched open-source models equipped with DeepSeek Harness (DSH) (DeepSeek-AI, 2026), DeepSeek-V4.1-Flash + our MathMap harness, and GPT-5.6-Sol + Codex as a strong proprietary long-horizon reference. This comparison measures the gains from extended interaction beyond direct reasoning across three research levels, while positioning MathMap against both open-source harnessing and a strong proprietary agent.

Progressive discovery systems. We further focus on L2 Progressive Discovery, where open-ended mathematical search is naturally expressed as iterative optimization over feasible constructions. We compare three optimization-oriented discovery systems, SimpleTES (Ye et al., 2026), OpenEvolve (Sharma, 2025), and ShinkaEvolve (Lange et al., 2026), against our MathMap harness. All four use DeepSeek-V4.1-Flash as the backbone with a two-hour budget per task, enabling a controlled comparison of long-horizon search strategies under matched model and wall-clock settings.

4.2 LONG-HORIZON GAINS ACROSS RESEARCH LEVELS

How much does long-horizon interaction improve over direct reasoning? We compare reasoning models with their long-horizon counterparts and MathMap across the three research levels. Tables 2 and 3 report the results on L1 Rediscovery and L2 Progressive Discovery, respectively, while no evaluated configuration produces an accepted result on L3 Decisive Discovery.

Table 2: **L1 Rediscovery results.** Models are ordered by pass rate within each group. Cells are labeled as Passed (passes verification), Failed (parsed output fails verification), Invalid (invalid output format), or None (no submitted solution); the final row reports the pass rate over all 10 tasks.

	Reasoning Models						*Long-Horizon Agents*						*Ours*
L1 Tasks	**GPT-5.6 Sol**	**Qwen3.8 Max**	**DeepSeek V4.1**	**GLM 5.3**	**Kimi K3**	**Qwen3.5 35B**	**GPT-5.6 + Codex**	**DeepSeek + DSH**	**Kimi-K3 + DSH**	**Qwen3.8 + DSH**	**GLM-5.3 + DSH**	**Qwen3.5 + DSH**	**DeepSeek + MathMap**
Focused Contribution													
R01 Erdős #633	Invalid	Invalid	Invalid	Invalid	Invalid	Invalid	Invalid	Failed	Invalid	Invalid	Invalid	Invalid	Invalid
R02 Erdős #1044	**Passed**	**Passed**	**Passed**	Invalid	**Passed**	Invalid	**Passed**	**Passed**	**Passed**	None	None	Invalid	**Passed**
R03 Erdős #1051	**Passed**	**Passed**	Failed	**Passed**	**Passed**	Invalid	**Passed**	Invalid	**Passed**	**Passed**	**Passed**	Invalid	**Passed**
R04 Erdős #1089	**Passed**	**Passed**	**Passed**	**Passed**	**Passed**	Invalid	**Passed**	**Passed**	**Passed**	**Passed**	**Passed**	Invalid	**Passed**
R05 Graffiti-154	**Passed**	**Passed**	**Passed**	**Passed**	Failed	Failed	**Passed**	**Passed**	**Passed**	**Passed**	**Passed**	Invalid	**Passed**
R06 Hadamard-668	Invalid	Invalid	Invalid	Invalid	Invalid	Invalid	None	Invalid	Invalid	Invalid	Invalid	Invalid	Invalid
R07 WOW-284	**Passed**	**Passed**	**Passed**	Invalid	Invalid	Invalid	**Passed**	**Passed**	Invalid	**Passed**	**Passed**	Invalid	**Passed**
R08 Jackson Decomp.	**Passed**	Failed	Failed	Invalid	Failed	Failed	**Passed**	**Passed**	**Passed**	**Passed**	**Passed**	Invalid	**Passed**
R09 Unsplittable Flow	None	None	Failed	Invalid	None	Failed	None	Failed	Invalid	Invalid	Invalid	Invalid	Failed
Substantial Advance													
R10 3D Jacobian	Invalid	Invalid	Failed	Failed	Invalid	Invalid	None	Invalid	Failed	Failed	Failed	Invalid	Failed
Pass rate (%)	**60**	**50**	**40**	**30**	**30**	**0**	**60**	**50**	**50**	**50**	**50**	**0**	**60**

AI capability changes markedly from rediscovery to discovery. Table 2 shows that frontier models can already recover several recently established research results on L1, with GPT-5.6-Sol reaching a 60% pass rate and several open models also succeeding on multiple tasks. On L2, the picture shifts from binary recovery to graded progress: as shown in Table 3, long-horizon configurations frequently obtain stronger feasible constructions and dominate the Rank-1 results. This progression does not yet extend to L3. Across all evaluated L3 submissions, none passes exact verification, so we omit a separate result table. Together, the three levels reveal a clear capability boundary: current AI can reconstruct research-level results and make measurable progress on unresolved problems, but verified resolution of open questions remains substantially harder.

Table 3: **L2 Progressive Discovery results.** Arrows indicate the optimization direction. Rank 1, Rank 2, and Rank 3 denote the top three results, respectively; ties share the same rank and color. Invalid denotes invalid output format, while None denotes no submitted solution. Columns within each group are ordered by the number of Rank-1 results, reported in the final row.

	Reasoning Models						*Long-Horizon Agents*						*Ours*
L2 Tasks	**GPT-5.6 Sol**	**Qwen3.8 Max**	**DeepSeek V4.1**	**GLM 5.3**	**Kimi K3**	**Qwen3.5 35B**	**GPT-5.6 + Codex**	**DeepSeek + DSH**	**GLM-5.3 + DSH**	**Qwen3.8 + DSH**	**Kimi-K3 + DSH**	**Qwen3.5 + DSH**	**DeepSeek + MathMap**
Instance-Level Result													
P01 TSP-512 ↓	28.442	295.427	29.341	Invalid	350.216	350.216	Invalid	**21.254**	**21.254**	21.263	21.728	25.161	**21.254**
P02 Covering Array ↓	220	Invalid	55416	Invalid	585	Invalid	**136**	173	None	1137	202	6342	149
P03 Graceful Tree ↓	7	13	9	Invalid	Invalid	Invalid	**0**	**0**	**0**	3	5	None	**0**
P04 CVRP-100 ↓	20106	23176	20484	20506	69434	41884	18786	**18774**	18858	**18774**	18898	None	18800
P05 RCPSP-60 ↓	8502265	8702216	9202397	9102280	8902378	9202341	8202140	8102180	8202172	8302204	8202150	8602205	**8102179**
P06 Weighted $K_{2,3}$ ↓	1.36864e11	1.36866e11	1.46587e11	Invalid	1.56303e11	1.56306e11	**1.34696e11**	1.34697e11	1.34699e11	1.34702e11	1.34702e11	None	**1.34696e11**
Focused Contribution													
P07 Sidon Set ↑	102	Invalid	71	Invalid	71	Invalid	**105**	**105**	**105**	**105**	**105**	78	**105**
P08 LABS-256 ↓	10880	10880	10880	Invalid	17440	Invalid	**5260**	5732	5728	**5260**	6384	None	**5260**
P09 Golomb Ruler ↓	703	703	Invalid	Invalid	703	Invalid	**623**	**623**	994	1158	639	1158	**623**
P10 Near-Perfect Cuboid ↓	1.108e16	6.867e15	Invalid	Invalid	None	Invalid	1.205e16	1.882e13	**1.769e13**	**1.769e13**	None	None	**1.769e13**
P11 Graph Reconstruction ↑	**6**	2	Invalid	Invalid	Invalid	0	**6**	3	None	None	2	None	2
P12 Union-Closed ↓	42	**38**	41	Invalid	Invalid	Invalid	**38**	41	None	48	None	48	46
P13 Sidorenko Search ↓	6.468e7	1.471e8	9.283e7	Invalid	None	Invalid	**5.626e7**	5.923e7	2.147e9	5.890e7	5.697e7	None	**5.626e7**
Substantial Advance													
P14 MaxDet-23 ↑	0.5475	0.00434	0	Invalid	0.5475	Invalid	**0.9726**	0.8507	0.5475	0.5475	0.5475	None	0.9336
P15 Projective Plane ↓	1413	2198	1570	Invalid	4082	Invalid	**942**	1099	**942**	1256	1099	Invalid	**942**
P16 Three MOLS ↓	Invalid	40	70	Invalid	54	150	29	56	41	None	150	73	**10**
P17 Ramsey (5,5) ↓	583	396	Invalid	1694	Invalid	Invalid	**2**	125	**2**	**2**	207	Invalid	**2**
P18 Schur Coloring ↓	Invalid	Invalid	1093	Invalid	Invalid	Invalid	**2**	**2**	32	16	**2**	Invalid	12
P19 Conway 99 ↓	Invalid	Invalid	Invalid	Invalid	Invalid	Invalid	**902**	1076	1101	1040	7029	1683	1017
P20 W(2,7) ↑	Invalid	155	180	Invalid	Invalid	Invalid	**3702**	615	500	500	600	300	796
# Rank 1	**1**	**1**	**0**	**0**	**0**	**0**	**15**	**6**	**6**	**5**	**2**	**0**	**12**

Long-horizon interaction benefits progressive discovery much more than rediscovery. Across the five matched open-source backbones, adding DSH raises the aggregate L1 pass rate from 30% to 40%. The effect is stronger on L2, where feasible-solution coverage increases from 51% to 82% and long-horizon variants achieve substantially more top-ranked results than their reasoning-only counterparts. This contrast suggests that additional test-time computation is most effective when the task provides a graded objective and repeated verification can guide search, whereas exact rediscovery depends more strongly on the underlying reasoning capability.

Strong backbones remain important even with long-horizon search. The strongest proprietary configuration, GPT-5.6-Sol + Codex, attains 15 Rank-1 results on L2, compared with at most 6 for any open-source model with DSH. Among open models, the larger backbones benefit substantially from harnessing, while Qwen3.5-35B-A3B obtains no Rank-1 result and produces substantially more invalid or missing outputs. These results indicate that long-horizon interaction can amplify a capable backbone, but does not fully compensate for weaker reasoning and constraint handling. The gap between open and proprietary configurations therefore remains visible even as both benefit from additional test-time computation.

MathMap shows that how computation is organized also matters. Using DeepSeek-V4.1-Flash as its backbone, MathMap reaches a 60% L1 pass rate, matching GPT-5.6-Sol and GPT-5.6-Sol + Codex, and improving over the 50% obtained by DeepSeek-V4.1-Flash with DSH. The difference is larger on L2: MathMap achieves Rank 1 on 12 of 20 tasks in the broad comparison of Table 3, twice the count of DeepSeek-V4.1-Flash + DSH and second only to GPT-5.6-Sol + Codex in this comparison. Thus, stronger long-horizon performance is not explained by additional computation alone. Organizing search around diverse exploration, external verification, and verified selection can convert the same backbone and time budget into substantially stronger progressive discovery.

4.3 Progressive Discovery Under Matched Backbone and Time Budget

How do discovery systems compare under the same backbone and time budget? We compare SimpleTES, OpenEvolve, ShinkaEvolve, and MathMap on all 20 L2 Progressive Discovery tasks using DeepSeek-V4.1-Flash with a two-hour budget per task. Table 4 shows that MathMap achieves the broadest coverage of top-performing results, while Table 5 shows that it does so with substantially fewer completion tokens and solution submissions.

Table 4: **L2 Progressive Discovery system comparison.** All methods use DeepSeek-V4.1-Flash with the same two-hour budget. Red cells denote Rank 1 among the four systems; ties share Rank 1. Invalid denotes invalid output format. The final row reports the number of Rank-1 results.

	Discovery Systems			*Ours*
L2 Tasks	**SimpleTES**	**ShinkaEvolve**	**OpenEvolve**	**MathMap**
Instance-Level Result				
P01 TSP-512 ↓	21.256	**21.254**	21.289	**21.254**
P02 Covering Array ↓	**130**	168	159	149
P03 Graceful Tree ↓	1	**0**	3	**0**
P04 CVRP-100 ↓	18804	18878	19240	**18800**
P05 RCPSP-60 ↓	8202150	8202140	8202150	**8102179**
P06 Weighted $K_{2,3}$ ↓	**1.34696e11**	1.34700e11	1.34700e11	**1.34696e11**
Focused Contribution				
P07 Sidon Set ↑	**105**	103	**105**	**105**
P08 LABS-256 ↓	5312	5728	5728	**5260**
P09 Golomb Ruler ↓	**623**	**623**	**623**	**623**
P10 Near-Perfect Cuboid ↓	**1.769e13**	Invalid	**1.769e13**	**1.769e13**
P11 Graph Reconstruction ↑	**4**	2	2	2
P12 Union-Closed ↓	Invalid	56	48	**46**
P13 Sidorenko Search ↓	5.668e7	5.824e7	5.732e7	**5.626e7**
Substantial Advance				
P14 MaxDet-23 ↑	**0.9450**	0.8151	0.9336	0.9336
P15 Projective Plane ↓	1015	1413	1256	**942**
P16 Three MOLS ↓	**10**	**10**	42	**10**
P17 Ramsey (5,5) ↓	**2**	140	178	**2**
P18 Schur Coloring ↓	**4**	23	6	12
P19 Conway 99 ↓	**834**	1010	1188	1017
P20 $W(2,7)$ ↑	**3703**	541	3702	796
# Rank 1	**12**	**4**	**3**	**14**

MathMap delivers the strongest overall progressive-discovery performance. Within the same-backbone system comparison, MathMap attains Rank 1 on 14 of 20 tasks, compared with 12 for SimpleTES, 4 for ShinkaEvolve, and 3 for OpenEvolve; ties arise when objectives are discrete or systems reach the same verified value. On projective-plane search, for example, MathMap achieves the best observed incidence defect of 942, improving over 1,015 for SimpleTES, 1,256 for OpenEvolve, and 1,413 for ShinkaEvolve. This broad performance aligns with its design: MathMap explores parallel tool-using branches, externally verifies intermediate results, and continues from the strongest verified checkpoint, whereas SimpleTES follows a continuous reasoning trajectory with verifier feedback and OpenEvolve and ShinkaEvolve rely on evolutionary candidate updates.

