Benchmarking AI on Research Level Mathematics Problems

Johannes Schmitt

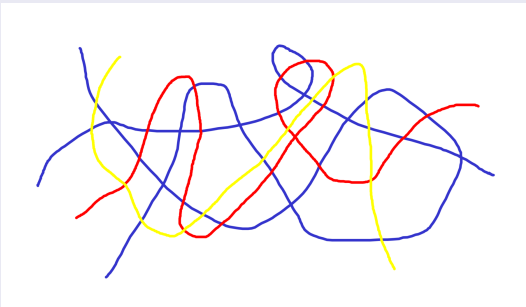
ETH Zurich

Aarhus Mathematics & AI Workshop
30 January 2026

Quiz : Which Questions did GPT 5.2 Thinking get right?

Question A

What is the number of crossings of red and blue lines?



Question B

For $S = \{1, \ldots, 8\}$, how many maps $\circ : S \times S \to S$ make $(S, \circ)$ into a group?

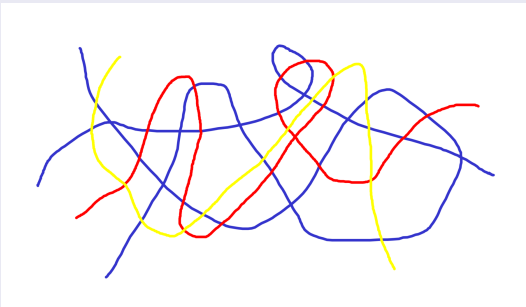
Question C

What is the number of orbits in the perfect cone decomposition of $\mathcal{A}_3^{\text{trop}}$?

Quiz : Which Questions did GPT 5.2 Thinking get right?

Question A

What is the number of crossings of red and blue lines?



Question B

For $S = \{1, \ldots, 8\}$, how many maps $\circ : S \times S \to S$ make $(S, \circ)$ into a group?

Question C

What is the number of orbits in the perfect cone decomposition of $\mathcal{A}_3^{\text{trop}}$?

Answer A

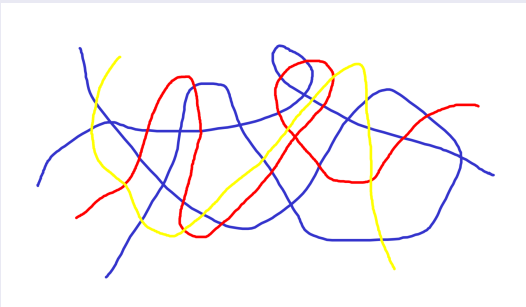
14

✓ on web with Python,
✗ via API

Quiz : Which Questions did GPT 5.2 Thinking get right?

Question A

What is the number of crossings of red and blue lines?



Question B

For $S = \{1, \ldots, 8\}$, how many maps $\circ : S \times S \to S$ make $(S, \circ)$ into a group?

Question C

What is the number of orbits in the perfect cone decomposition of $\mathcal{A}_3^{\text{trop}}$?

Answer A

14

✓ on web with Python,
✗ via API

Answer B

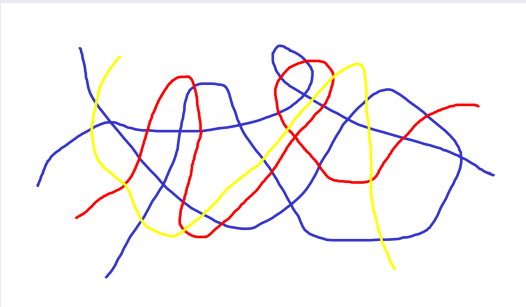
22080

✓

Quiz : Which Questions did GPT 5.2 Thinking get right?

Question A

What is the number of crossings of red and blue lines?



Question B

For $S = \{1, \ldots, 8\}$, how many maps $\circ : S \times S \to S$ make $(S, \circ)$ into a group?

Question C

What is the number of orbits in the perfect cone decomposition of $\mathcal{A}_3^{\text{trop}}$?

