Frontier Math Problem Creation

Rabdos AI Team[1]

November 17, 2025

1 The Context

AI is making remarkable strides in mathematical reasoning with recent breakthroughs demonstrating that LLMs can engage with complex mathematical problems and even contribute to mathematical discovery. Central to realizing its full potential is the creation of abundant high-quality mathematical problems to train and evaluate models. Human-generated datasets remain valuable but relying on humans alone is untenable as models become increasingly capable. AI-driven data generation offers a possible solution, but existing approaches suffer from rapid saturation, limited coverage of mathematical domains, and lack of precise control over difficulty.

2 Our Approach

Rabdos AI creates novel mathematical problems that systematically challenge current frontier models. Our approach combines expert knowledge with automated synthesis tools to produce problems requiring genuine mathematical insight rather than pattern matching or computation alone.

Technical Framework. The creation pipeline integrates human expertise, retrieval systems, large language models, and symbolic verification tools operating at research-grade mathematical rigor. Problems are validated through both computer algebra systems and executable code, ensuring semantic correctness, structural integrity, and computational verifiability.

Coverage and Diversity. Our methodology samples from a broad spectrum of mathematical domains, ensuring that benchmark coverage reflects the genuine breadth of mathematical reasoning rather than concentration in narrow subfields.

Insight Requirements. Unlike problems solvable through direct application of standard algorithms, our creation process enforces non-trivial transformations through deliberate mathematical composition. The resulting problems require solvers to identify and execute sequences of domain-appropriate techniques, with complexity barriers preventing brute-force approaches.

Unambiguous Validation. Our problems are closed-form with unique, rigorously verifiable answers rather than open-ended exercises admitting multiple interpretations. Each problem includes automated verification protocols that eliminate ambiguity in correctness evaluation.

Calibrated Difficulty. The framework supports difficulty calibration through adjustable parameters to enable precise targeting of model capabilities. This permits systematic exploration of the frontier between problems that models can and cannot solve reliably.

3 Evaluation

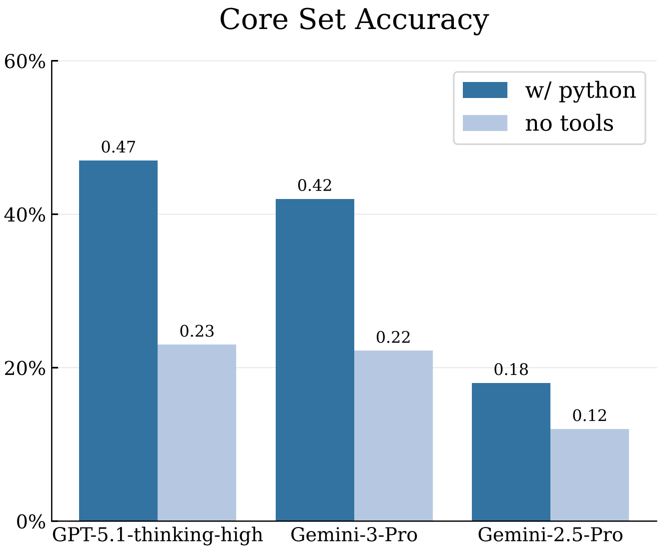
We evaluate on a total of 73 problems created using our approach that break SOTA reasoning models. We partition the problems into two sets, *core* and *diamond*, based on the difficulty of solving them using such models.

[1] Corresponding author: Mayur Naik (mayur@rabdos.ai)

The core set comprises 55 problems that break models GPT-5.1-thinking-high, Gemini-3-Pro, and Gemini-2.5-Pro. The diamond set comprises 18 problems that are even harder and break models GPT-5.1-Pro, Gemini Deep Think, GPT-5.1-thinking-high, and Gemini-3-Pro. All experiments were conducted between November 17-24, 2025.

3.1 Core Set

Figure 1 summarizes the performance of GPT-5.1-thinking, Gemini-3-Pro, and Gemini-2.5-Pro on the core set comprising 55 problems. We report average pass@1 over 8 runs, and for GPT-5.1-thinking we use a high–reasoning-effort configuration with a 20-minute timeout per problem. Both models fail on a large portion of the problems. The sizable gap between pass@1 and pass@8 for both models shows a significant headroom for performing reinforcement learning on the dataset. Even with access to a coding tool, both models still fail on many questions, suggesting that the problems pose genuinely challenging reasoning tasks.