MathMap converts computation into progress more efficiently. It uses only 2.08M completion tokens per task on average, about 29% of SimpleTES's 7.14M and 30% of OpenEvolve's 6.88M, while achieving the most Rank-1 results. MathMap also submits only 15 solutions per task, compared with 120 for SimpleTES, 72 for ShinkaEvolve, and 26 for OpenEvolve. Its tool-using branches can construct, inspect, and repair candidates before submission, consistent with the substantially lower submission count. These results suggest that verifier-grounded, tool-assisted search can improve both discovery quality and test-time efficiency.

Table 5: **Average interaction statistics on L2 Progressive Discovery over 20 tasks.**

System	Interaction Steps	Submitted Solutions	Valid Solutions	Completion Tokens
SimpleTES	188	120	109	7,137,609
ShinkaEvolve	2,570	72	33	9,263,893
OpenEvolve	**112**	26	18	6,877,943
MathMap	423	**15**	**15**	**2,079,752**

4.4 Long-Horizon Dynamics of Verified Progress

L2 makes long-horizon mathematical progress directly observable. The trajectories in Fig. 4 show when useful constructions emerge and how subsequent computation improves them. With an external verifier and task-native metric, long-horizon agents can repeatedly refine verified solutions, but at markedly different rates, showing that more computation does not necessarily translate into faster frontier progress.

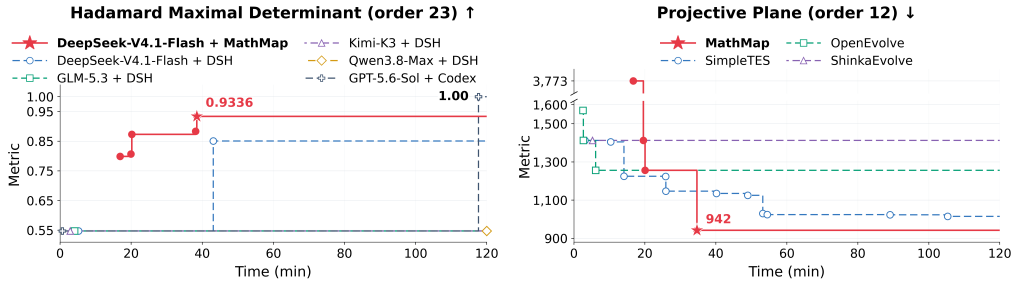


Figure 4: **Long-horizon dynamics of verified progress.** Hadamard Maximal Determinant (order 23) compares long-horizon agents across backbones, while Projective Plane (order 12) compares same-backbone discovery systems. Each staircase tracks the best verified solution over time, revealing how differently systems convert additional computation into verified progress.

MathMap's explore–verify–select loop turns verifier feedback into stronger progress than open-source DSH. As shown in Fig. 4 (left), on Hadamard Maximal Determinant (order 23), MathMap starts at 0.7988 and improves to 0.8732 at 20 minutes and 0.9336 at 39 minutes. By contrast, DeepSeek-V4.1-Flash + DSH reaches only 0.8507 at 43 minutes, while the other open-source DSH variants show little improvement. The explore–verify–select loop therefore extracts substantially more progress than the vanilla DSH setup with the same backbone. GPT-5.6-Sol + Codex eventually reaches 1.0 near the two-hour limit, indicating further headroom from stronger proprietary agents.

Early progress does not necessarily translate into stronger long-horizon improvement. As shown in Fig. 4 (right), on Projective Plane (order 12), SimpleTES obtains a strong verified solution within the first few minutes but improves only gradually thereafter, reaching a final metric 1,015. MathMap takes longer to obtain its first verified solution, but then improves rapidly from 3,773 to 942 within 35 minutes, surpassing the two-hour endpoints of all three discovery systems. This contrast shows that first-solution quality and subsequent improvement rate capture different aspects of long-horizon search. L2 therefore makes both final quality and the dynamics of verifiable progress measurable.

5 Conclusion and Future Work

OpenProblemBench provides a result-verifiable platform for measuring AI progress on open mathematical problems across three capability levels: **L1 Rediscovery**, **L2 Progressive Discovery**, and **L3 Decisive Discovery**. They distinguish recovering established results, making measurable progress on unresolved objectives, and resolving specified open questions. Across eight mathematical domains, the benchmark combines task-native objectives, explicit acceptance criteria, and executable verification, making intermediate progress measurable without conflating it with decisive discovery. Together with an open evaluation protocol and leaderboard, it provides a common platform for tracking reasoning models, long-horizon agents, and discovery systems as the frontier advances.

Our evaluation reveals a clear capability transition across the three levels. Frontier models can already rediscover several recent research results, while long-horizon interaction yields much larger gains on L2 Progressive Discovery; however, no evaluated configuration reaches L3 Decisive Discovery. **MathMap** further shows that a simple combination of diverse exploration, external verification, and verified selection can convert additional test-time computation into stronger progress with competitive efficiency. Its time-stamped trajectories also separate early solution quality, sustained refinement, and final performance, showing that endpoint scores alone cannot fully characterize long-horizon research behavior. Together, these results suggest that effective use of verifiable intermediate progress is a key step toward stronger mathematical discovery.

These findings also expose the next gap: verifying mathematical outcomes is only a first step toward genuine discovery. Future work will focus on four directions: **(1)** expanding to broader domains and more open problems with finitely verifiable positive constructions or counterexamples; **(2)** extending result-level verification to step-level evaluation of intermediate reasoning; **(3)** assessing the novelty of accepted results to distinguish independent rediscovery from genuinely new mathematical contributions; and **(4)** developing long-horizon systems that can move beyond progressive improvement and cross into L3 Decisive Discovery, producing verified new results on genuinely open mathematics.

AI USE STATEMENT

In this work, we used generative AI tools to polish the manuscript, improve language clarity and presentation, and assist in developing verification code for the benchmark. AI-generated suggestions were used only as supporting material and were reviewed by the authors before being incorporated into the paper or codebase. We manually checked all AI-assisted manuscript revisions for accuracy and consistency with our intended claims, and the verification code was inspected and tested against the task specifications and expected outputs. We have reviewed all AI-assisted work and take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies AI systems on mathematical research problems and does not involve human subjects, personal data, or sensitive user information. The benchmark is constructed from publicly documented mathematical problems and research results, with task definitions and evaluation criteria designed to preserve the scope of the underlying mathematical claims. Because results on open problems may be interpreted as new mathematical discoveries, we take particular care to distinguish Rediscovery, Progressive Discovery, and Decisive Discovery, and to avoid overstating the implications of benchmark outcomes. Executable verification establishes only that a submitted result satisfies the specified task conditions; claims of historical novelty or broader mathematical significance require separate literature review and independent expert verification. We will release the benchmark, evaluation protocol, and supporting materials to encourage transparent comparison and responsible validation of future results.

REPRODUCIBILITY STATEMENT

We provide comprehensive information to support reproduction of both the benchmark and experimental results. The main paper describes the three evaluation levels, task objectives, verification protocol, MathMap design, and experimental settings. The appendix documents the complete task inventory, model and harness configurations, evaluation budgets, and additional implementation details. The supplementary materials provide the task definitions used in the benchmark. All reported endpoint values and time-stamped trajectories are derived from recorded experimental outputs; missing intermediate checkpoints are not synthetically reconstructed.

REFERENCES

Noga Alon, Thomas F. Bloom, W. T. Gowers, Daniel Litt, Will Sawin, Arul Shankar, Jacob Tsimerman, Victor Wang, and Melanie Matchett Wood. Remarks on the disproof of the unit distance conjecture. *arXiv preprint arXiv:2605.20695*, 2026. URL https://arxiv.org/abs/2605.20695.

Eiichi Bannai, Etsuko Bannai, and Dennis Stanton. An upper bound for the cardinality of an s-distance subset in real Euclidean space, II. *Combinatorica*, 3(2):147–152, 1983. doi: 10.1007/BF02579288. URL https://doi.org/10.1007/BF02579288.

Kevin Barreto, Jiwon Kang, Sang-hyun Kim, Vjekoslav Kovač, and Shengtong Zhang. Irrationality of rapidly converging series: a problem of Erdős and Graham. *arXiv preprint arXiv:2601.21442*, 2026. URL https://arxiv.org/abs/2601.21442.

Michael Beeson, Miklós Laczkovich, and Yan X. Zhang. Solution of Erdős problem 633. *arXiv preprint arXiv:2604.03609*, 2026. URL https://arxiv.org/abs/2604.03609.

Federico Bianchi, Yongchan Kwon, Aneesh Pappu, and James Zou. Harnessing the collective intelligence of AI agents in the wild for new discoveries. *arXiv preprint arXiv:2606.10402*, 2026. URL https://arxiv.org/abs/2606.10402.

Thomas F. Bloom. Erdős problem #203. Erdős Problems, online problem record, n.d.a. URL https://www.erdosproblems.com/203. Accessed 2026-09-26.

Thomas F. Bloom. Erdős problem #273. Erdős Problems, online problem record, n.d.b. URL https://www.erdosproblems.com/273. Accessed 2026-09-26.

Thomas F. Bloom. Erdős problem #274. Erdős Problems, online problem record, n.d.c. URL https://www.erdosproblems.com/274. Accessed 2026-09-26.

Henning Bruhn and Oliver Schaudt. The journey of the union-closed sets conjecture. *Graphs and Combinatorics*, 31(6):2043–2074, 2015. doi: 10.1007/s00373-014-1515-0.

Mason R. Calbert and Ryan E. Dougherty. Upper bounds for covering arrays of higher index. *arXiv preprint arXiv:2211.01209*, 2022. URL https://arxiv.org/abs/2211.01209.

Patrick G. Cesarz and Andrew J. Woldar. On the automorphism group of a putative Conway 99-graph. *Algebraic Combinatorics*, 8(2):379–398, 2025. doi: 10.5802/alco.418. URL https://alco.centre-mersenne.org/articles/10.5802/alco.418/.

Xujin Chen, Zhuo Diao, Xiaodong Hu, and Zhongzheng Tang. Sufficient conditions for Tuza's conjecture on packing and covering triangles. *arXiv preprint arXiv:1605.01816*, 2016. URL https://arxiv.org/abs/1605.01816.

Luís Cruz-Filipe, Marijn JH Heule, Warren A Hunt Jr, Matt Kaufmann, and Peter Schneider-Kamp. Efficient certified RAT verification. In Leonardo de Moura (ed.), *Automated Deduction – CADE 26*, volume 10395 of *Lecture Notes in Computer Science*, pp. 220–236. Springer, 2017. doi: 10.1007/978-3-319-63046-5_14. URL https://arxiv.org/abs/1612.02353.

Cristina Dalfó. A survey on the missing Moore graph. *Linear Algebra and its Applications*, 569: 1–14, 2019. doi: 10.1016/j.laa.2018.12.035.

G. B. Dantzig and J. H. Ramser. The truck dispatching problem. *Management Science*, 6(1):80–91, 1959. doi: 10.1287/mnsc.6.1.80. URL https://doi.org/10.1287/mnsc.6.1.80.

George B. Dantzig, D. Ray Fulkerson, and Selmer M. Johnson. Solution of a large-scale traveling-salesman problem. *Journal of the Operations Research Society of America*, 2(4):393–410, 1954. doi: 10.1287/opre.2.4.393.

DeepSeek-AI. DeepSeek Harness: Everything is a plugin. https://github.com/deepseek-ai/deepseek-harness, 2026.

Peter F. Dembowski. *Finite Geometries*. 1968.

Demonstrandum Research. Graffiti 154: exact counterexamples and verification artifacts. Research artifact repository, 2026. URL https://github.com/demonstrandum-research/artifacts/tree/main/problems/p2-factory/kills/graffiti-154.

Konstantinos Drakakis. A review of the available construction methods for Golomb rulers. *Advances in Mathematics of Communications*, 3(3):235–250, 2009. doi: 10.3934/amc.2009.3.235. URL https://doi.org/10.3934/amc.2009.3.235.

Judith Egan and Ian M. Wanless. Enumeration of MOLS of small order. *Mathematics of Computation*, 85:799–824, 2016. URL https://arxiv.org/abs/1406.3681.

Epoch AI. FrontierMath: Open problems. Online resource, 2026a. URL https://epoch.ai/frontiermath/open-problems. Accessed 2026-09-26.

Epoch AI. Hadamard matrix of order 668, 2026b. URL https://epoch.ai/frontiermath/open-problems/hadamard. Accessed 2026-09-26.