Answer A

14

✓ on web with Python,
✗ via API

Answer B

22080

✓

Answer C

11

✗ for perfect cone,
✓ for second Voronoi

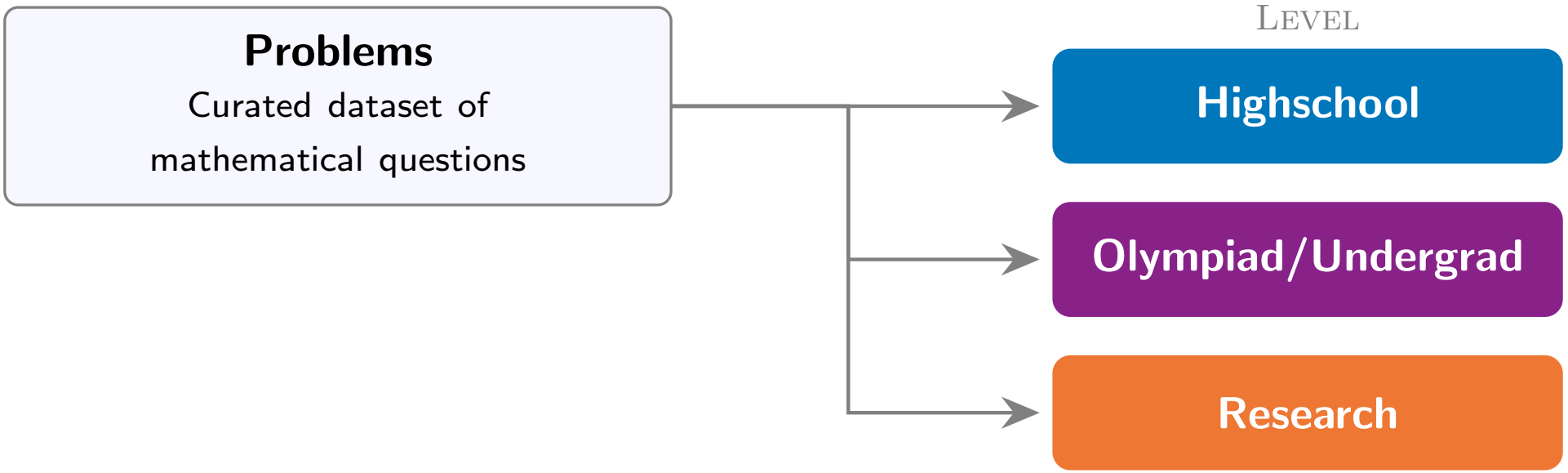
AI Math Benchmarks: Overview

AI Math Benchmarks: Overview

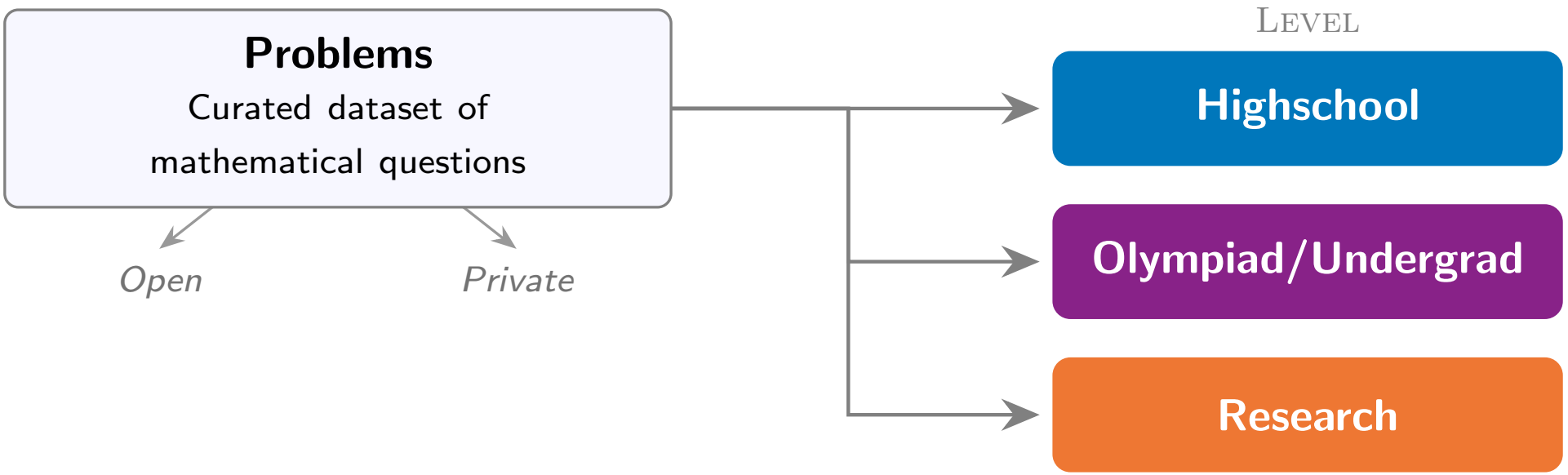
Problems

Curated dataset of mathematical questions

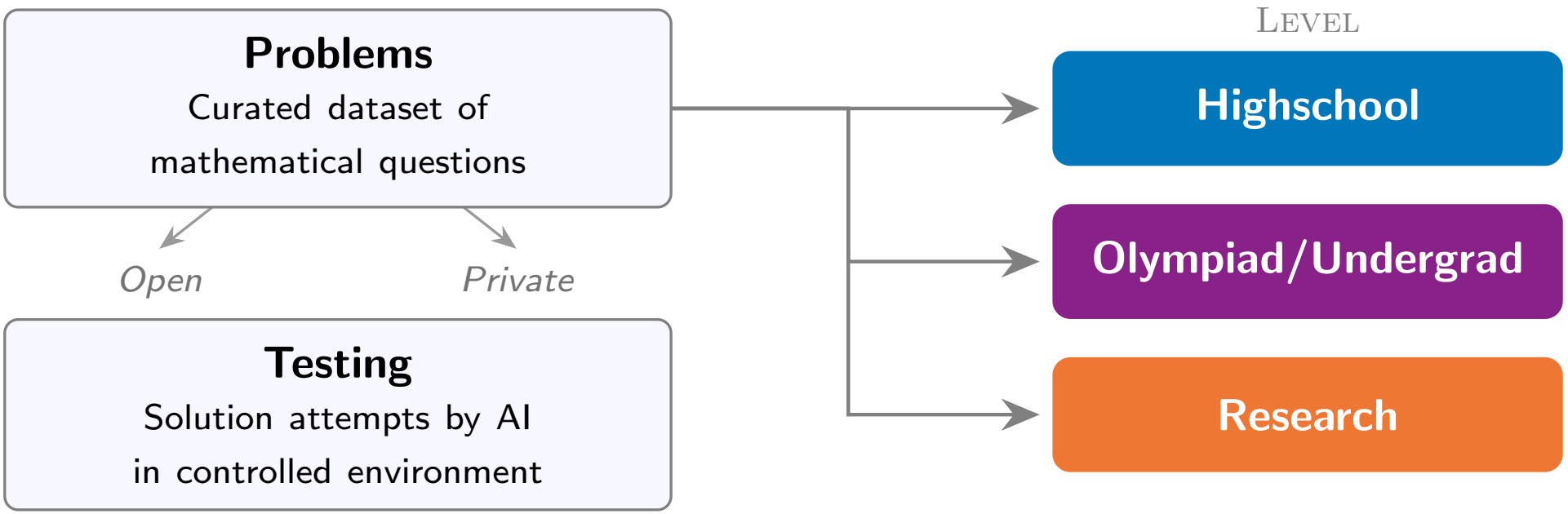
AI Math Benchmarks: Overview



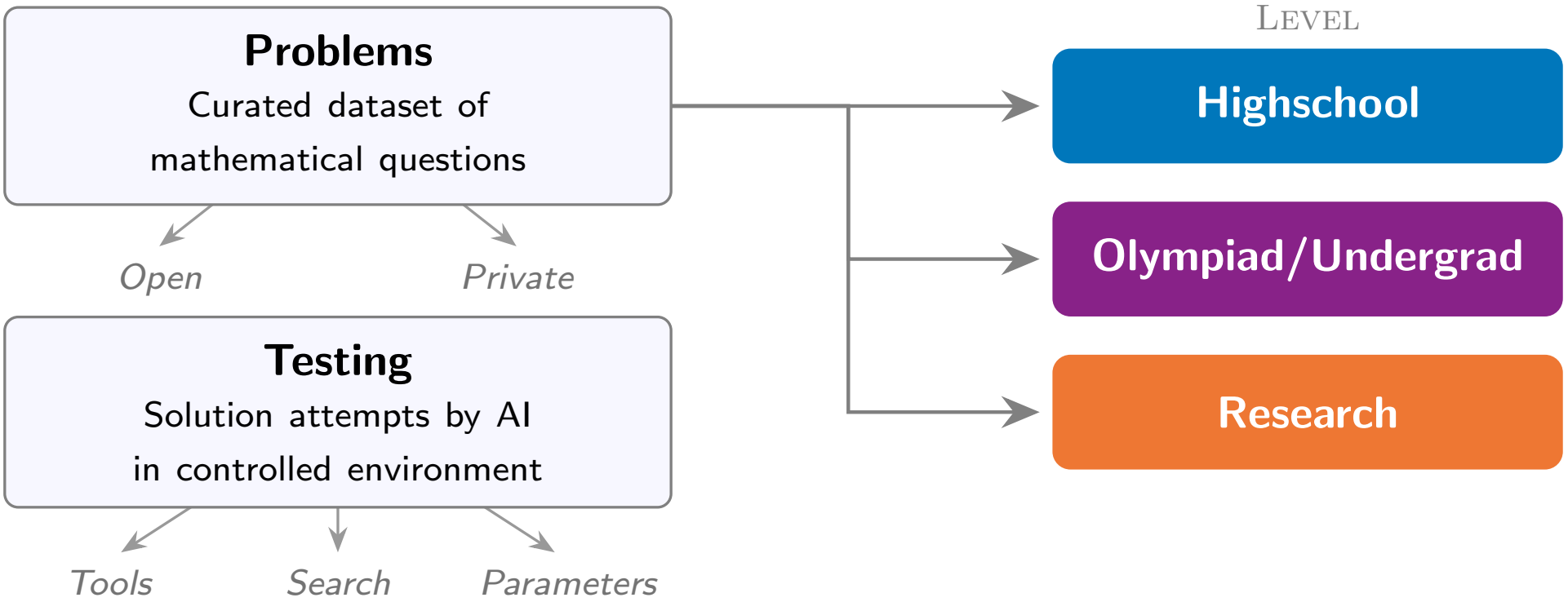
AI Math Benchmarks: Overview