	GPT-5.1-thinking	
	w/ Python	no tools
pass@1	.472	.229
pass@8	.600	.364

	Gemini-3-Pro	
	w/ Python	no tools
pass@1	.423	.222
pass@8	.627	.357

	Gemini-2.5-Pro	
	w/ Python	no tools
pass@1	.184	.120
pass@8	.491	.291

Figure 1: Performance of GPT-5.1-thinking-high, Gemini-3-Pro, and Gemini-2.5-Pro on 55 problems from the core set.

3.2 Diamond Set

Our diamond set comprises 18 problems that require substantially more design effort and are of higher difficulty than the core set. Figure 2 summarizes the performance of GPT-5.1-Pro, Gemini DeepThink, GPT-5.1-thinking-high, and Gemini-3-Pro on the diamond set. For this evaluation, we accessed all four models through the web portal and ran each model once on the full set. Even SOTA reasoning models such as GPT-5.1-Pro and Gemini DeepThink obtain only 2 and 3 of the 18 problems correct, respectively.

4 Team

Our team comprises four researchers at the University of Pennsylvania specializing in mathematics, deep learning, and automated symbolic reasoning.

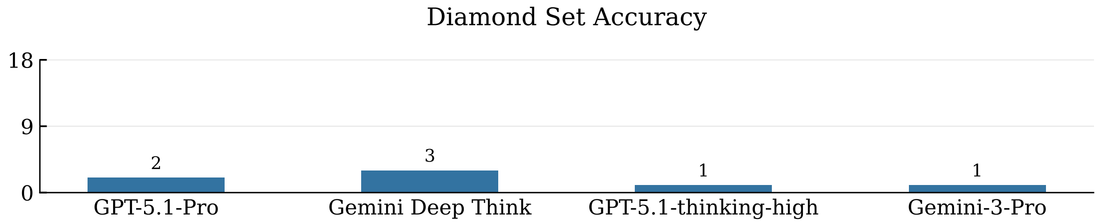


Figure 2: Performance of GPT-5.1-Pro, Gemini Deep Think, GPT-5.1-thinking-high, and Gemini-3-Pro on 18 problems from the diamond set.

Robert Ghrist is Andrea Mitchell University Professor in the departments of Mathematics and Electrical & Systems Engineering. His expertise spans multiple areas of mathematics. His current interests include applications of LLMs to advance mathematics research and education. He serves as Associate Dean of Undergraduate Education in the School of Engineering. He is a recipient of numerous awards in research and teaching, including the Presidential Early Career Award for Scientists and Engineers (PECASE) in 2004, the Chauvenet Prize by the Mathematical Association of America in 2013, and Penn's Lindback Award for Distinguished Teaching in 2015. His online calculus course at Coursera ("Single Variable Calculus") has illustrated the subject to over 100,000 people around the world. He earned his PhD (1995) in applied mathematics from Cornell.

Mayur Naik is Misra Family Professor in the department of Computer and Information Science. His expertise spans the fields of programming languages and artificial intelligence. His current research focuses on neurosymbolic programming, which integrates deep learning with symbolic reasoning to make AI applications safe, interpretable, efficient, and easier to develop. He has published over 80 papers in top-tier conferences in programming languages, deep learning, and related fields. His work has earned four Distinguished Paper awards, three Test-of-Time honors, and four Spotlight Paper recognitions. His online course on software analysis has been taken by over 10,000 students in Georgia Tech's OMSCS program. He obtained his PhD (2008) in computer science from Stanford University.

Zhiqiu (Oscar) Xu is a third year PhD student in the department of Computer and Information Science advised by Prof. Naik. He is an expert in deep learning representations, training algorithms, and benchmarking. He has five publications in top-tier deep learning conferences. He holds bachelors and Masters degrees from UC Berkeley in Computer Science and Applied Mathematics.

Shreya Arya is a postdoctoral scholar and Hans Rademacher Instructor in the department of Mathematics. She is interested in geometry, topology and category theory and its applications to statistics, computer science and mathematical physics. She holds a PhD in Mathematics from Duke University in 2024 where she received the Rudin Prize for Outstanding PhD Dissertation.

A Sample Problems (Core Set)

We present 4 sample problems from the core set. Each problem required an hour of steering by a math expert. For each problem, we provide a ground truth numeric answer and the number of correct runs out of 5 for GPT-5.1-thinking-high and Gemini-2.5-Pro as of November 17, 2025.