Louis Esperet and Giuseppe Mazzuoccolo. On the maximum fraction of edges covered by t perfect matchings in a cubic bridgeless graph. *arXiv preprint arXiv:1306.2828*, 2013. URL https://arxiv.org/abs/1306.2828.

Nevena Francetić, Sarada Herke, Brendan D. McKay, and Ian M. Wanless. On Ryser's conjecture for linear intersecting multipartite hypergraphs. *European Journal of Combinatorics*, 61:91–105, 2017. doi: 10.1016/j.ejc.2016.10.004. URL https://arxiv.org/abs/1508.00951.

Joseph A. Gallian. A dynamic survey of graph labeling. The Electronic Journal of Combinatorics, Dynamic Survey DS6, 2024. URL https://www.combinatorics.org/files/Surveys/ds6/ds6v27-2024.pdf. Version 27, 15 November 2024.

Shuhong Gao. Counterexamples to the Jacobian conjecture in dimensions greater than two. *arXiv preprint arXiv:2608.00222*, 2026. URL https://arxiv.org/abs/2608.00222.

Daniel M. Gordon. The prime power conjecture is true for $n < 2,000,000$. *The Electronic Journal of Combinatorics*, 1(1):R6, 1994. doi: 10.37236/1186. URL https://doi.org/10.37236/1186.

Bertille Granet. Hamilton decompositions of regular bipartite tournaments. *arXiv preprint arXiv:2209.02988*, 2022. URL https://arxiv.org/abs/2209.02988.

Richard K. Guy. *Unsolved Problems in Number Theory*. Springer, New York, 3 edition, 2004. doi: 10.1007/978-0-387-26677-0.

Anand Shripad Hegde, R. B. Sandeep, and P. Shashank. Erdős–Gyárfás conjecture on graphs without long induced paths. *arXiv preprint arXiv:2410.22842*, 2024. URL https://arxiv.org/abs/2410.22842.

Owen Henderschedt, Jessica McDonald, and Songling Shan. Total coloring graphs with large minimum degree. *arXiv preprint arXiv:2507.05548*, 2025. URL https://arxiv.org/abs/2507.05548.

Marijn J. H. Heule. Schur number five. *arXiv preprint arXiv:1711.08076*, 2017. URL https://arxiv.org/abs/1711.08076.

D Iannucci and R Sorli. On the total number of prime factors of an odd perfect number. *Mathematics of computation*, 72(244):2077–2084, 2003.

Rainer Kolisch. Serial and parallel resource-constrained project scheduling methods revisited: Theory and computation. *European Journal of Operational Research*, 90(2):320–333, 1996. doi: 10.1016/0377-2217(95)00357-6. URL https://doi.org/10.1016/0377-2217(95)00357-6.

Jeffrey C. Lagarias. The 3x+13x+1 problem and its generalizations. *The American Mathematical Monthly*, 92(1):3–23, 1985. doi: 10.1080/00029890.1985.11971528.

Robert Tjarko Lange, Yuki Imajuku, and Edoardo Cetin. ShinkaEvolve: Towards open-ended and sample-efficient program evolution. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/7886b9bafe76c52fd568db10ff9772df-Abstract-Conference.html. ICLR 2026.

Junlin Liu, Shengnan An, Shuang Zhou, Dan Ma, Yehao Lin, Xinxuan Lv, Xuanlin Wang, Xiaoyu Li, Ziwen Wang, Xuezhi Cao, and Xunliang Cai. AMO-bench: Large language models still struggle in high school math competitions. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 2120–2137, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.101. URL https://aclanthology.org/2026.findings-acl.101/.

Thang Luong, Dawsen Hwang, Hoang H. Nguyen, Golnaz Ghiasi, Yuri Chervonyi, Insuk Seo, Junsu Kim, Garrett Bingham, Jonathan Lee, Swaroop Mishra, Alex Zhai, Huiyi Hu, Henryk Michalewski, Jimin Kim, Jeonghyun Ahn, Junhwi Bae, Xingyou Song, Trieu Hoang Trinh, Quoc V. Le, and Junehyuk Jung. Towards robust mathematical reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 35418–35442, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1794. URL https://aclanthology.org/2025.emnlp-main.1794/.

R. Daniel Mauldin. The Beal conjecture and prize. Author's problem statement, University of North Texas, n.d. URL https://sites.itservices.cas.unt.edu/~mauldin/beal1198.html. Accessed 2026-09-26.

Lech Mazur. A computer-assisted proof of Sendov's conjecture, 2026. URL https://www.proofatlas.ai/formalizations/sendov-conjecture/. Public version of 5 August 2026.

Brendan D. McKay. Reconstruction of small graphs and digraphs. *arXiv preprint arXiv:2102.01942*, 2021. URL https://arxiv.org/abs/2102.01942.

Richard Montgomery. Transversals in latin squares. *arXiv preprint arXiv:2406.19873*, 2024.

Petra Mutzel and Lutz Oettershagen. The crossing number of seq-shellable drawings of complete graphs. *arXiv preprint arXiv:1803.07515*, 2018. URL https://arxiv.org/abs/1803.07515.

Kevin O'Bryant. A complete annotated bibliography of work related to Sidon sequences. *The Electronic Journal of Combinatorics*, pp. DS11, 2004. doi: 10.37236/32. URL https://doi.org/10.37236/32.

Tomás Oliveira e Silva, Siegfried Herzog, and Silvio Pardi. Empirical verification of the even Goldbach conjecture and computation of prime gaps up to $4 \cdot 10^{18}$. *Mathematics of Computation*, 83(288):2033–2060, 2014. doi: 10.1090/S0025-5718-2013-02787-1. URL https://doi.org/10.1090/S0025-5718-2013-02787-1.

OpenAI. On the Navier–Stokes millennium prize problem. Research publication, September 2026. URL https://openai.com/index/navier-stokes-solution/.

William P Orrick, Bruce Solomon, Roland Dowdeswell, and Warren D Smith. New lower bounds for the maximal determinant problem. *arXiv preprint math/0304410*, 2003.

Tom Packebusch and Stephan Mertens. Low autocorrelation binary sequences. *Journal of Physics A: Mathematical and Theoretical*, 49(16):165001, 2016. doi: 10.1088/1751-8113/49/16/165001.

R. E. A. C. Paley. On orthogonal matrices. *Journal of Mathematics and Physics*, 12(1–4): 311–320, 1933. doi: 10.1002/sapm1933121311. URL https://doi.org/10.1002/sapm1933121311.

Samuil Petkov. Counterexamples, spectral obstructions, and deletion stability for WOW-284. *arXiv preprint arXiv:2607.27452*, 2026. URL https://arxiv.org/abs/2607.27452.

John Rabung and Mark Lotts. Improving the use of cyclic zippers in finding lower bounds for van der waerden numbers. *the electronic journal of combinatorics*, pp. P35–P35, 2012.

Stanisław Radziszowski. Small Ramsey numbers. *The Electronic Journal of Combinatorics*, pp. DS1, 1994. doi: 10.37236/21. URL https://doi.org/10.37236/21. Dynamic survey.

Arthur Freitas Ramos, David Barros Hulak, and Ruy Jose Guerra Barretto de Queiroz. A formal counterexample to the cost-preserving single-source unsplittable flow conjecture. *Archive of Formal Proofs*, July 2026. URL https://isa-afp.org/entries/Dinitz_Garg_Goemans_Counterexample.html.

Matthieu Rosenfeld. The lonely runner conjecture holds for nine runners. *arXiv preprint arXiv:2512.01912*, 2025. URL https://arxiv.org/abs/2512.01912.

Serge E. Salez. The Erdős–Straus conjecture: New modular equations and checking up to $n = 10^{17}$. *arXiv preprint arXiv:1406.6307*, 2014. URL https://arxiv.org/abs/1406.6307.

Lisa Sauermann. Rota's basis conjecture holds for random bases of vector spaces. *arXiv preprint arXiv:2203.17121*, 2022. URL https://arxiv.org/abs/2203.17121.

Tobias Schnieders. Barnette graphs with faces up to size 8 are Hamiltonian. *arXiv preprint arXiv:2508.03531*, 2025. URL https://arxiv.org/abs/2508.03531.

Tyler Seacrest. Seymour's second neighborhood conjecture for subsets of vertices. *arXiv preprint arXiv:1808.06293*, 2018. URL https://arxiv.org/abs/1808.06293.

Paul Seymour. Hadwiger's conjecture. In Jr. Nash, John Forbes and Michael Th. Rassias (eds.), *Open Problems in Mathematics*, pp. 417–437. Springer, 2016. doi: 10.1007/978-3-319-32162-2_13.

Asankhaya Sharma. OpenEvolve: an open-source evolutionary coding agent, 2025. URL https://github.com/algorithmicsuperintelligence/openevolve. Software repository.

Jian Shen. On the girth of digraphs. *Discrete Mathematics*, 211(1–3):167–181, 2000. doi: 10.1016/S0012-365X(99)00323-4. URL https://doi.org/10.1016/S0012-365X(99)00323-4.

A. F. Sidorenko. Extremal problems for k-colour graphs and exact inequalities for pairs of random elements. *Discrete Mathematics and Applications*, 1(3):271–278, 1991. doi: 10.1515/dma.1991.1.3.271. URL https://doi.org/10.1515/dma.1991.1.3.271.

Jonathan Sorenson and Jonathan Webster. An algorithm and computation to verify legendre's conjecture up to 7· 10 13. *Research in Number Theory*, 11(1):4, 2025.

Eckhard Steffen. Tutte's 5-flow conjecture for highly cyclically connected cubic graphs. *arXiv preprint arXiv:math/0607077*, 2006. URL https://arxiv.org/abs/math/0607077.

Haoxiang Sun, Yingqian Min, Zhipeng Chen, Xin Zhao, and Ji-Rong Wen. Challenging the boundaries of reasoning: An olympiad-level math benchmark for large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17438–17457, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.792. URL https://aclanthology.org/2026.acl-long.792/.

Quanyu Tang. On Erdős problem #1044. Author's research note and source repository, 2026. URL https://github.com/QuanyuTang/erdos-problem-1044.

Terence Tao. Mathematics in the age of AI. *arXiv preprint arXiv:2608.16753*, 2026. URL https://arxiv.org/abs/2608.16753.

Erik Y. Wang, Sumeet Motwani, James V. Roggeveen, Eliot Hodges, Dulhan Jayalath, Charles London, Kalyan Ramakrishnan, Flaviu Cipcigan, Philip Torr, and Alessandro Abate. HorizonMath: Measuring AI progress toward mathematical discovery with automatic verification. *arXiv preprint arXiv:2603.15617*, 2026. URL https://arxiv.org/abs/2603.15617.

Haotian Ye, Haowei Lin, Jingyi Tang, Yizhen Luo, Rahul Thapa, Caiyin Yang, Chang Su, Rui Yang, Ruihua Liu, Rundao Li, Zeyu Li, Pengwei Sun, Chong Gao, Dachao Ding, Guangrong He, Miaolei Zhang, Lina Sun, Wenyang Wang, Yuchen Zhong, Zhuohao Shen, Puheng Li, Pan Lu, Bianxiao Cui, Di He, Jianzhu Ma, Junfeng Li, Hexi Baoyin, Yejin Choi, Stefano Ermon, Xiaowen Chu, Tongyang Li, Yuzhi Xu, and James Zou. Structured scaling of AI discovery across diverse scientific domains. *arXiv preprint arXiv:2604.19341*, 2026. URL https://arxiv.org/abs/2604.19341. Version 2; revised 27 July 2026.

A Implementation Details

This section summarizes the implementation settings used throughout our experiments. We first describe the inference configuration for **reasoning-only models**, followed by the setup for **long-horizon agents** based on DeepSeek Harness. We then specify the matched-backbone configuration for the **Progressive Discovery systems**, and finally report the configuration of our **MathMap** harness. Unless otherwise stated, each task–configuration pair is evaluated once under the settings below.

A.1 Reasoning-Only Inference Settings

All reasoning-only models use temperature 1.0 and top-p 0.95. For each API, we set `max_tokens` to the maximum supported value: `131072` for GLM-5.3 and Qwen3.8-Max-0902, `393216` for DeepSeek-V4.1-Flash, and the corresponding API maximum for the remaining models. We also use the highest reasoning-effort setting available from each API. GLM-5.3, Kimi-K3, and DeepSeek-V4.1-Flash use `max`, while Qwen3.8-Max-0902, Qwen3.5-35B-A3B, and GPT-5.6-Sol use `xhigh`. These settings are used to expose the strongest available direct-reasoning capability of each model without long-horizon scaffolding.

A.2 Long-Horizon Agent Settings

The five matched open-source long-horizon configurations use the native DeepSeek Harness (DSH) (DeepSeek-AI, 2026) with its default tool set and no Internet access. For each backbone, the decoding configuration is aligned with its reasoning-only counterpart: temperature 1.0, top-p 0.95, the maximum generation length supported by the API, and the same highest available reasoning-effort setting. This keeps the underlying model configuration fixed while adding long-horizon interaction and tool use through DSH.