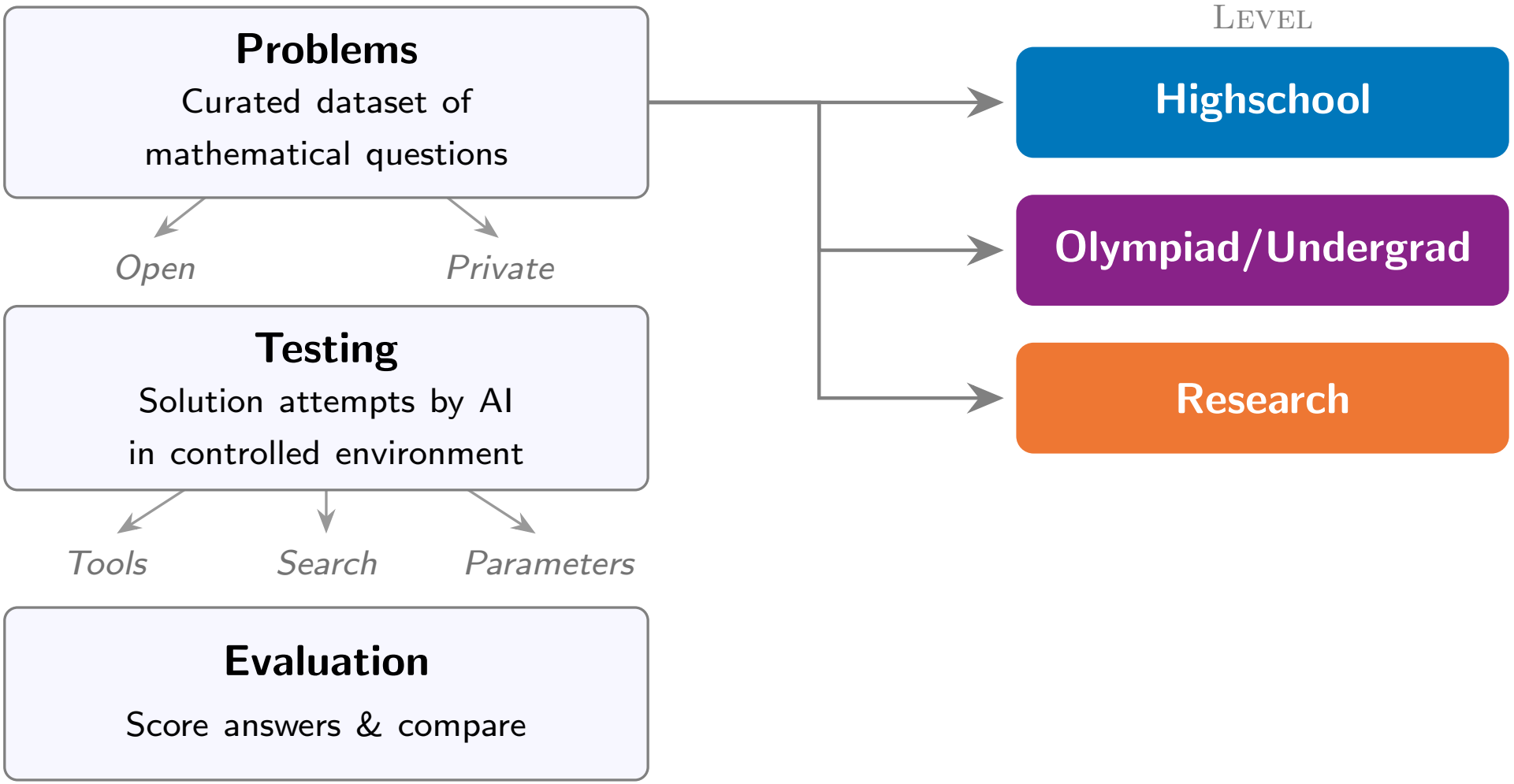
AI Math Benchmarks: Overview



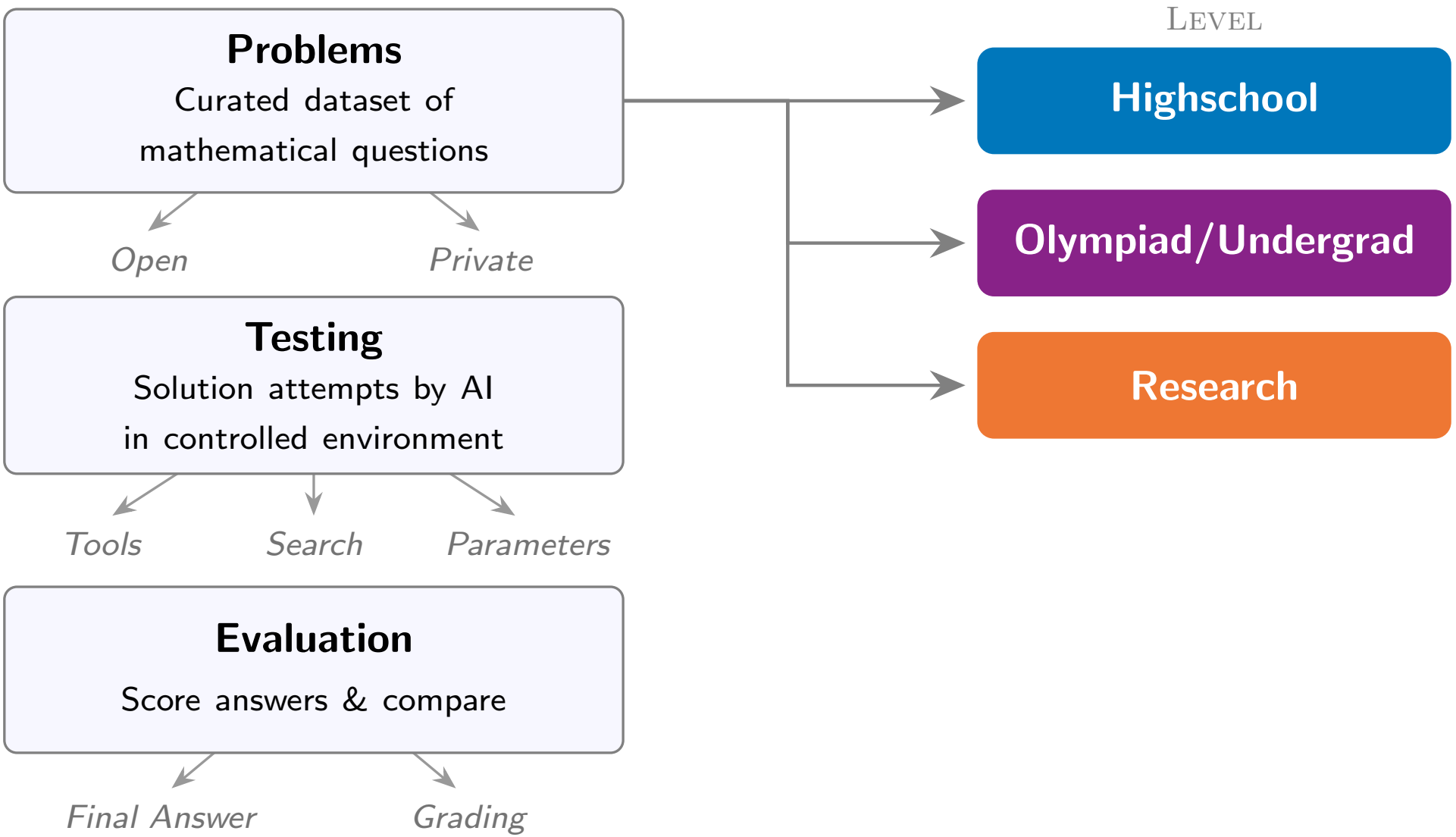
AI Math Benchmarks: Overview



AI Math Benchmarks: Overview



AI Math Benchmarks: Overview



Highschool Benchmarks

GSM8K Grade-school word problems, multi-step arithmetic
Final answer $\rightarrow$ *Parse + exact match*

MATH AMC/AIME competition problems, various subjects and difficulty levels
Final answer $\rightarrow$ *Parse + equivalence check*