A.1 Discrete Geometry

Let H be the set of all hyperplanes in $\mathbb{R}^{48}$ given by $x_i - x_j = 0$ or $x_i - x_j = 1$ for every pair of indices with $1 \le i < j \le 48$. Intersect H with the hyperplane $x_1 + x_2 + \cdots + x_{48} = 0$ to obtain an arrangement inside that 47-dimensional subspace.
Let r be the number of regions of this arrangement, and let b be the number of relatively bounded regions. What is the remainder of $(r - b)$ when divided by 48?

Answer (mod 48**):** $\boxed{2}$

Evaluation: `[Gemini-2.5-Pro: 1/5]` `[GPT-5.1-thinking-high: 1/5]`

A.2 Topology

Let X be the set of real numbers and define

$$A = \left\{ \frac{1}{n} \;\middle|\; n = 1, 2, 3, \dots \right\}.$$

Define a topology τ on X by declaring that a set $O \subset X$ belongs to τ if and only if

$$O = U - B,$$

where U is an open set in the usual Euclidean topology on $\mathbb{R}$ and $B \subset A$.
Determine whether the topological space (X, τ) is **countably paracompact.** Write `1` for true and `0` for false.

Answer: $\boxed{0}$

Evaluation: `[Gemini-2.5-Pro: 3/5]` `[GPT-5.1-thinking-high: 1/5]`

A.3 Enumerative Combinatorics

Consider the set $S = \{0, 2, 3, \dots, 50\}$ (so $|S| = 50$). Count the number F of ordered 25-tuples of distinct elements of S whose sum is congruent to $11 \pmod{23}$. Compute $F \bmod 1031$.

Answer: $\boxed{421}$

Evaluation: `[Gemini-2.5-Pro: 0/5]` `[GPT-5.1-thinking-high: 0/5]`

A.4 Calculus

Evaluate the double integral

$$I = \frac{\pi}{4}\int_0^1\int_0^1 \frac{16\sin^2(\pi s)\sin^2(\pi t) - 16\sin^2(\pi t) + 4}{\left(-8\sin^2(\pi s)\sin^2(\pi t) + 8\sin^2(\pi t) + 1\right)^{3/2}}\, ds\, dt.$$

Give your answer as an exact real number.

Answer: $\boxed{1}$

Evaluation: `[Gemini-2.5-Pro: 0/5]   [GPT-5.1-thinking-high: 3/5]`

B Sample Problems (Diamond Set)

We present three sample problems from the Diamond Set. For each problem, both GPT-5.1-Pro and Gemini DeepThink fail under our single-run evaluation.

B.1 Group Theory

Let n beads be arranged on a circle and indexed $0, 1, \ldots, n-1$. Each bead is colored by an element of the finite field $\mathbb{F}_7 = \mathbb{Z}/7\mathbb{Z}$. Two colorings $(c_0, \ldots, c_{n-1}) \in \mathbb{F}_7^n$ are considered equivalent if one can be obtained from the other by a rotation or reflection of the necklace (i.e. by the usual dihedral action D_n on indices).

A coloring is *admissible* if it satisfies all three linear constraints in $\mathbb{F}_7$:

$$\text{(C0)}\ \sum_{i=0}^{n-1} c_i \equiv 0, \qquad \text{(C1)}\ \sum_{i=0}^{n-1} i\, c_i \equiv 0, \qquad \text{(C2)}\ \sum_{\substack{0\le i\le n-1\\ i\equiv 0 \pmod 3}} c_i \equiv 1.$$

For this problem, take $n = 30$.

How many D_n–equivalence classes of admissible colorings are there for $n = 30$? Report your answer modulo 1,000,003 as a positive integer.

Answer (mod 1,000,003**):** $\boxed{587{,}104}$

Evaluation: `[Gemini Deep Think: False ]   [GPT-5.1-Pro: False ]`

B.2 Probability

Imagine you're studying the geometry of random cubic curves in the projective plane over a finite field.

Over $\mathbb{F}_{101}$ (a prime field with 101 elements), a random homogeneous cubic polynomial in three variables X, Y, Z is given by

$$f(X, Y, Z) = aX^3 + bY^3 + cZ^3 + dX^2Y + eX^2Z + gY^2X + hY^2Z + iZ^2X + jZ^2Y + kXYZ,$$

where each coefficient $a, b, c, d, e, g, h, i, j, k \in \mathbb{F}_{101}$ is chosen independently and uniformly at random (including zero).