A.3 Progressive Discovery System Settings

For the controlled L2 Progressive Discovery comparison, SimpleTES (Ye et al., 2026), OpenEvolve (Sharma, 2025), and ShinkaEvolve (Lange et al., 2026) all use DeepSeek-V4.1-Flash as the backbone and receive an independent two-hour wall-clock budget per task. No reference solution or seed program is provided, and all submitted solutions are evaluated by the same task verifier. Native iteration or generation limits are left effectively unbounded, with the external wall-clock budget determining termination.

SimpleTES uses its upstream search configuration with `num_chains=4`, `k_candidates=4`, reflection enabled, and five inspirations. Request concurrency is set to four, and the checkpoint interval is reduced to 25 evaluations so that intermediate checkpoints are recorded within the two-hour budget.

OpenEvolve uses its factory-default configuration with five islands, population size 1000, archive size 100, and `parallel_evaluations=4`. The number of iterations is unbounded and termination is determined by the wall-clock budget.

ShinkaEvolve reports task-specific rather than global hyperparameters. We adopt the Circle Packing configuration, using two islands, archive size 40, `patch_type_probs=[0.45, 0.45, 0.1]`, `max_patch_attempts=3`, `max_patch_resamples=3`, four archive inspirations, two top-k inspirations, temperatures $[0.0, 0.5, 1.0]$, meta-recommendation interval 10, migration interval 10 with migration rate 0.0, and elite selection ratio 0.3. We set `max_evaluation_jobs=5` and `max_proposal_jobs=6` for intra-task concurrency. Because all experiments use a single backbone, its model-selection mechanism reduces to a single available model. The `embedding_model` is set to `null` because our gateway provides no embedding endpoint.

The systems retain their native search structures and concurrency settings rather than forcing these parameters to be identical. Accordingly, they may consume different numbers of model calls and tokens within the same wall-clock budget; we therefore additionally report interaction steps, solution submissions, valid submissions, and completion-token usage. Interaction steps include both reasoning and tool-use turns; for systems without tool calls, they reduce to assistant turns.

A.4 MATHMAP CONFIGURATION

MathMap uses DeepSeek-V4.1-Flash as the backbone and is built on the native DeepSeek Harness tool-use environment. Its search follows an *explore–verify–select* loop that forms a persistent **Verified Discovery Tree**. Starting from the current verified parent, each round launches three independent branches in parallel. Each branch is a DSH-based tool-using session that may explore a new construction, repair the current candidate, or pursue an alternative optimization direction.

Candidate solutions are evaluated by the external OpenProblemBench verifier rather than by the agent itself. The verifier determines feasibility and returns the task-native metric when applicable. After the three branches complete, MathMap compares the verified candidates with the current parent and retains the best result according to the task objective. The parent is therefore preserved unless a verified candidate improves it. Three new branches are then expanded from the selected checkpoint, yielding a repeated $1 \to 3 \to 1$ search pattern. For L1 Rediscovery, MathMap terminates once an accepted result is verified; for L2 Progressive Discovery, checkpoint selection follows the task-native objective described above.

This design concentrates subsequent computation on the strongest verified state while retaining parallel exploration at each round. The parent–child relations among verified candidates form the Verified Discovery Tree used to track long-horizon progress. MathMap uses the same two-hour per-task wall-clock budget as the other discovery systems in the controlled L2 comparison.

B Detailed Descriptions of Problems

Table 6 lists all 65 tasks across the three research levels, with links to their corresponding descriptions.

Table 6: **Complete list of problems in OpenProblemBench across the three research levels.** Problem titles link to their corresponding descriptions, and significance indicates the benchmark's qualitative assessment of each problem's mathematical consequence.

Task	Problem Title	Significance
L1-Rediscovery		
R01	Non-Square Congruent Tilings of Triangles (Erdős #633)	Focused Contribution
R02	Polynomial Lemniscate Component Perimeter (Erdős #1044)	Focused Contribution
R03	Critical Growth Base for an Irrational Series (Erdős #1051)	Focused Contribution
R04	High-Dimensional Distinct-Distance Limit (Erdős #1089)	Focused Contribution
R05	Graffiti 154 Spectral Counterexample	Focused Contribution
R06	Hadamard Matrix of Order 668	Focused Contribution
R07	WOW-284 Dual-Degree Counterexample	Focused Contribution
R08	Jackson Hamilton-Decomposition Counterexample	Focused Contribution
R09	Cost-Preserving Unsplittable-Flow Counterexample	Focused Contribution
R10	Three-Dimensional Jacobian Conjecture Counterexample	Substantial Advance
L2-Progressive Discovery		
P01	Manhattan Traveling Salesperson Problem with 512 Cities	Instance-Level Result
P02	Ternary Covering Array with 220 Rows	Instance-Level Result
P03	Graceful Labeling of a Fixed 127-Vertex Tree	Instance-Level Result
P04	Capacitated Vehicle Routing with 100 Customers	Instance-Level Result
P05	Resource-Constrained Project Scheduling with 60 Jobs	Instance-Level Result
P06	Weighted $K_{2,3}$ Homomorphism Minimization	Instance-Level Result
P07	Sidon Sets in an Interval of Length 10,000	Focused Contribution
P08	Low-Autocorrelation Binary Sequences of Length 256	Focused Contribution
P09	Golomb Rulers with 29 Marks	Focused Contribution
P10	Primitive Near-Perfect Cuboids in a Dyadic Range	Focused Contribution
P11	Vertex-Deleted Deck Overlap for 14-Vertex Graphs	Focused Contribution
P12	Union-Closed Frequency Minimization on 13 Elements	Focused Contribution
P13	Sidorenko Homomorphism Search on a 16-Vertex Host	Focused Contribution
P14	Hadamard Maximal Determinant (order 23)	Substantial Advance
P15	Projective Plane (order 12)	Substantial Advance
P16	Three Mutually Orthogonal Latin Squares of Order 10	Substantial Advance
P17	Ramsey $(5, 5)$ Colorings on 43 Vertices	Substantial Advance
P18	Schur Six-Coloring on 537 Integers	Substantial Advance
P19	Conway 99-Graph Common-Neighbor Defect	Substantial Advance
P20	Two-Colorings without Seven-Term Arithmetic Progressions	Substantial Advance

Continued on the next page

Table 6 (continued): Complete list of OpenProblemBench tasks

Task	Problem Title	Significance
L3-Decisive Discovery		
D01	Hadamard Matrix of Order 2092	Substantial Advance
D02	Perfect Cuboid Problem	Major Resolution
D03	Lander–Parkin–Selfridge Conjecture Counterexample	Major Resolution
D04	Lehmer's Totient Problem	Major Resolution
D05	Erdős–Straus Conjecture Counterexample	Major Resolution
D06	Beal Conjecture Counterexample	Major Resolution
D07	Legendre Conjecture Counterexample by Composite Coverage	Major Resolution
D08	Odd Perfect Number	Major Resolution
D09	Tuza Triangle Packing–Covering Counterexample	Major Resolution
D10	Conway 99-Graph	Major Resolution
D11	Erdős–Gyárfás Cycle-Length Counterexample	Major Resolution
D12	Barnette Hamiltonicity Counterexample	Major Resolution
D13	Moore Graph of Degree 57	Major Resolution
D14	Caccetta–Häggkvist Counterexample	Major Resolution
D15	Seymour Second-Neighborhood Counterexample	Major Resolution
D16	Harary–Hill Counterexample via a Straight-Line Drawing	Major Resolution
D17	Graph Reconstruction Counterexample	Major Resolution
D18	Graceful-Tree Conjecture Counterexample	Major Resolution
D19	Total-Coloring Conjecture Counterexample	Major Resolution
D20	Tutte Five-Flow Conjecture Counterexample	Major Resolution
D21	Berge–Fulkerson Counterexample	Major Resolution
D22	Perfect Difference Set of Non-Prime-Power Order	Major Resolution
D23	Ryser Latin-Transversal Counterexample	Major Resolution
D24	Union-Closed-Sets Conjecture Counterexample	Major Resolution
D25	Lonely-Runner Counterexample with Integer Relative Speeds	Major Resolution
D26	Ryser Hypergraph Counterexample	Major Resolution
D27	Projective Plane of Non-Prime-Power Order	Major Resolution
D28	Rota Basis-Rearrangement Counterexample	Major Resolution
D29	Two-Prime Covering Construction (Erdős #203)	Major Resolution
D30	Prime-Minus-One Covering System (Erdős #273)	Major Resolution
D31	Exact Coset Cover with Unequal Sizes (Erdős #274)	Major Resolution
D32	Nontrivial Collatz Cycle	Broad-Scope Resolution
D33	Strong Goldbach Counterexample by Symbolic Coverage	Broad-Scope Resolution
D34	Hadwiger Conjecture Counterexample	Broad-Scope Resolution
D35	Sidorenko Counterexample via a Rational Step Graphon	Broad-Scope Resolution

Reading the task descriptions. Each card separates context, question, submission, and significance. References document the parent mathematics; all instance data, limits, encodings, and scoring rules remain those of the benchmark. Significance denotes the intended mathematical consequence, not difficulty, score weights, or the novelty of every feasible candidate.

Card colors: Instance-Level Focused Substantial Major Broad-Scope

Shared submission and verification conventions. Submit one schema-compliant JSON object under `OUTPUT:`, followed by a nonempty `SOLUTION:` explaining the derivation and checks. Unless stated otherwise, graphs are finite and simple, and finite objects must be explicit: prose or unexpanded generators do not suffice. L1/L3 require exact acceptance; L2 reports a native metric only after feasibility checks. Incomplete verification has no score; search exhaustion is indeterminate, not proof of nonexistence. Verification concerns artifacts and required certificates, not all accompanying prose. Specified unsatisfiability certificates use the benchmark's RUP/LRAT interface (Cruz-Filipe et al., 2017).

B.1 L1 Rediscovery Problems

These ten tasks reconstruct designated results or constructions. Exact answers are compared with audited targets; finite constructions are checked through identities or certificates. Acceptance establishes recovery under the task contract, not independence from prior exposure or historical novelty.

R01 Non-Square Congruent Tilings of Triangles (Erdős #633)

Focused Contribution

Context. A non-square congruent tiling partitions a triangle T into N mutually congruent triangles, with N not a perfect square. The classification comprises eight possibly overlapping clauses involving the angles and arithmetic parameters (Beeson et al., 2026).

Question. Recover the eight clauses whose union is necessary and sufficient. Within each clause, A, B, C may denote any ordering of the angles of T; a ratio M/K uses coprime positive integers.

Submission and verification. Submit eight objects in `cases`. Each contains an `angle_relation` string and an `arithmetic_condition` string in the declared mathematical syntax. Acceptance requires the complete audited classification, modulo the permitted angle permutations and equivalent angle relations, with no duplicated or additional clause.

Significance. The target recovers a geometric classification linking congruent dissection to angle identities and arithmetic restrictions, rather than a construction for a single triangle.

R02 Polynomial Lemniscate Component Perimeter (Erdős #1044)

Focused Contribution

Context. For a monic polynomial $f(z) = \prod_{i=1}^{n}(z - z_i)$ with $|z_i| \leq 1$, define

$$\Lambda(f) = \max_{U} \operatorname{length}(\partial U),$$

where U ranges over connected components of $\{z \in \mathbb{C} : |f(z)| < 1\}$ and length denotes Euclidean boundary length (Tang, 2026).

Question. Determine the infimum of $\Lambda(f)$ over all positive degrees n and all admissible roots.

Submission and verification. Submit the reduced rational value using `infimum_numerator` and a positive `infimum_denominator`, together with an explanation of the lower bound and a construction approaching the infimum. Acceptance requires exact agreement with the audited extremal constant, not numerical proximity.

Significance. This identifies a sharp extremal constant governing the boundary geometry of polynomial lemniscates.

R03 Critical Growth Base for an Irrational Series (Erdős #1051)

Focused Contribution

Context. Let $S = \sum_{n\geq 1}(a_n a_{n+1})^{-1}$ for a strictly increasing sequence of positive integers (a_n). Define β_* as the infimum of bases $\beta > 1$ for which

$$\limsup_{n\to\infty} a_n^{1/\beta^n} = \infty \quad \Longrightarrow \quad S \notin \mathbb{Q}$$

holds for every such sequence (Barreto et al., 2026).

Question. Determine β_* by an exact algebraic-number certificate.

Submission and verification. Submit `minimal_polynomial` as primitive integer coefficients in descending powers, with positive leading coefficient, and `isolating_interval` with reduced rational endpoints. Acceptance checks the audited polynomial and that the interval isolates the intended root greater than one.

Significance. The target recovers the sharp growth threshold for a universal irrationality criterion, rather than a sufficient but non-sharp bound.

R04 High-Dimensional Distinct-Distance Limit (Erdős #1089)

Focused Contribution

Context. For fixed $n \geq 2$, let $g_d(n)$ be the smallest m such that every m-point set in $\mathbb{R}^d$ determines at least n distinct pairwise distances. The problem concerns the asymptotic size of few-distance configurations (Bannai et al., 1983).

Question. Determine the exact function

$$L(n) = \lim_{d\to\infty} \frac{g_d(n)}{d^{n-1}}.$$

Submission and verification. Submit `limit_expression`, an exact ASCII expression in n using the permitted arithmetic and factorial syntax, with a derivation from matching asymptotic bounds. Acceptance compares the expression with the audited formula by exact symbolic normalization; agreement at sampled values is insufficient.