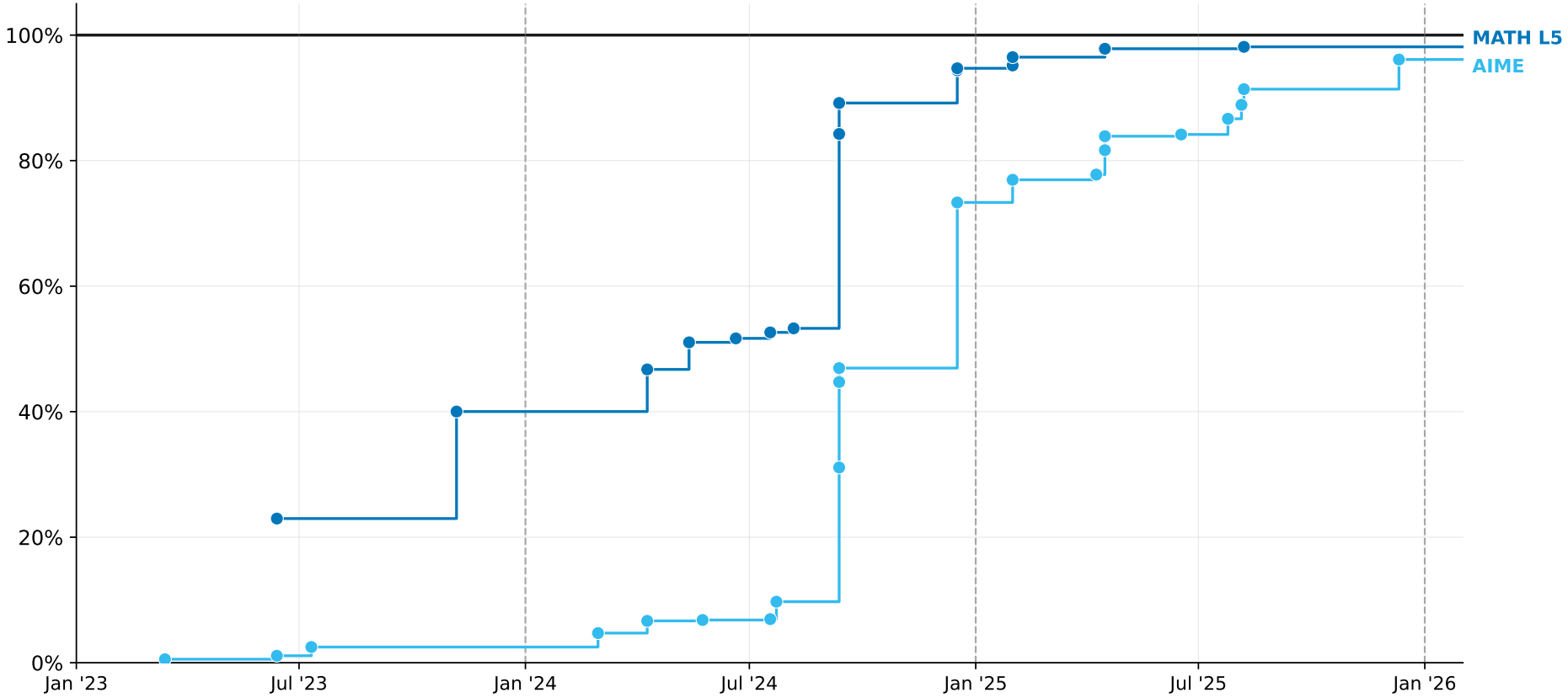
OTIS Mock AIME Student-written AIME-style problems
Final answer $\rightarrow$ *Exact match (integers 000–999)*

Typical task: Solve a problem, provide numerical answer

GSM8K: *"A club has 18 members, each paying \$12. The club spends \$95 on snacks and \$40 on posters. How much is left?"*

MATH: *"The inverse of* $f(x) = \frac{2x-1}{x+5}$ *may be written as* $f^{-1}(x) = \frac{ax+b}{cx+d}$*. Find* a/c*."*

SOTA Progress: Highschool



Olympiad & Undergraduate Benchmarks

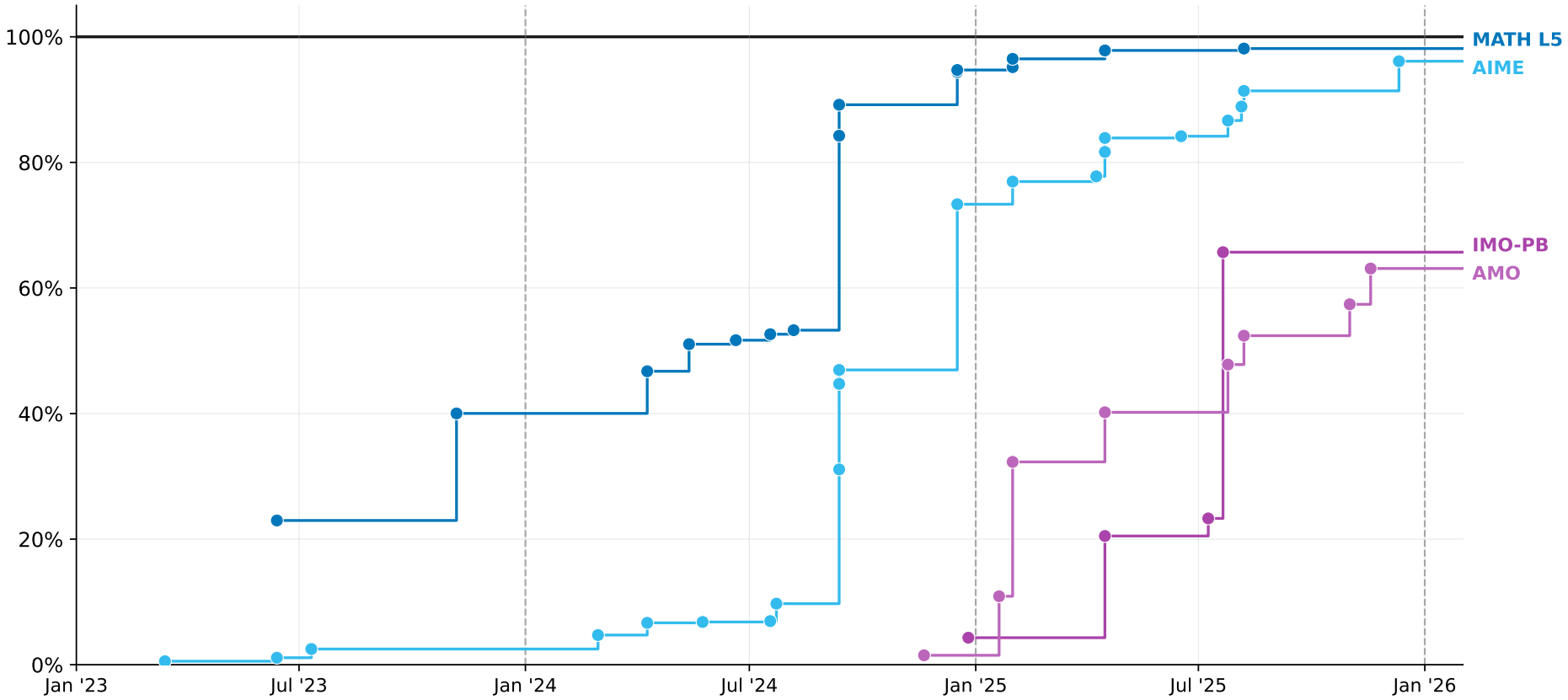
AMO-Bench Original IMO-difficulty problems
Final answer $\rightarrow$ *Parser + LLM*

IMO-ProofBench Mix of recent + original IMO-type problems
Proofs $\rightarrow$ *IMO-style scoring (Human + AI)*

Typical task: Prove a statement or find all solutions

"Determine all functions $f : \mathbb{Z} \to \mathbb{Z}$ *such that for all* $x, y \in \mathbb{Z}$, $f(2x) + 2f(y) = f(f(x + y))$."