Let $N(f)$ be the number of projective points $[X : Y : Z] \in \mathbb{P}^2(\mathbb{F}_{101})$ satisfying $f(X, Y, Z) = 0$. (There are exactly $101^2 + 101 + 1 = 10303$ projective points in total.)

We say the random cubic hypersurface is *balanced* if

$$\boxed{N(f) \equiv 1 \pmod 3}.$$

Compute the exact probability that a random cubic is balanced. Express the answer as a reduced fraction $\frac{r}{s}$ in lowest terms.

Answer: $\boxed{3423841/1030301}$

Evaluation: `[Gemini Deep Think: False ]` `[GPT-5.1-Pro: False ]`

B.3 Epistemic Logic

Consider a multi-agent epistemic model with three agents A, B, C operating in the multi-agent logic $\mathsf{S5}$. The set of possible worlds is $W = \mathbb{F}_2^{12}$, the set of all 12-bit binary strings $x = (x_1, x_2, \dots, x_{12})$ with arithmetic modulo 2 (XOR). The designated actual world is $w_\star = 0$ (the all-zeros string). Each agent i has an equivalence relation R_i on W representing indistinguishability. We define these via linear subspaces of $\mathbb{F}_2^{12}$:

Agent A: Let $U_A = \operatorname{span}\{e_1, e_2\}$ where e_j is the jth standard basis vector. Define $x\, R_A\, y$ iff $x - y \in U_A$ (equivalently, x and y differ only in coordinates 1 and 2).

Agent B: Let $U_B = \operatorname{span}\{e_3, e_4\}$. Define $x\, R_B\, y$ iff $x - y \in U_B$.

Agent C: Let $U_C = \operatorname{span}\{e_1 + e_3,\ e_2 + e_4\}$. Define $x\, R_C\, y$ iff $x - y \in U_C$.

Epistemic operators. Write $K_i\varphi$ for "agent i knows φ," $EG\varphi := K_A\varphi \wedge K_B\varphi \wedge K_C\varphi$ for "everyone knows φ," and $CG\varphi$ for "common knowledge of φ among $G = \{A, B, C\}$." Common knowledge $CG\varphi$ holds at world w iff φ holds at every world reachable from w by any finite sequence of steps along $R_A \cup R_B \cup R_C$.

Component structure and tags. Let $H = U_A + U_B + U_C$ (the subspace sum, computed with XOR addition). The connected components of the undirected graph with edge set $R_A \cup R_B \cup R_C$ are precisely the cosets $x + H$ for $x \in \mathbb{F}_2^{12}$. For each component X, define its tag as the 8-bit integer formed by the last eight coordinates of any world in X:

$$\operatorname{tag}(X) := \text{binary value of } (x_5 x_6 x_7 x_8 x_9 x_{10} x_{11} x_{12}) \in \{0, 1, \dots, 255\}.$$

This is well-defined since H acts only on the first four coordinates. We consider valuations V assigning truth values to three propositional atoms p, q, r at each world.

A valuation is *admissible* iff it satisfies two conditions:
(F1) At $w_\star$: $\mathcal{M}, w_\star \models CG(EGp \wedge EG\neg q)$.
(F2) Among all components $X \neq X_\star$ (where $X_\star$ is the component containing $w_\star$), call X *good* if for any (equivalently, every) $x \in X$:

$$\mathcal{M}, x \models CGEGr \quad \wedge \quad CGEG(p \leftrightarrow r \oplus \operatorname{parity}(\operatorname{tag}(X))),$$

where $\operatorname{parity}(t)$ is the sum of bits in the binary representation of t modulo 2, and $\oplus$ denotes XOR. Let $S = \{X \neq X_\star : X \text{ is good}\}$. Then,

$$|S| \equiv 7 \pmod{13}, \qquad \sum_{X \in S} \operatorname{tag}(X) \equiv 45 \pmod{97}.$$

Count the number of admissible valuations. Report your answer modulo 10,007.

Answer (mod 10,007**):** $\boxed{4{,}814}$

Evaluation: `[Gemini Deep Think: False ]` `[GPT-5.1-Pro: False ]`