Significance. The result determines the leading asymptotic scale of high-dimensional configurations with few distinct distances.

R05 Graffiti 154 Spectral Counterexample

Focused Contribution

Context. For a connected graph G with n vertices, m edges, and Wiener index $W = \sum_{u<v} d(u,v)$, the adjacency eigenvalues have population standard deviation $\sigma_A = \sqrt{2m/n}$. The two average-distance conventions are $\bar{d}_D = 2W/[n(n-1)]$ and $\bar{d}_A = 2W/n^2$ (Demonstrandum Research, 2026).

Question. Construct a graph satisfying both $\sigma_A > n/\bar{d}_D$ and $\sigma_A > n/\bar{d}_A$.

Submission and verification. Submit `order`, with $3 \leq n \leq 256$, and the complete `edges` list. Binary acceptance requires a simple connected graph and exact verification of both strict inequalities; distances and edge counts are recomputed from the artifact.

Significance. One witness disproves the proposed spectral–distance inequality under both average-distance conventions, without relying on a choice of normalization.

R06 Hadamard Matrix of Order 668

Focused Contribution

Context. A Hadamard matrix of order n is a sign matrix $H \in \{-1, +1\}^{n \times n}$ with $HH^{\mathsf{T}} = nI_n$. This rediscovery task fixes $n = 668$ (Epoch AI, 2026b).

Question. Construct such a matrix for $n = 668$, normalized so that the first row and first column contain only $+1$.

Submission and verification. Submit `order`, the prescribed `encoding`, and all 668 packed rows in `rows_b64`. Each row is packed into 84 bytes, most-significant-bit first, with 0 denoting $+1$, 1 denoting -1, and four zero padding bits. Acceptance checks canonical decoding, normalization, and exact orthogonality of every distinct row pair; a generator alone is insufficient.

Significance. The artifact recovers an orthogonal design at one specified order. It does not establish the Hadamard existence conjecture for all admissible orders.

R07 WOW-284 Dual-Degree Counterexample

Focused Contribution

Context. The minimum dual degree is

$$\delta^*(G) = \min_v \frac{1}{d(v)} \sum_{u \in N(v)} d(u).$$

WOW-284 asserts $\delta^*(G) \leq -\lambda_{\min}(D(G))$ for connected graphs of girth at least five; $D(G)$ is the shortest-path distance matrix (Petkov, 2026).

Question. Construct a strongly regular graph with parameters $(50, 7, 0, 1)$: degree seven, no common neighbor for adjacent vertices, and exactly one for nonadjacent vertices.

Submission and verification. Submit the complete 50×50 binary `adjacency_matrix`. Exact verification of these parameters certifies girth five, $\delta^* = 7$, and $\lambda_{\min}(D) = -4$, hence the required strict violation $7 > 4$.

Significance. The construction certifies a spectral counterexample using finite combinatorial identities.

R08 Jackson Hamilton-Decomposition Counterexample

Focused Contribution

Context. A regular bipartite tournament orients $K_{2n,2n}$ so that every vertex has indegree and outdegree n. A Hamilton decomposition partitions all arcs into n directed Hamilton cycles, each visiting every vertex once (Granet, 2022).

Question. Construct a regular orientation of $K_{6,6}$ that has no decomposition into three directed Hamilton cycles.

Submission and verification. Submit a 6×6 binary `orientation_matrix`: entry one means $L_i \to R_j$ and zero means $R_j \to L_i$. Acceptance checks that every row and column sums to three, then exactly excludes all Hamilton decompositions of the digraph.

Significance. A valid orientation refutes the universal decomposition assertion within a small, explicitly specified family of regular digraphs.

R09 Cost-Preserving Unsplittable-Flow Counterexample

Focused Contribution

Context. A single source routes demands d_i to terminals in a directed acyclic graph. A feasible splittable arc flow x costs $C_f = \sum_a c_a x_a$. An unsplittable routing chooses one path per terminal; its arc loads ℓ_a sum the demands using each arc, and its cost is $\sum_a c_a \ell_a$ (Ramos et al., 2026).

Question. Construct a seven-vertex, nine-arc instance with three positive integer demands and nonnegative integer arc flows and costs, such that every routing satisfying $\ell_a \le x_a + d_{\max}$ has cost strictly greater than C_f, where $d_{\max} = \max_i d_i$.

Submission and verification. Submit the complete `vertices`, `source`, `terminals`, and `arcs` records. Acceptance checks acyclicity and flow conservation, then checks the load and cost conditions over all possible unsplittable routings.

Significance. The instance demonstrates that the stated congestion allowance cannot always be combined with cost-preserving rounding.

R10 Three-Dimensional Jacobian Conjecture Counterexample

Substantial Advance

Context. For $F = (F_1, F_2, F_3) \in \mathbb{Z}[x, y, z]^3$, a nonzero constant Jacobian determinant implies local invertibility of the induced map $\mathbb{C}^3 \to \mathbb{C}^3$. Distinct inputs with the same image instead witness failure of global injectivity (Gao, 2026).

Question. Construct F with $\det J_F = c \neq 0$ constant and at least two distinct rational points p, q satisfying $F(p) = F(q)$.

Submission and verification. Submit three complete sparse term lists in `polynomials`, each term specifying an integer coefficient and its three exponents in $[0, 32]$, and reduced rational `collision_points`. Acceptance requires exact symbolic verification of the determinant and exact evaluation at the supplied points.

Significance. The required certificate separates local nonsingularity from global polynomial invertibility. The frozen L1 designation awards credit for reconstructing the specified counterexample, not for historical priority.

B.2 L2 Progressive Discovery Problems

These twenty tasks measure improvement among feasible constructions: sixteen minimize a native objective and four maximize one. Each card specifies its score direction and normalization. A positive defect can be a valid optimization result without satisfying the corresponding exact discovery criterion.

P01 Manhattan Traveling Salesperson Problem with 512 Cities

Instance-Level Result

Context. The metric traveling salesperson problem seeks a shortest closed tour through all cities (Dantzig et al., 1954). This instance fixes 512 integer-coordinate cities $p_i = (x_i, y_i)$ and Manhattan distances $d(i, j) = |x_i - x_j| + |y_i - y_j|$.

Question. Find a permutation π of $\{0, \ldots, 511\}$ minimizing the closed-tour length

$$L(\pi) = \sum_{r=0}^{511} d\big(\pi_r, \pi_{(r+1) \bmod 512}\big).$$

Submission and metric. Submit the 512 labels in `tour`. After checking the permutation, the evaluator recomputes every edge, including the return to the starting city. The metric is $L(\pi)/10^6$, with lower values preferred.

Significance. This fixed routing instance tests search over a large finite space. A shorter tour improves the instance objective but is not, by itself, a new general mathematical result.

P02 Ternary Covering Array with 220 Rows

Instance-Level Result

Context. For $X \in \{0, 1, 2\}^{220 \times 10}$, let $c_{J,a}$ count rows that equal $a \in \{0, 1, 2\}^4$ on a four-column set J. Index-two coverage requires every such interaction at least twice (Calbert & Dougherty, 2022). Repeated rows count with multiplicity.

Question. Minimize the index-two coverage loss

$$L(X) = \sum_{\substack{J \subseteq \{1,\ldots,10\} \\ |J|=4}} \sum_{a \in \{0,1,2\}^4} \max\{0, 2 - c_{J,a}\}^2.$$

Submission and metric. Submit all 220 row vectors in `rows`. The evaluator checks the dimensions and alphabet and reports the exact integer loss; lower is better. An interaction occurring zero, one, or at least two times contributes 4, 1, or 0, respectively.

Significance. Zero loss certifies a ternary strength-four covering array of index two at the fixed size. It does not certify that 220 rows are necessary or optimal.

P03 Graceful Labeling of a Fixed 127-Vertex Tree

Instance-Level Result

Context. A graceful labeling makes all positive edge differences distinct (Gallian, 2024). Here T is the 127-vertex tree obtained by smallest-leaf decoding of the Prüfer sequence $p_i = (37i^2 + 11i + 5) \bmod 127$, $0 \le i \le 124$.

Question. Choose a bijection $f : V(T) \to \{0, \dots, 126\}$ minimizing

$$D(f) = 126 - \big|\{|f(u) - f(v)| : uv \in E(T)\}\big|.$$

Submission and metric. Submit the vertex-label permutation in `labels`. The metric is the exact defect $D(f)$, with lower values preferred. A zero defect means that the 126 edge differences are precisely $1, \dots, 126$.

Significance. The objective measures labeling quality on one specified tree. A graceful labeling here does not resolve the universal graceful-tree conjecture.

P04 Capacitated Vehicle Routing with 100 Customers

Instance-Level Result

Context. Capacitated vehicle routing serves customer demands by depot-to-depot tours (Dantzig & Ramser, 1959). This instance fixes 100 customers with integer coordinates and positive demands, a depot at $(500, 500)$, Manhattan costs, and capacity 40 per vehicle.

Question. Choose a customer permutation. A prescribed greedy decoder scans it from left to right, starting a new route exactly when adding the next demand would exceed capacity; every route starts and ends at the depot.

Submission and metric. Submit `tour`, a permutation of labels $1, \dots, 100$. The metric is the exact total length of the decoded routes, including every depot departure and return; lower is better. Route boundaries are determined by the decoder, not separately optimized or submitted.

Significance. This is a fixed-instance test of ordering under a capacity-constrained decoder. Its score concerns that representation, rather than a certificate of global optimality over all vehicle-routing solutions.

P05 Resource-Constrained Project Scheduling with 60 Jobs

Instance-Level Result

Context. Resource-constrained project scheduling combines precedence constraints with limited renewable resources (Kolisch, 1996). The fixed project has 58 real jobs, two dummy jobs, and resource capacities $(10, 12, 8)$; each real job has a specified duration and demand vector.

Question. Choose a priority permutation of the real jobs. A serial schedule generator repeatedly selects the highest-priority job whose predecessors have all been scheduled, then assigns its earliest integer start time satisfying predecessor completion and resource capacities. Jobs are nonpreemptive.

Submission and metric. Submit `priority`, a permutation of $1, \dots, 58$. The evaluator constructs the schedule and minimizes

$$J = 100000\, C_{\max} + \sum_{j=1}^{58} s_j,$$

where $C_{\max}$ is the makespan and s_j is a real job's start time. The priority order itself need not be topological.

Significance. The score combines makespan with a secondary start-time preference on one fixed instance; it does not establish a general scheduling result.

P06 Weighted $K_{2,3}$ Homomorphism Minimization

Instance-Level Result

Context. For an 18-vertex graph G with exactly 45 edges, set $c(u, v) = |N(u) \cap N(v)|$. Its $K_{2,3}$ homomorphism count is (Sidorenko, 1991)

$$H(G) = \sum_{u,v \in V(G)} c(u, v)^3.$$

Diagonal terms $u = v$ are included. Write $W(G) = \sum_{uv \in E(G)} w(u, v)$ for the total prescribed edge weight.

Question. Minimize the lexicographically encoded objective $J(G) = 45000136\, H(G) - W(G)$.

Submission and metric. Submit exactly 45 distinct edges in `edges`, with $0 \le u < v \le 17$. The evaluator recomputes the homomorphism count and the prescribed label-dependent total weight and reports the exact integer J; lower is better.

Significance. The task calibrates extremal graph search with a deterministic tie-breaker. Improving this weighted fixed-instance objective is not itself a counterexample to Sidorenko's conjecture.

P07 Sidon Sets in an Interval of Length 10,000

Focused Contribution

Context. A Sidon set $A = \{a_1 < \cdots < a_k\} \subseteq \{1, \ldots, 10000\}$ has distinct sums $a_i + a_j$ over all index pairs $i \le j$. Thus the uniqueness condition includes sums of an element with itself (O'Bryant, 2004).

Question. Construct an admissible set with the largest possible cardinality k.

Submission and metric. Submit the nonempty increasing list in `values`. The evaluator checks the bounds and every pair-sum collision, then reports $|A|$; higher is better. Any collision between pair sums makes the set infeasible.

Significance. Each feasible construction certifies a lower bound on the maximum Sidon-set size in the fixed interval. An improved construction does not establish the matching upper bound.

P08 Low-Autocorrelation Binary Sequences of Length 256

Focused Contribution

Context. For $s \in \{-1, +1\}^{256}$, the aperiodic autocorrelation at lag k is

$$C_k(s) = \sum_{i=0}^{255-k} s_i s_{i+k}, \qquad 1 \le k \le 255.$$

These sums do not wrap around the sequence (Packebusch & Mertens, 2016).

Question. Minimize the autocorrelation energy $E(s) = \sum_{k=1}^{255} C_k(s)^2$.

Submission and metric. Submit all 256 signs in `sequence`. After checking the length and alphabet, the evaluator reports the exact integer energy; lower is better.

Significance. The task seeks improved finite sequences with low aggregate off-peak autocorrelation. A low-energy sequence is a constructive bound, not an optimality proof.

P09 Golomb Rulers with 29 Marks

Focused Contribution

Context. A 29-mark Golomb ruler is an integer sequence $0 = a_0 < a_1 < \cdots < a_{28}$ with all $\binom{29}{2} = 406$ positive differences $a_j - a_i$, $i < j$, distinct (Drakakis, 2009).

Question. Construct an admissible ruler minimizing its total length $L = a_{28}$.