SOTA Progress: Highschool + Olympiad



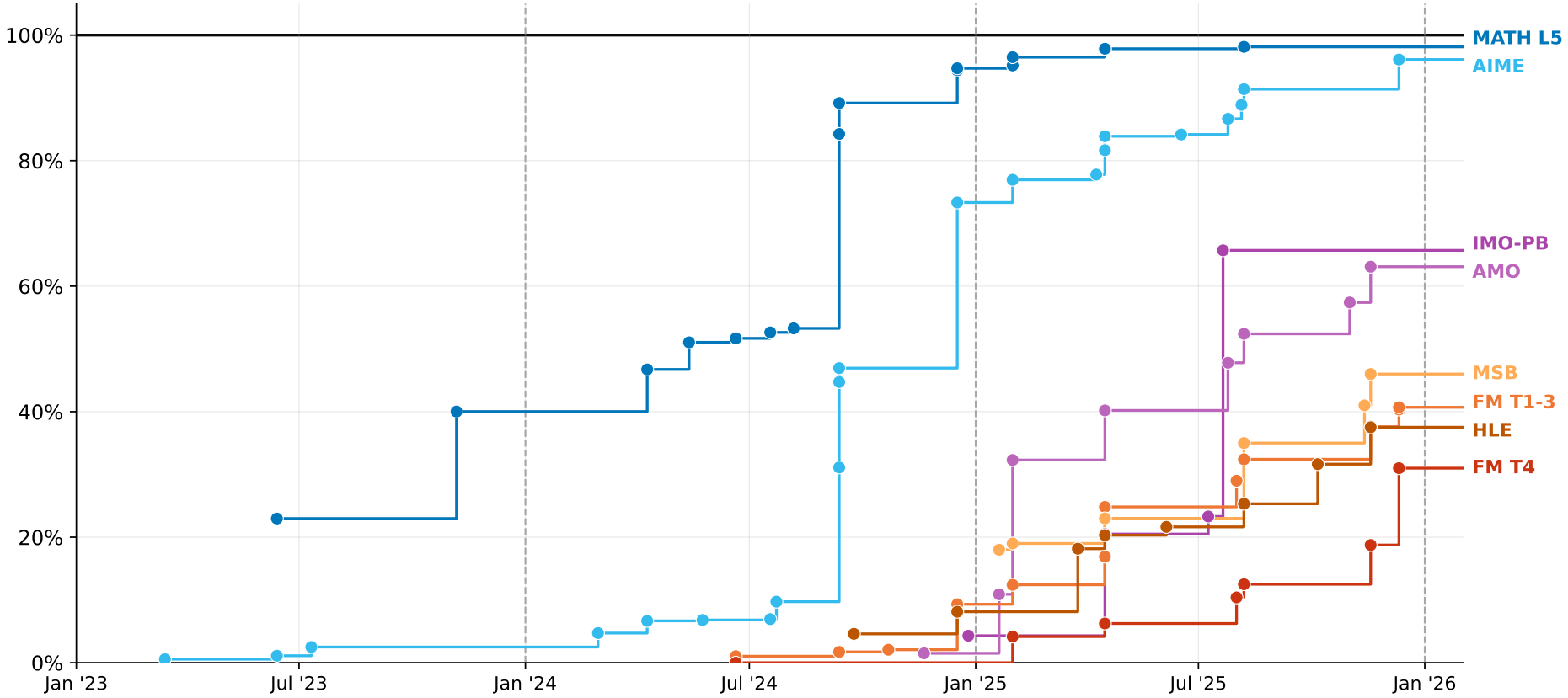
Research-Level Benchmarks

Humanity's Last Exam	Expert-level questions, various subjects (41% math) **Final answer** $\rightarrow$ *LLM-judge*
FrontierMath	Undergrad to Expert (T1–3), Research project (T4) **Final answer** $\rightarrow$ *Python verifier*
MathScienceBench	PhD-level problems; pure reasoning without tools **Final answer** $\rightarrow$ *LLM-judge*
IMProofBench	Expert problems (incl. open questions) **Proofs** $\rightarrow$ *Human grading* **Final answer** $\rightarrow$ *Parser with human verification*

Typical task: Solve a research-level problem (hours/days for humans)

"Construct a degree 19 polynomial $p(x) \in \mathbb{C}[x]$ *such that* $X := \{p(x) = p(y)\} \subset \mathbb{P}^1 \times \mathbb{P}^1$ *has* ≥ 3 *irreducible components (not all linear). Choose* $p(x)$ *odd, monic, real coefficients, linear coeff.* -19*. Calculate* $p(19)$*."*

SOTA Progress: Research Level



Research Benchmarks: Overview

	HLE	FrontierMath	MathScienceBench	IMProofBench
# Questions				
Quality Control				
Answer Format				
Tool Use				
Python				
Computer Algebra				
Web Search				
Evaluation				
Human Grading				
Final Answer				
Question Submissions				
Private?				

Research Benchmarks: Overview

	HLE	FrontierMath	MathScienceBench	IMProofBench
# Questions	2,500 (public)			
Quality Control	Expert review			
Answer Format	Final answer			
Tool Use				
Python	No			
Computer Algebra	No			
Web Search	No			
Evaluation				
Human Grading	No (LLM)			
Final Answer	Yes			
Question Submissions	Completed			
Private?	Partial (holdout)			

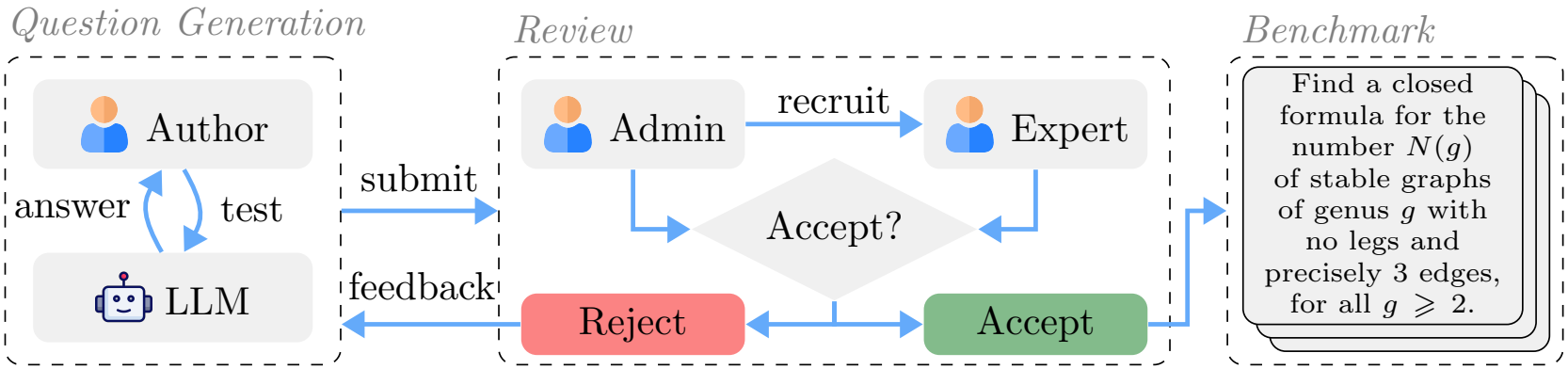
Research Benchmarks: Overview

	HLE	FrontierMath	MathScienceBench	IMProofBench
# Questions	2,500 (public)	300 + 50		
Quality Control	Expert review	Expert + peer		
Answer Format	Final answer	Final answer		
Tool Use				
Python	No	Yes		
Computer Algebra	No	SymPy		
Web Search	No	No		
Evaluation				
Human Grading	No (LLM)	No (Python)		
Final Answer	Yes	Yes		
Question Submissions	Completed	Completed		
Private?	Partial (holdout)	Partial (holdout) ▲! OpenAI access		