Submission and metric. Submit all 29 integers in `marks`. The evaluator checks the zero origin, strict ordering, and uniqueness of every positive difference, then reports L; lower is better.

Significance. A shorter ruler improves a constructive upper bound on the minimum length of a 29-mark Golomb ruler. The ruler alone does not certify optimality.

P10 Primitive Near-Perfect Cuboids in a Dyadic Range

Focused Contribution

Context. An Euler brick has integer edges and integer face diagonals (Guy, 2004). The search is restricted to primitive triples with

$$2^{53} \le a < b < c < 2^{54}, \qquad \gcd(a, b, c) = 1,$$

for which $a^2 + b^2$, $a^2 + c^2$, and $b^2 + c^2$ are perfect squares.

Question. Set $Q = a^2 + b^2 + c^2$ and $s = \lfloor \sqrt{Q} \rfloor$. Minimize the nearest-square defect

$$\Delta = \min\{Q - s^2, (s + 1)^2 - Q\}.$$

Submission and metric. Submit exactly the three integer fields `a`, `b`, and `c`. The evaluator verifies every hard constraint and recomputes Δ with exact integer arithmetic; lower is better.

Significance. The metric measures the square defect of the space diagonal within a bounded primitive family. Only $\Delta = 0$ constructs a perfect cuboid; a small positive square defect does not.

P11 Vertex-Deleted Deck Overlap for 14-Vertex Graphs

Focused Contribution

Context. A graph's vertex-deleted deck is the multiset of its one-vertex-deleted subgraphs up to isomorphism (McKay, 2021). Seek nonisomorphic graphs G, H with 14 vertices and 32 edges each, equal degree multisets, and $G, H, \overline{G}, \overline{H}$ all connected.

Question. Form a bipartite graph on the deletion choices, joining v to w when $G - v \cong H - w$. Maximize its matching number M, which measures deck overlap with multiplicity.

Submission and metric. Submit complete `G` and `H` objects, each with `n` and `edges`. After exact feasibility and card-isomorphism checks, the metric is the maximum matching size M; higher is better.

Significance. A score of 14 would give equal decks for nonisomorphic graphs and refute reconstruction. Smaller scores measure deck overlap without refuting reconstruction.

P12 Union-Closed Frequency Minimization on 13 Elements

Focused Contribution

Context. Let $\mathcal{F}$ consist of 64 distinct subsets of $\{0, \ldots, 12\}$. It must contain $\varnothing$, have full union, and be closed under pairwise union. It must also separate elements: each pair is distinguished by a set containing exactly one of them (Bruhn & Schaudt, 2015).

Question. For $f_i = |\{A \in \mathcal{F} : i \in A\}|$, minimize $f_{\max} = \max_i f_i$.

Submission and metric. Submit `family_hex`, a strictly increasing list of 64 four-digit lowercase hexadecimal masks, using only bits zero through 12. The evaluator verifies all family constraints and reports $f_{\max}$; lower is better.

Significance. A score at most 31 would make every element occur in strictly fewer than half of the sets, giving a union-closed-sets counterexample. A score of 32 does not violate the conjectured half-frequency bound.

P13 Sidorenko Homomorphism Search on a 16-Vertex Host

Focused Contribution

Context. This Sidorenko-type homomorphism search (Sidorenko, 1991) fixes the bipartite graph H with parts $\{L_0, \ldots, L_4\}$ and $\{R_0, \ldots, R_4\}$, joining L_i to R_j exactly when $(j - i) \bmod 5 \in \{0, 1, 2\}$. The fixed source graph has ten vertices and 15 edges.

Question. Choose a simple graph G with 16 vertices and exactly 64 edges to minimize $\hom(H, G)$, the number of edge-preserving vertex maps, including noninjective maps.

Submission and metric. Submit 16 packed adjacency rows in `adjacency_rows_hex`, each a four-digit lowercase hexadecimal string. The evaluator verifies the host constraints and recomputes the exact homomorphism count; lower is better.

Significance. The host density is $2|E(G)|/16^2 = 1/2$. A count strictly below $16^{10}(1/2)^{15} = 2^{25}$ would refute Sidorenko's inequality for this fixed H; a count at or above the threshold is only an optimization result.

P14 Hadamard Maximal Determinant (order 23)

Substantial Advance

Context. The maximal determinant problem seeks a sign matrix with largest determinant magnitude (Orrick et al., 2003). At order 23, the target is

$$D(23) = \max_{H \in \{-1,+1\}^{23 \times 23}} |\det H|.$$

The feasible objects are discrete matrices, not continuous relaxations.

Question. Construct a 23×23 sign matrix maximizing its determinant magnitude.

Submission and metric. Submit the complete matrix in `matrix`. The evaluator computes $|\det H|$ exactly. The metric is $|\det H|/D_{\mathrm{ref}}$, using the frozen reference $D_{\mathrm{ref}} = 2\,779\,447\,296\,000\,000$. Higher is better: a score of one matches the reference, while a score above one exceeds it.

Significance. A stronger determinant construction advances the specified extremal matrix problem. Neither the reference value nor an improved candidate alone certifies the exact maximum $D(23)$.

P15 Projective Plane (order 12)

Substantial Advance

Context. A projective plane of order 12 has 157 points and 157 lines, with 13 incidences at each (Dembowski, 1968). Represent a candidate matrix N by 13 permutations π_t of $\{0, \dots, 156\}$, whose images are distinct at each position. Every row and column of N then sums to 13.

Question. For $\lambda_{ij} = \sum_c N_{ic} N_{jc}$, minimize the four-cycle count

$$C_4(N) = \sum_{i<j} \binom{\lambda_{ij}}{2}.$$

Submission and metric. Submit 13 objects in `matchings`, either explicit permutation images or affine maps $\pi(i) = ai + b \pmod{157}$ with $a \neq 0$. The evaluator reconstructs N and reports the exact C_4; lower is better.

Significance. Zero defect certifies a projective plane of this non-prime-power order. Positive defect measures a feasible near-design and establishes neither existence nor nonexistence of a plane.

P16 Three Mutually Orthogonal Latin Squares of Order 10

Substantial Advance

Context. A Latin square of order 10 is a 10×10 array whose rows and columns are permutations of $\{0, \dots, 9\}$. Two squares are orthogonal if superimposing them gives 100 distinct ordered symbol pairs (Egan & Wanless, 2016).

Question. Choose three distinct Latin squares L_0, L_1, L_2 minimizing

$$D = \sum_{0 \le a < b \le 2} \left(100 - \left|\{(L_a(i,j), L_b(i,j)) : 0 \le i, j < 10\}\right|\right).$$

Submission and metric. Submit all three arrays in `squares`. Latin validity and distinctness are hard constraints; the metric is the exact pairwise orthogonality defect D, with lower values preferred.

Significance. A zero score constructs three mutually orthogonal Latin squares of order 10. Positive defect quantifies approximation to this design without establishing its existence.

P17 Ramsey $(5, 5)$ Colorings on 43 Vertices

Substantial Advance

Context. A graph on 43 vertices encodes a red–blue coloring of the complete graph: edges receive one color and nonedges the other. A monochromatic five-set is either a clique or an independent set (Radziszowski, 1994).

Question. Minimize the total number of five-cliques and independent five-sets,

$$L(G) = \#K_5(G) + \#K_5(\overline{G}).$$

Submission and metric. Submit a symmetric, zero-diagonal 43×43 binary `adjacency_matrix`. The evaluator counts both types of forbidden subset exactly; lower is better.

Significance. A zero score certifies $R(5, 5) > 43$, equivalently $R(5, 5) \geq 44$. Reducing a positive violation count improves the objective but does not certify this Ramsey bound.

P18 Schur Six-Coloring on 537 Integers

Substantial Advance

Context. A set is sum-free if it contains no triple $x, y, x + y$, including $x = y$ (Heule, 2017). The task colors $\{1, \ldots, 537\}$ with six colors and measures violations of this condition within the color classes.

Question. Choose $c_x \in \{0, \ldots, 5\}$ to minimize

$$L(c) = \sum_{\substack{1 \le x \le y \\ x+y \le 537}} \mathbf{1}\{c_x = c_y = c_{x+y}\}.$$

Submission and metric. Submit `kind` equal to `explicit` and `colors_digits`, a string of exactly 537 digits from zero through five. The evaluator reports the exact number of monochromatic additive triples; lower is better.

Significance. Zero violations partition the interval into six sum-free sets, certifying $S(6) \ge 537$ where $S(6)$ denotes the largest colorable endpoint. Positive scores do not establish that bound.

P19 Conway 99-Graph Common-Neighbor Defect

Substantial Advance

Context. In a simple 14-regular graph on 99 vertices, let $c(u, v) = |N(u) \cap N(v)|$. The Conway parameters prescribe $t(u, v) = 1$ for adjacent vertices and $t(u, v) = 2$ for nonadjacent vertices (Cesarz & Woldar, 2025).

Question. Minimize the common-neighbor *shortage*

$$D(G) = \sum_{u<v} \max\{t(u, v) - c(u, v), 0\}.$$

This is a one-sided shortage, not an absolute or squared discrepancy.

Submission and metric. Submit `adjacency_rows_hex`, comprising 99 rows of 25 lowercase hexadecimal digits, using only bits zero through 98. Simplicity and exact 14-regularity are hard constraints; the exact integer D is minimized.

Significance. The fixed total common-neighbor count makes $D = 0$ equivalent to an $\mathrm{srg}(99, 14, 1, 2)$. A positive score is a near-construction, not a Conway 99-graph.

P20 Two-Colorings without Seven-Term Arithmetic Progressions

Substantial Advance

Context. The van der Waerden number $W(2, 7)$ is the least N such that every two-coloring of $\{1, \ldots, N\}$ contains a monochromatic seven-term arithmetic progression (Rabung & Lotts, 2012).

Question. Construct the longest binary string $c_1, \ldots, c_N$ such that no progression $a, a + d, \ldots, a + 6d$ with $d \ge 1$ and $a + 6d \le N$ is monochromatic.

Submission and metric. Submit the nonempty binary string in `colors_digits`, with its first digit normalized to zero. The evaluator checks all relevant progressions and maximizes the valid string length N; any violation makes the candidate infeasible.

Significance. Every feasible string certifies $W(2, 7) > N$. Longer valid strings strengthen this lower bound without providing a matching upper bound.

B.3 L3 Decisive Discovery Problems

These thirty-five tasks require exact witnesses: eight seek positive constructions and twenty-seven seek counterexamples. Acceptance requires every stated condition; nearby constructions receive no decisive-discovery credit. Numerical ranges and certificate restrictions belong to the frozen interfaces. Significance increases from Substantial Advance to Major Resolution and Broad-Scope Resolution.

D01 Hadamard Matrix of Order 2092

Substantial Advance

Context. A Hadamard matrix is a sign matrix with pairwise orthogonal rows (Paley, 1933). This interface asks for order 2092, a single case of the existence question for orders divisible by four.

Question. Construct $H \in \{-1, +1\}^{2092\times 2092}$ satisfying $HH^{\mathsf{T}} = 2092I_{2092}$, with first row and column all $+1$.

Submission and verification. Submit `order` equal to 2092, the prescribed `encoding`, and all 2092 strings in `rows_b64`. Each row uses 262 packed bytes, most-significant-bit first, with zero for $+1$, one for -1, and four zero padding bits. Acceptance checks the entire decoded matrix and every exact orthogonality condition.

Significance. The result establishes existence at one specified order in the Hadamard construction problem. It does not prove existence at all orders divisible by four.

D02 Perfect Cuboid Problem

Major Resolution

Context. A perfect cuboid is a rectangular box with positive integer edges and all four diagonals integral: three face diagonals and one space diagonal. An Euler brick satisfies only the three face-diagonal conditions (Guy, 2004).

Question. Construct positive integers $a \le b \le c$ and $d_{ab}, d_{ac}, d_{bc}, d_{abc}$ satisfying

$$a^2 + b^2 = d_{ab}^2, \quad a^2 + c^2 = d_{ac}^2, \quad b^2 + c^2 = d_{bc}^2, \quad a^2 + b^2 + c^2 = d_{abc}^2.$$

Submission and verification. Submit the seven integers as `a`, `b`, `c`, `d_ab`, `d_ac`, `d_bc`, and `d_abc`. Acceptance checks positivity, edge ordering, and all four equations exactly. Unlike the L2 interface, this task imposes neither primitivity nor a dyadic size range.

Significance. One valid tuple establishes the existence of a perfect cuboid. Integral face diagonals alone, or a nearly integral space diagonal, do not establish the target.

D03 Lander–Parkin–Selfridge Conjecture Counterexample

Major Resolution

Context. The Lander–Parkin–Selfridge conjecture predicts at least k terms in a nontrivial equality of positive kth powers after cancelling bases common to both sides (Guy, 2004).

Question. Find an integer $k \ge 3$ and nonempty finite lists L, R of positive integers such that

$$\operatorname{supp}(L) \cap \operatorname{supp}(R) = \varnothing, \qquad |L| + |R| < k, \qquad \sum_{a\in L} a^k = \sum_{b\in R} b^k.$$

Repetitions within either list are allowed and count as separate terms.

Submission and verification. Submit `k`, `left`, and `right`. The evaluator checks support disjointness, the strict term-count bound, and equality by exact integer exponentiation.