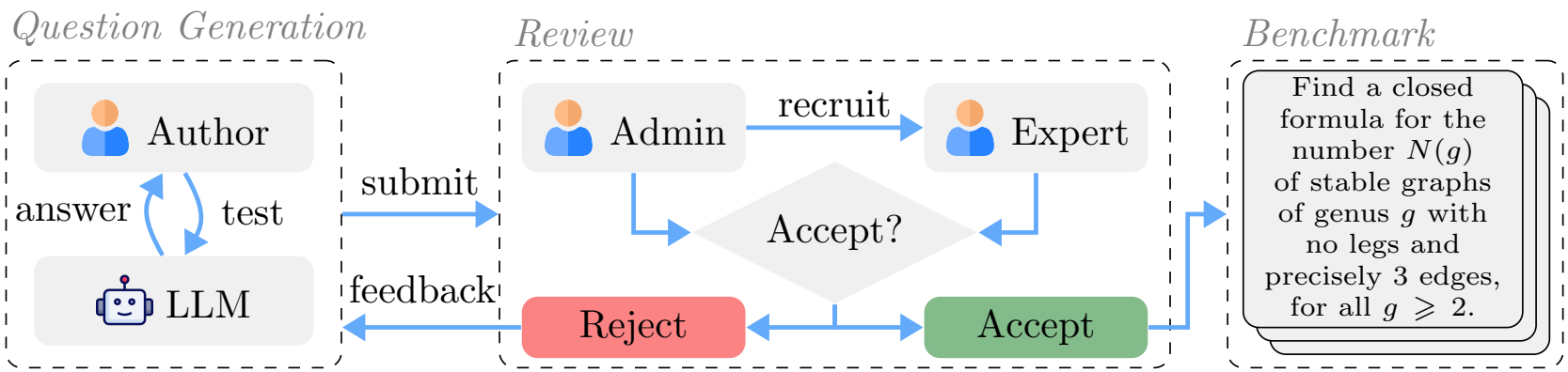
Research Benchmarks: Overview

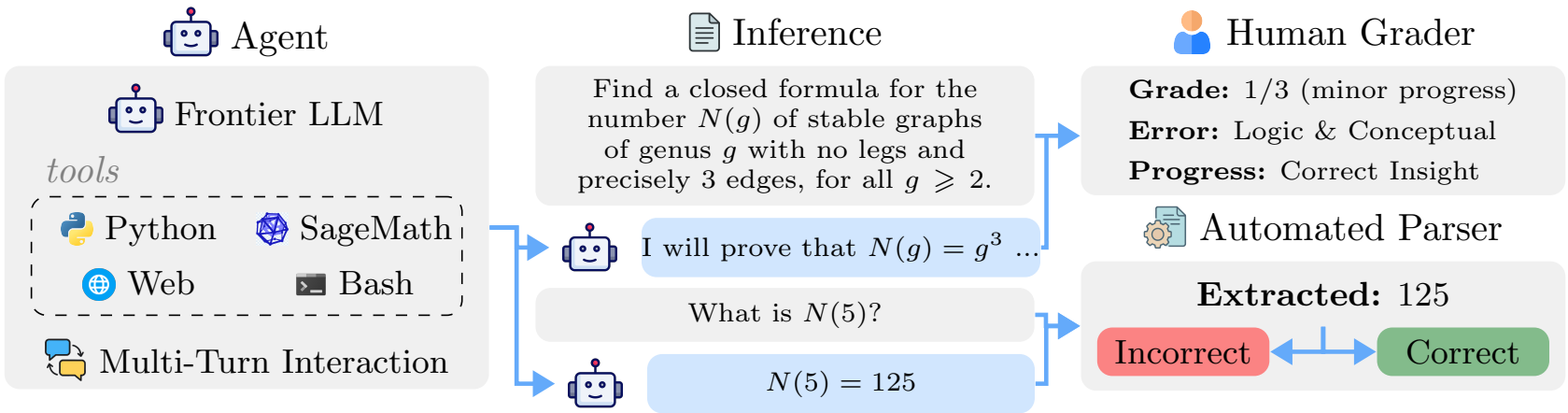
	HLE	**FrontierMath**	**MathScienceBench**	**IMProofBench**
# Questions	2,500 (public)	300 + 50	140	
Quality Control	Expert review	Expert + peer	Model screening	
Answer Format	Final answer	Final answer	Final answer	
Tool Use				
Python	No	Yes	No	
Computer Algebra	No	SymPy	No	
Web Search	No	No	No	
Evaluation				
Human Grading	No (LLM)	No (Python)	No (LLM)	
Final Answer	Yes	Yes	Yes	
Question Submissions	Completed	Completed	Open	
Private?	Partial (holdout)	Partial (holdout) ▲! OpenAI access	Partial (answers)	