Significance. The equality refutes the conjectured term-count bound. A counterexample to a stronger sum-of-powers claim need not meet this task's stricter target.

D04 Lehmer's Totient Problem

Major Resolution

Context. Euler's totient $\varphi(N)$ counts integers in $\{1, \ldots, N\}$ coprime to N. For primes, $\varphi(N) = N - 1$. Lehmer's problem asks whether $\varphi(N) \mid (N - 1)$ can also hold for composite N (Guy, 2004).

Question. Construct a composite integer $N \geq 4$ for which $\varphi(N) \mid (N - 1)$.

Submission and verification. Submit `N` and its complete prime-power factorization in `factors`, using records `p`, `e`. The evaluator verifies the factors and computes

$$\varphi(N) = \prod_i p_i^{e_i - 1}(p_i - 1)$$

exactly. Every prime factor must be below 2^{64}; this interface checks primality internally and accepts no separate primality-certificate fields.

Significance. A valid composite witness disproves the proposed implication from totient divisibility to primality.

D05 Erdős–Straus Conjecture Counterexample

Major Resolution

Context. The Erdős–Straus conjecture represents $4/n$ as three positive unit fractions. With ordered denominators $x \leq y \leq z$, the equivalent integer equation is $4xyz = n(xy + xz + yz)$ (Salez, 2014).

Question. Find $10^{17} < n < 2^{63}$ for which no positive ordered solution exists.

Submission and verification. Submit `n` and a mandatory `proof`: an RUP/LRAT refutation bound by `cnf_sha256` to the benchmark's canonical bit-vector CNF encoding. The encoding uses the task's mathematically complete denominator bounds. The proof is supplied as ASCII LRAT text or gzip-compressed Base64 chunks. Acceptance requires successful exact proof checking, not an unsuccessful search for denominators.

Significance. A certified unsatisfiable instance refutes the universal three-unit-fraction representation claim. The numerical interval is a frozen interface restriction, not a claim about the latest literature bound.

D06 Beal Conjecture Counterexample

Major Resolution

Context. Beal's conjecture states that $A^x + B^y = C^z$ with positive integer bases and integer exponents exceeding two forces a prime divisor common to A, B, C (Mauldin, n.d.).

Question. Construct positive integers A, B, C and integers $x, y, z > 2$ satisfying

$$A^x + B^y = C^z, \qquad \gcd(A, B, C) = 1.$$

The coprimality condition concerns the three-way greatest common divisor; pairwise coprimality is not separately required.

Submission and verification. Submit the six integer fields `A`, `B`, `C`, `x`, `y`, and `z`. Acceptance checks positivity, exponent bounds, the greatest common divisor, and the power identity exactly.

Significance. One primitive equality refutes the requirement of a prime divisor common to all three bases.

D07 Legendre Conjecture Counterexample by Composite Coverage

Major Resolution

Context. Legendre's conjecture requires a prime strictly between n^2 and $(n+1)^2$ (Sorenson & Webster, 2025). The integers to be checked are $n^2 + 1 + i$, $0 \le i < 2n$.

Question. Find $n > 70500000000000$ and certify that every integer in this interval is composite.

Submission and verification. Submit `n` and a residue-partition tree in `nodes`. The root $(M, r) = (1, 0)$ represents all indices; a node (M, r) splits into $(kM, r + jM)$ for $0 \le j < k$, with $2 \le k \le 64$. Each leaf provides a proper divisor d satisfying $d \mid M$ and $d \mid n^2 + 1 + r$. Acceptance checks complete interval coverage and that the divisor is proper for every represented integer.

Significance. The certificate establishes an entire prime-free interval between consecutive squares. A list of composites that leaves any index unchecked is insufficient.

D08 Odd Perfect Number

Major Resolution

Context. A positive integer N is perfect if its divisor sum satisfies $\sigma(N) = 2N$. The existence question here restricts N to be odd (Iannucci & Sorli, 2003).

Question. Construct such an N together with its complete prime-power factorization.

Submission and verification. Submit `N` and `factors`, with one `p`, `e` record per distinct prime. The evaluator checks the factorization and oddness and computes

$$\sigma(N) = \prod_i \frac{p_i^{e_i+1} - 1}{p_i - 1}$$

exactly. Every prime factor must be below 2^{64}; primality is checked internally, without separate primality-certificate fields.

Significance. An accepted factorization establishes the existence of an odd perfect number through an exact divisor-sum identity, rather than numerical evidence of near-perfectness.

D09 Tuza Triangle Packing–Covering Counterexample

Major Resolution

Context. Let $\nu(G)$ be the maximum number of pairwise edge-disjoint triangles, and $\tau(G)$ the minimum number of edges meeting every triangle. Tuza's conjecture asserts $\tau(G) \le 2\nu(G)$ (Chen et al., 2016).

Question. Construct a graph with $\tau(G) > 2\nu(G)$.

Submission and verification. Submit `n` and the complete `edges` list. Acceptance requires exact computation of both optimization quantities and verification of the strict inequality; heuristic packing or covering values do not certify the inequality.

Significance. A valid graph disproves the factor-two relation between triangle packing and edge covering. The covering objects are edges, not vertices.

D10 Conway 99-Graph

Major Resolution

Context. A strongly regular graph with parameters (v, k, λ, μ) is k-regular; adjacent vertex pairs have exactly λ common neighbors and nonadjacent pairs exactly μ. Conway's 99-graph asks for $(99, 14, 1, 2)$ (Cesarz & Woldar, 2025).

Question. Construct a graph with parameters $(99, 14, 1, 2)$.

Submission and verification. Submit `n` equal to 99 and the complete `edges` list. Acceptance checks simplicity, degree 14 at every vertex, and the exact common-neighbor count for every unordered pair. Unlike the L2 shortage objective, this interface requires all equalities simultaneously.

Significance. An accepted graph settles the specified strongly regular graph existence case. A nonzero common-neighbor defect, however small, does not meet the target.

D11 Erdős–Gyárfás Cycle-Length Counterexample

Major Resolution

Context. The Erdős–Gyárfás conjecture asserts that every graph with minimum degree at least three contains a cycle whose length is a power of two (Hegde et al., 2024).

Question. Construct a simple graph G with $\delta(G) \geq 3$ and no simple cycle of length 2^k for any $k \geq 2$ with $2^k \leq |V(G)|$.

Submission and verification. Submit `n` and `edges`. Acceptance checks the degree condition and exactly excludes every relevant cycle length, beginning with four. An interrupted search that finds no such cycle is not an exclusion certificate.

Significance. One accepted graph disproves the claimed link between a local minimum-degree condition and the presence of a cycle in the prescribed length family.

D12 Barnette Hamiltonicity Counterexample

Major Resolution

Context. Barnette's conjecture asserts that every cubic, bipartite, planar, 3-connected graph has a Hamilton cycle. All four structural conditions are essential to this target (Schnieders, 2025).

Question. Construct a graph satisfying all four structural conditions but admitting no Hamilton cycle.

Submission and verification. Submit `n`, `edges`, cyclic face lists in `faces`, and `proof`. The face lists certify a planar embedding. A non-null proof is a SHA-256-bound LRAT refutation of the canonical Hamilton-cycle CNF; `null` requests bounded internal exact search. Acceptance requires the structural checks and a completed non-Hamiltonicity verification.

Significance. A valid witness refutes Hamiltonicity in this restricted planar graph class. All four structural conditions must hold alongside certified non-Hamiltonicity.

D13 Moore Graph of Degree 57

Major Resolution

Context. A degree-57 Moore graph of diameter two has $57^2 + 1 = 3250$ vertices. Equivalently, it is strongly regular with parameters $(3250, 57, 0, 1)$ (Dalfó, 2019).

Question. Construct a 3250-vertex graph in which every vertex has degree 57, adjacent vertices have no common neighbor, and nonadjacent vertices have exactly one.

Submission and verification. Submit `n` equal to 3250 and the complete `edges` list, containing 92,625 edges. Acceptance checks every degree and common-neighbor identity exactly.

Significance. The artifact settles this extremal graph existence case by attaining the degree–diameter bound with the specified Moore structure.

D14 Caccetta–Häggkvist Counterexample

Major Resolution

Context. The Caccetta–Häggkvist conjecture bounds the length of a shortest directed cycle by $\lceil n/r \rceil$ in a loopless n-vertex digraph with positive minimum outdegree r (Shen, 2000).

Question. Construct a digraph whose directed girth $g^+(G)$ satisfies

$$g^+(G) > \left\lceil \frac{n}{r} \right\rceil .$$

Directed girth is the shortest directed cycle length, with value ∞ if no directed cycle exists.

Submission and verification. Submit `n` and the directed arc list in `edges`. Acceptance recomputes the minimum outdegree and directed girth and checks the strict violation exactly.

Significance. Such a digraph refutes the proposed sharp relation between minimum outdegree and short directed cycles.

D15 Seymour Second-Neighborhood Counterexample

Major Resolution

Context. For an oriented graph, let $N_1^+(v)$ denote the out-neighbors and define the strict second out-neighborhood as (Seacrest, 2018)

$$N_2^+(v) = \left(\bigcup_{u \in N_1^+(v)} N_1^+(u) \right) \setminus \left(N_1^+(v) \cup \{v\} \right).$$

Oriented graphs have neither loops nor antiparallel arcs.

Question. Construct an oriented graph for which $|N_2^+(v)| < |N_1^+(v)|$ at every vertex.

Submission and verification. Submit `n` and the directed `edges`. Acceptance verifies the orientation constraints and computes both neighborhoods exactly for every vertex; the second neighborhood excludes the vertex itself and its immediate out-neighbors.

Significance. The artifact refutes the assertion that at least one vertex has a second out-neighborhood no smaller than its first. A violation at only some vertices is insufficient.

D16 Harary–Hill Counterexample via a Straight-Line Drawing

Major Resolution

Context. The Harary–Hill conjecture gives the crossing number of K_n as (Mutzel & Oettershagen, 2018)

$$H(n) = \frac{1}{4}\left\lfloor \frac{n}{2} \right\rfloor \left\lfloor \frac{n-1}{2} \right\rfloor \left\lfloor \frac{n-2}{2} \right\rfloor \left\lfloor \frac{n-3}{2} \right\rfloor .$$

This interface restricts to straight-line drawings from rational planar points in the specified general position.

Question. Construct such a drawing of K_n with strictly fewer than $H(n)$ crossings.

Submission and verification. Submit `coordinates`, with both coordinates of every point encoded as exact rational strings. Acceptance checks general position and recomputes segment crossings using exact geometric predicates.

Significance. A straight-line witness below $H(n)$ would also refute the proposed crossing-number lower bound for unrestricted drawings. Failure to find such a straight-line drawing does not settle the unrestricted problem.

D17 Graph Reconstruction Counterexample

Major Resolution

Context. The graph reconstruction conjecture states that every graph of order at least three is determined, up to isomorphism, by its multiset of vertex-deleted subgraphs (McKay, 2021).

Question. Construct nonisomorphic graphs G, H of the same order $n \geq 3$ with identical vertex-deleted decks, including multiplicities.

Submission and verification. Submit full `G` and `H` graph objects and `deck_bijection`. Each entry identifies one deleted vertex from each graph and supplies an isomorphism between the remaining vertices. Acceptance checks that every deletion is used once, verifies every supplied card isomorphism, and exactly confirms $G \ncong H$.

Significance. The result gives nonisomorphic graphs with identical complete decks. Partial deck overlap, even when large, does not refute reconstruction.

D18 Graceful-Tree Conjecture Counterexample

Major Resolution

Context. An m-edge tree is graceful if a bijection $f : V(T) \to \{0, \ldots, m\}$ yields edge differences $|f(u) - f(v)|$ equal to $1, \ldots, m$, each exactly once (Gallian, 2024).

Question. Construct a tree that admits no graceful labeling.

Submission and verification. Submit `n` and the complete `edges` list. Acceptance first verifies that the graph is a tree, then requires an exact exclusion of every graceful labeling. One labeling with repeated differences does not exclude all graceful labelings.

Significance. An accepted tree refutes the universal graceful-tree assertion. This is distinct from the L2 task, which minimizes the defect of a labeling on one fixed tree.

D19 Total-Coloring Conjecture Counterexample

Major Resolution

Context. A total coloring colors both vertices and edges: adjacent vertices, edges sharing an endpoint, and incident vertex–edge pairs must differ in color. The total-coloring conjecture gives the bound $\Delta(G) + 2$ (Henderschedt et al., 2025).

Question. Construct a graph having no total coloring with $\Delta(G) + 2$ colors.

Submission and verification. Submit `n`, `edges`, and `proof`. A non-null proof is a SHA-256-bound LRAT refutation of the canonical total-coloring CNF; `null` invokes bounded internal exact search. Acceptance requires verified unsatisfiability for the palette size computed from the submitted graph.

Significance. The result refutes the maximum-degree-plus-two bound for simultaneous vertex and edge coloring, not merely the quality of a particular coloring.

D20 Tutte Five-Flow Conjecture Counterexample

Major Resolution

Context. Orient the edges and let B be the signed incidence matrix over $\mathbb{F}_5$. A nowhere-zero five-flow is a vector $f \in \ker B$ with every $f_e \neq 0$; the equations $Bf = 0$ impose conservation at every vertex (Steffen, 2006).

Question. Construct a connected bridgeless graph with no such flow.

Submission and verification. Submit `n` and `edges`. Acceptance checks connectedness and the absence of bridges, then exactly excludes all nowhere-zero solutions of $Bf = 0$ over $\mathbb{F}_5$.

Significance. An accepted graph disproves the universal five-flow assertion within the specified graph interface. The certificate must exclude edge flows satisfying conservation, not vertex colorings.

D21 Berge–Fulkerson Counterexample

Major Resolution

Context. The Berge–Fulkerson conjecture asks for six perfect matchings of every bridgeless cubic graph, covering each edge exactly twice. The matchings form a multiset, so repetition is allowed (Esperet & Mazzuoccolo, 2013).

Question. Construct a bridgeless cubic graph admitting no such six-matching double cover.

Submission and verification. Submit a positive even `n` and the complete `edges` list. Acceptance verifies the graph constraints, enumerates the relevant perfect matchings, and exactly excludes every admissible multiset of six with edge multiplicity two.

Significance. A valid graph refutes the proposed universal perfect-matching cover. Excluding only collections of six distinct matchings is insufficient: repetitions are allowed.

D22 Perfect Difference Set of Non-Prime-Power Order

Major Resolution

Context. For $v = q^2 + q + 1$, a perfect difference set $D \subseteq \mathbb{Z}_v$ has $q+1$ elements and represents each nonzero residue exactly once as an ordered difference $d - d'$ with $d \neq d'$ (Gordon, 1994).

Question. Construct such a D with q not a prime power.

Submission and verification. Submit `order`, a complete `order_factors` list proving that q has at least two distinct prime factors, and `difference_set`. Factor records include the prescribed optional Pocklington certificates when needed. Acceptance verifies the order and all cyclic difference multiplicities exactly.

Significance. The artifact refutes the prime-power restriction for this cyclic design. The factorization must certify the non-prime-power condition.

D23 Ryser Latin-Transversal Counterexample

Major Resolution

Context. A transversal in an order-n Latin square consists of n cells, one in each row and column, containing all n symbols. Ryser's odd-order conjecture predicts such a transversal whenever n is odd (Montgomery, 2024).

Question. Construct a Latin square of odd order with no transversal.

Submission and verification. Submit the complete array in `square`, using symbols $0, \ldots, n-1$. Acceptance checks odd order and the Latin property and exactly excludes every selection satisfying the row, column, and symbol constraints.

Significance. An accepted square refutes the universal odd-order transversal assertion. A partial transversal that fails to extend along one search path does not certify nonexistence.

D24 Union-Closed-Sets Conjecture Counterexample

Major Resolution

Context. A family $\mathcal{F}$ is union-closed when $A \cup B \in \mathcal{F}$ for every $A, B \in \mathcal{F}$. The conjecture predicts that an element of its nonempty union occurs in at least half of its sets (Bruhn & Schaudt, 2015).

Question. Construct a duplicate-free family with union $\{0, \ldots, u-1\}$, $u > 0$, such that

$$|\{A \in \mathcal{F} : i \in A\}| < \frac{|\mathcal{F}|}{2} \qquad \text{for every } i.$$

Submission and verification. Submit `universe_size` and the explicit sets in `family`. Acceptance checks the universe, distinctness, every pairwise union, and every strict frequency inequality. The fixed size and separation requirements of the L2 interface do not apply here.

Significance. One valid family refutes the union-closed half-frequency assertion. A frequency equal to one half does not refute the conjecture.

D25 Lonely-Runner Counterexample with Integer Relative Speeds

Major Resolution

Context. Write $\|x\| = \min_{z\in\mathbb{Z}} |x - z|$. For distinct nonzero integer relative speeds $v_1, \ldots, v_n$, a lonely time satisfies $\|tv_i\| \geq 1/(n+1)$ simultaneously for every i (Rosenfeld, 2025).

Question. Construct speeds with $\sum_i |v_i| \leq 200000$ for which no such time exists; equivalently, for every real t at least one i satisfies $\|tv_i\| < 1/(n+1)$.

Submission and verification. Submit the integer list in `speeds`. Integer periodicity reduces verification to one time period, and the evaluator checks the critical times and intervening cells exactly. An arbitrary finite time grid is not an exact certificate.

Significance. The witness would refute the lonely-runner assertion through an integer-relative-speed instance, with a certificate covering the continuous time domain.

D26 Ryser Hypergraph Counterexample

Major Resolution

Context. In an r-partite r-uniform hypergraph, each edge contains exactly one vertex from each of r parts. Let $\nu(H)$ be the maximum number of vertex-disjoint edges and $\tau(H)$ the minimum vertex cover size (Francetić et al., 2017).

Question. Construct such a hypergraph with

$$\tau(H) > (r-1)\nu(H).$$

Submission and verification. Submit `parts`, specifying the labeled vertex classes, and the explicit `edges`. Acceptance verifies the partite uniform structure and computes both matching and covering optima exactly.

Significance. The result refutes the proposed matching–covering bound for multipartite hypergraphs. Here covers use vertices, unlike the edges meeting triangles in the Tuza task.

D27 Projective Plane of Non-Prime-Power Order

Major Resolution

Context. A projective plane of order q has $v = q^2 + q + 1$ points and lines, each incident with $q + 1$ objects of the other type. Two distinct points determine a unique line, and two distinct lines meet at a unique point (Dembowski, 1968).

Question. Construct such a plane with q not a prime power.

Submission and verification. Submit `order`, its complete `order_factors`, and all point lists in `lines`. At least two distinct prime factors of q must be verified, using the prescribed Pocklington certificates when needed. Acceptance checks all incidence sizes and pairwise uniqueness conditions exactly.

Significance. The construction disproves the prime-power restriction for finite projective planes. A positive-defect incidence structure does not establish existence.

D28 Rota Basis-Rearrangement Counterexample

Major Resolution

Context. Arrange n bases of an n-dimensional vector space as the rows of an array. Rota's basis conjecture asks whether each row can be permuted so that every column is also a basis (Sauermann, 2022).

Question. Construct n bases over $\mathbb{Q}$ or a prime field $\mathbb{F}_p$ for which no such rearrangement exists.

Submission and verification. Submit `field`, specifying the rational or prime field and any required primality certificate, and `bases`, an $n \times n \times n$ array of exact coordinates. Acceptance checks that each input row is a basis and exactly excludes all row-wise rearrangements yielding basis columns.

Significance. The artifact refutes the basis-rearrangement assertion in a representable vector-space instance, and therefore also refutes the general matroid formulation.

D29 Two-Prime Covering Construction (Erdős #203)

Major Resolution

Context. Erdős #203 asks for a positive m coprime to six such that $2^k 3^\ell m + 1$ is composite for every $k, \ell \geq 0$ (Bloom, n.d.a). A finite covering of exponent pairs can certify this infinite assertion.

Question. Construct m and congruence rules (a, M, b, N, p) covering all pairs, where a rule applies when $k \equiv a \pmod{M}$ and $\ell \equiv b \pmod{N}$. Require a prime p with

$$2^M \equiv 3^N \equiv 1 \pmod{p}, \qquad m 2^a 3^b + 1 \equiv 0 \pmod{p},$$

and ensure p is a proper divisor throughout the represented class.

Submission and verification. Submit `m` and at most 10,000 `rules`, with $1 \leq M, N \leq 10^4$ and $p < 2^{63}$. Acceptance checks the modular identities and complete coverage of the joint period grid, whose size is at most 10^6, using at most 5×10^6 coverage-marking operations.

Significance. The finite periodic artifact certifies compositeness for an entire two-parameter exponential family, not merely for a bounded list of exponents.

D30 Prime-Minus-One Covering System (Erdős #273)

Major Resolution

Context. A covering system is a finite family of congruence classes whose union is $\mathbb{Z}$. Erdős #273 restricts the distinct moduli to $p - 1$ with $p \geq 5$ prime (Bloom, n.d.b).

Question. Construct congruences $a_i \pmod{m_i}$ with distinct $4 \leq m_i \leq 10^6$, each $m_i + 1$ prime, that cover every integer.

Submission and verification. Submit `congruences`, containing at most 10,000 records `a`, `m`. The least common multiple $M = \operatorname{lcm}_i m_i$ must not exceed 10^6. Acceptance checks that each $m_i + 1$ is prime and that the moduli are distinct and verifies coverage of every residue modulo M.

Significance. One periodic certificate establishes an infinite covering by this restricted class of arithmetic progressions. Covering a long initial interval without the full period is insufficient.

D31 Exact Coset Cover with Unequal Sizes (Erdős #274)

Major Resolution

Context. For a finite group G, a left coset gH has size $|H|$ and index $[G : H] = |G|/|H|$. The task seeks an exact coset partition with pairwise distinct subgroup indices, equivalently pairwise distinct coset sizes (Bloom, n.d.c).

Question. Construct a permutation group of degree at most seven and more than one left coset forming a pairwise disjoint cover of G with pairwise unequal cardinalities.

Submission and verification. Submit `degree`, `group_generators`, and `cosets`, each with subgroup generators and a representative. The interface permits at most 64 group generators, 128 cosets, and 64 generators per subgroup. Acceptance enumerates the finite groups and cosets and checks subgroup containment, disjointness, full coverage, and unequal sizes.

Significance. The partition refutes the Herzog–Schönheim assertion that an exact coset cover must repeat an index, within the specified finite-group representation.

D32 Nontrivial Collatz Cycle

Broad-Scope Resolution

Context. The ordinary Collatz map is

$$T(n) = \begin{cases} n/2, & n \text{ even}, \\ 3n+1, & n \text{ odd}. \end{cases}$$

Its distinguished positive cycle is $1 \to 4 \to 2 \to 1$ (Lagarias, 1985).

Question. Construct a nonempty cyclic list of pairwise distinct positive integers $c_0, \ldots, c_{m-1}$ such that $T(c_i) = c_{(i+1) \bmod m}$ and its vertex set is not $\{1, 2, 4\}$.

Submission and verification. Submit the complete list in `cycle`. The evaluator checks every exact transition, including the last-to-first transition, and rejects the distinguished cycle.

Significance. A nontrivial positive cycle disproves universal convergence to the distinguished cycle. This interface targets finite cycles, not the separate possibility of an unbounded nonperiodic orbit.

D33 Strong Goldbach Counterexample by Symbolic Coverage

Broad-Scope Resolution

Context. The strong Goldbach conjecture represents every even integer greater than two as a sum of two primes (Oliveira e Silva et al., 2014). For an even candidate $N > 4 \times 10^{18}$, the odd splits are $a = 3 + 2i$ and $N - a$, with $a \le N/2$.

Question. Certify that every such split has a composite summand.

Submission and verification. Submit `N` and a residue-partition tree in `nodes`, rooted at $(M, r) = (1, 0)$. An internal split refines the index class into $(kM, r + jM)$, $2 \le k \le 64$. A leaf names a side and a proper divisor d with $d \mid 2M$, and either $d \mid 3 + 2r$ or $d \mid N - 3 - 2r$. Acceptance verifies complete coverage and that every selected divisor remains proper. The split involving two is excluded because $N - 2$ is even and greater than two.

Significance. The artifact refutes a universal additive statement by a finite certificate covering all prime-pair possibilities, rather than a limited search that happens to find none.

D34 Hadwiger Conjecture Counterexample

Broad-Scope Resolution

Context. Hadwiger's conjecture links chromatic number to complete minors. A K_t minor consists of t disjoint connected branch sets, with an edge joining every pair of branch sets (Seymour, 2016).

Question. Construct a graph G with $\chi(G) \geq t$ but no K_t minor, within $2 \leq t \leq 16$ and $t \leq n \leq 256$.

Submission and verification. Submit `n`, `edges`, `t`, and `proofs`. Two proof entries are mandatory: `noncolorability` and `minor_exclusion`. Each is a SHA-256-bound RUP/LRAT refutation of its canonical CNF. The first excludes a $(t-1)$-coloring; the second excludes a K_t minor. Neither may be replaced by `null`.

Significance. A fully certified graph disproves the universal chromatic-number–minor implication. Both exclusions are necessary; a displayed coloring cannot certify the chromatic lower bound.

D35 Sidorenko Counterexample via a Rational Step Graphon

Broad-Scope Resolution

Context. A finite step graphon consists of positive rational weights $\alpha_1, \ldots, \alpha_k$ summing to one and a symmetric rational matrix $W \in [0,1]^{k \times k}$. Its homomorphism density for H is (Sidorenko, 1991)

$$t(H, W) = \sum_{\phi : V(H) \to \{1,\ldots,k\}} \left(\prod_{v \in V(H)} \alpha_{\phi(v)} \right) \left(\prod_{uv \in E(H)} W_{\phi(u)\phi(v)} \right).$$

Question. Construct a bipartite graph H with at least one edge and a step graphon satisfying $t(H, W) < t(K_2, W)^{|E(H)|}$, where $t(K_2, W) = \sum_{i,j} \alpha_i \alpha_j W_{ij}$.

Submission and verification. Submit the graph in `H` and the rational `weights` and `matrix` in `graphon`. Acceptance verifies the structural constraints and the strict density inequality using exact rational arithmetic.

Significance. An accepted pair disproves Sidorenko's universal bipartite homomorphism-density inequality. A floating-point negative gap is not an exact certificate.