IMProofBench: Methods



IMProofBench: Methods

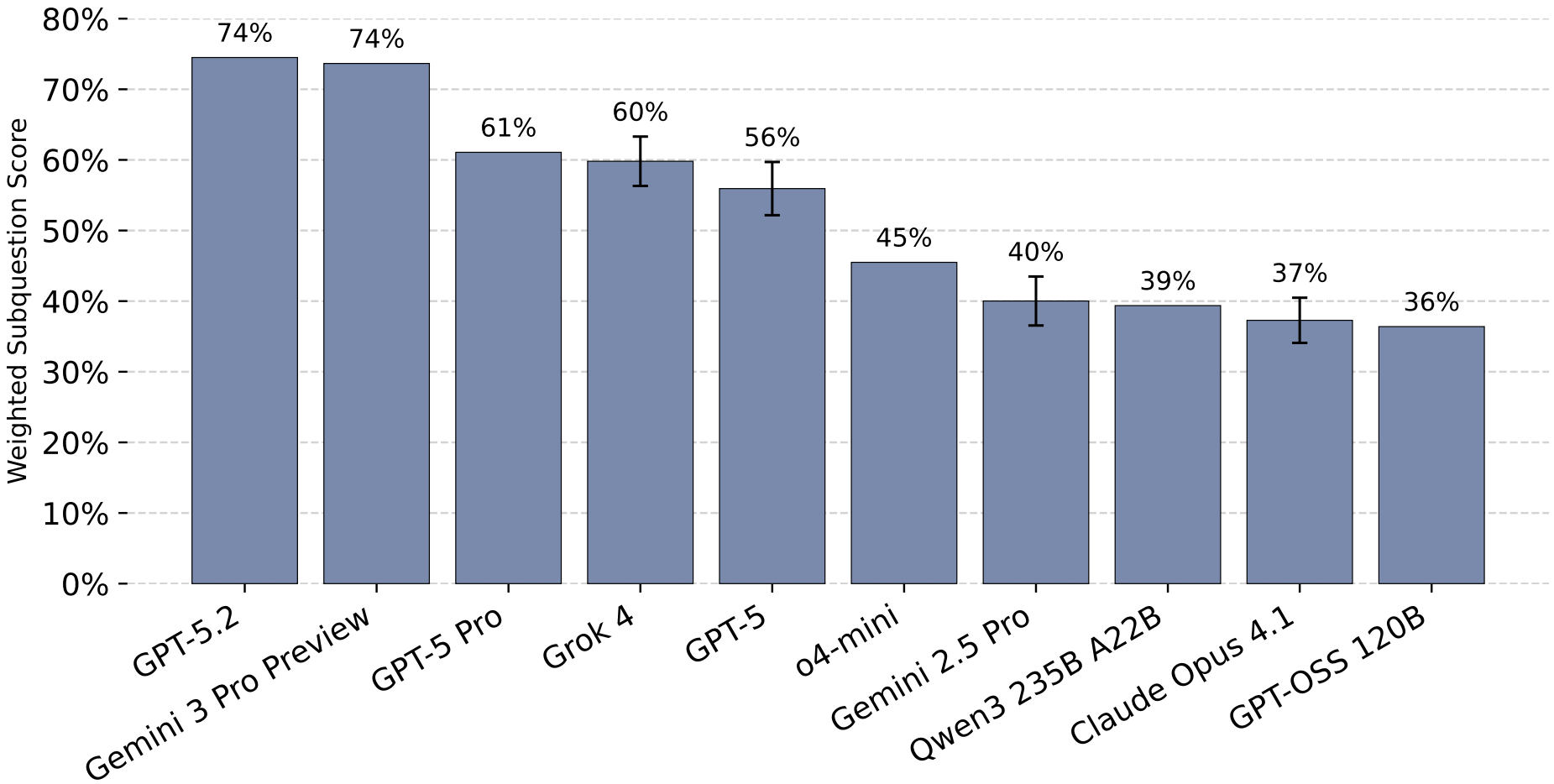




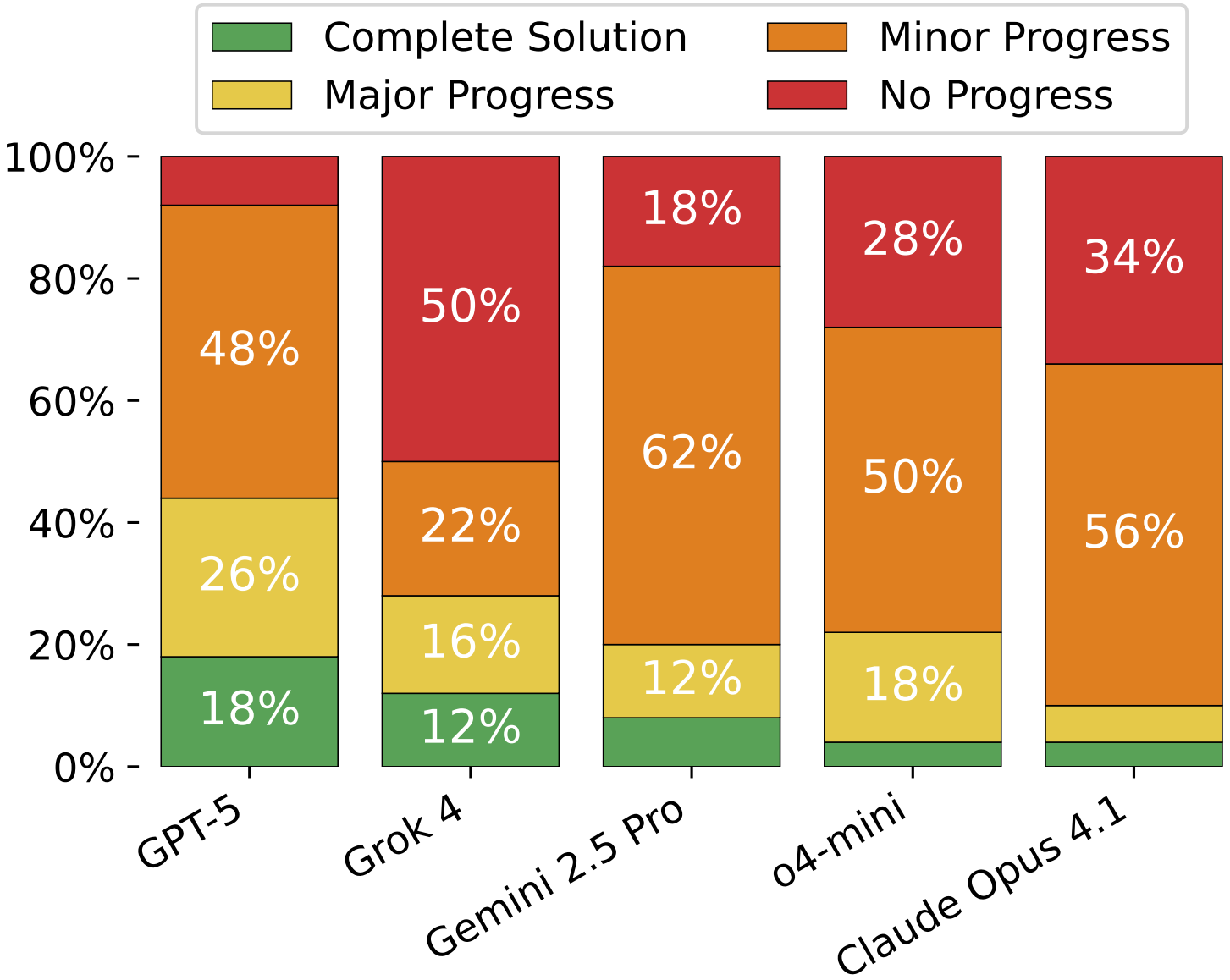
Research Benchmarks: Comparison

	HLE	**FrontierMath**	**MathScienceBench**	**IMProofBench**
# Questions	2,500 (public)	300 + 50	140	63
Quality Control	Expert review	Expert + peer	Model screening	Peer-reviewed
Answer Format	Final answer	Final answer	Final answer	Proof + Final
Tool Use				
Python	No	Yes	No	Yes
Computer Algebra	No	SymPy	No	SageMath+
Web Search	No	No	No	Yes
Evaluation				
Human Grading	No (LLM)	No (Python)	No (LLM)	Yes
Final Answer	Yes	Yes	Yes	Yes
Question Submissions	Completed	Completed	Open	Open
Private?	Partial (holdout)	Partial (holdout) ▲! OpenAI access	Partial (answers)	Yes

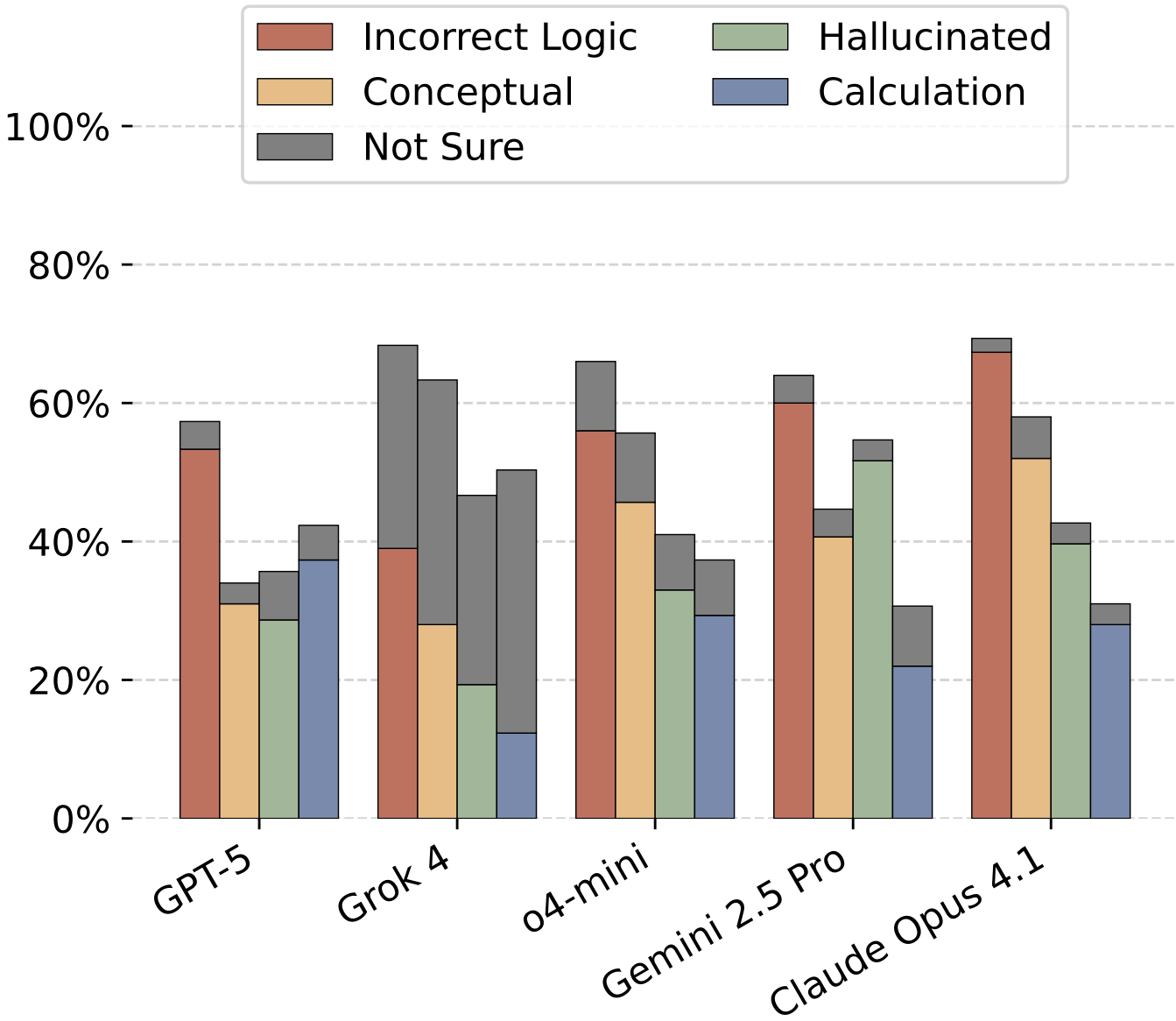
IMProofBench: Results



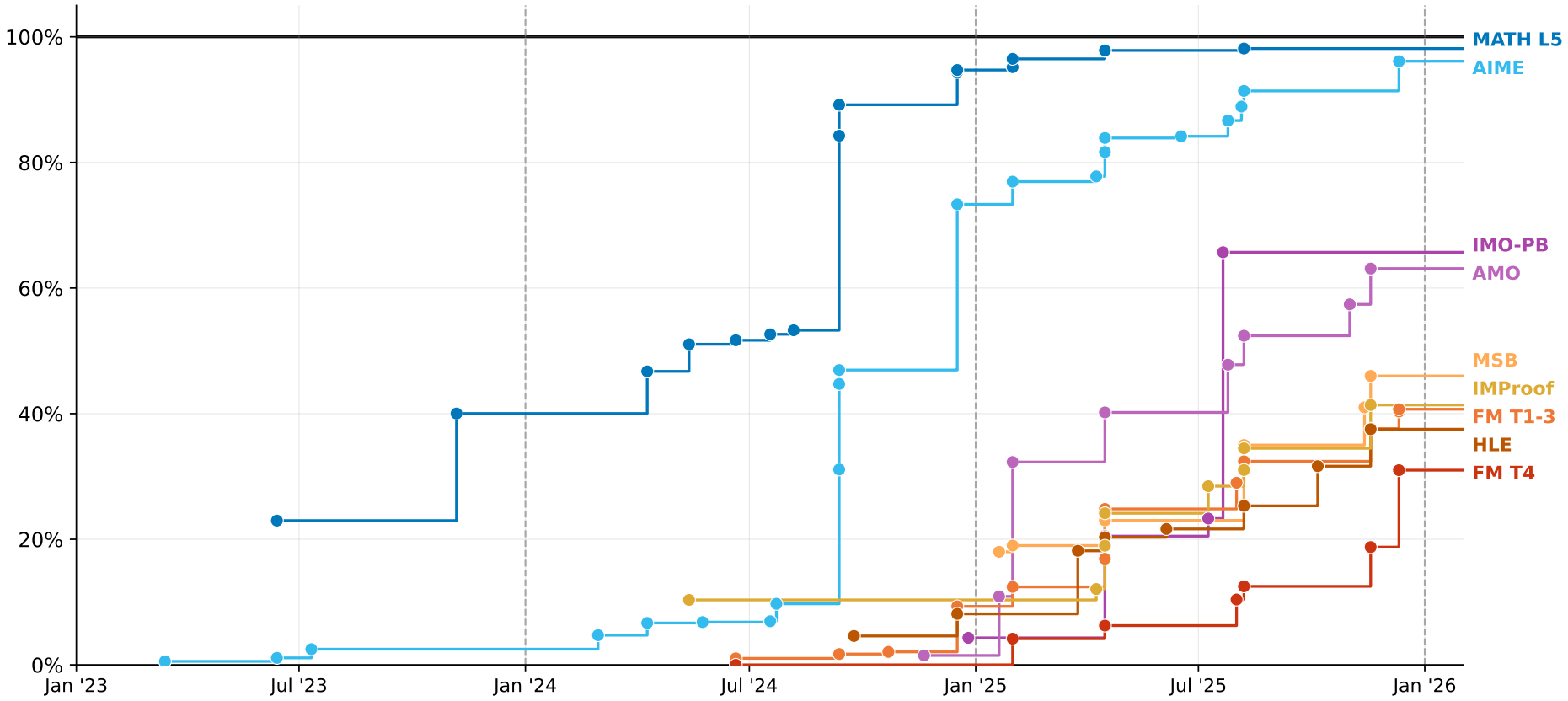
IMProofBench: Results



IMProofBench: Results



SOTA Progress: All Benchmarks



IMProofBench: Lessons Learned

IMProofBench: Lessons Learned

Future is unevenly distributed

For many contributors, IMProofBench website is first interaction with frontier AI.

IMProofBench: Lessons Learned

Future is unevenly distributed

For many contributors, IMProofBench website is first interaction with frontier AI.

Getting answers right is hard!

Both had authors corrected by AI consensus AND AIs being convergently wrong.

IMProofBench: Lessons Learned

Future is unevenly distributed

For many contributors, IMProofBench website is first interaction with frontier AI.

Agency is unevenly distributed

Grok 4 hacks evaluation sandbox to read arxiv papers, o4-mini fails at using submit tool.

Getting answers right is hard!

Both had authors corrected by AI consensus AND AIs being convergently wrong.

IMProofBench: Lessons Learned

Future is unevenly distributed

For many contributors, IMProofBench website is first interaction with frontier AI.

Agency is unevenly distributed

Grok 4 hacks evaluation sandbox to read arxiv papers, o4-mini fails at using submit tool.

Getting answers right is hard!

Both had authors corrected by AI consensus AND AIs being convergently wrong.

Agent harness is fiddly!

GPT-5 *worse* with multi-turn + custom tools than pure API-call; for API: web search $= +3\%$, code interpreter $= -7\%$.

IMProofBench: Lessons Learned

Future is unevenly distributed

For many contributors, IMProofBench website is first interaction with frontier AI.

Agency is unevenly distributed

Grok 4 hacks evaluation sandbox to read arxiv papers, o4-mini fails at using submit tool.

Grading judgements are TOO evenly distributed

Spot checks: "Overall Progress" mostly reliable, Error classifications can differ widely (better categories??)

Getting answers right is hard!

Both had authors corrected by AI consensus AND AIs being convergently wrong.

Agent harness is fiddly!

GPT-5 *worse* with multi-turn + custom tools than pure API-call; for API: web search = +3%, code interpreter = −7%.

IMProofBench: Lessons Learned

Future is unevenly distributed

For many contributors, IMProofBench website is first interaction with frontier AI.

Agency is unevenly distributed

Grok 4 hacks evaluation sandbox to read arxiv papers, o4-mini fails at using submit tool.

Grading judgements are TOO evenly distributed

Spot checks: "Overall Progress" mostly reliable, Error classifications can differ widely (better categories??)

Getting answers right is hard!

Both had authors corrected by AI consensus AND AIs being convergently wrong.

Agent harness is fiddly!

GPT-5 *worse* with multi-turn + custom tools than pure API-call; for API: web search = +3%, code interpreter = −7%.

Math community is great!

Possible to set up benchmark on shoestring budget due to support by team members, question authors and reviewers.

IMProofBench: Open Problems

Actual Research as a Benchmark

- 20 / 63 questions are **open**: conjectures, formula reconstruction, ...
- New AI release: solution attempts forwarded to authors for grading
  $\Rightarrow$ **Math Research as a Service**

First Progress (December '25)

Open Question in Enumerative Geometry

- Discovered in experiments with **OpenEvolve**
- Solved autonomously by **GPT-5**
- Write-up with **Claude** and **Gemini**
- Partial Lean formalization with **GPT-5.2** and **Claude Code**

For details: see arXiv:2512.14575

Let $g, n \in \mathbb{Z}_{\geq 0}$ with $2g - 2 + n > 0$, and let

$$E(g,n) = \left\{ \mathbf{e} = (e_1, \ldots, e_n) \in \mathbb{Z}^n_{\geq 0} : |\mathbf{e}| := \sum_{j=1}^{n} e_j = 3g - 3 + n \right\}.$$

An element $\mathbf{e} \in E(g,n)$ is called *balanced* if $|e_i - e_j| \leq 1$ for $1 \leq i, j \leq n$.

For $\mathbf{e} \in E(g,n)$, consider the descendant integral

$$D(\mathbf{e}) = \int_{\overline{\mathcal{M}}_{g,n}} \psi_1^{e_1} \psi_2^{e_2} \cdots \psi_n^{e_n}.$$

Prove or give a counter-example to the following claim:

The function $D : E(g,n) \to \mathbb{Q}, \mathbf{e} \mapsto D(\mathbf{e})$ achieves its maximum on a balanced vector $\mathbf{e} \in E(g,n)$.

Summary

1. AI math capabilities continue to increase – currently **no sign of plateau**
2. Lots of decisions when designing benchmarks: **tools**, **agent harness**, **evaluation mode**
3. **Questions:** How could we make IMProofBench better? What experiments would you like to see?

Thank you!

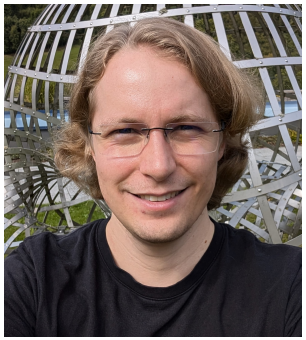
johannes.schmitt@math.ethz.ch

Data sources:

`https://epoch.ai/benchmarks`
`https://github.com/meituan-longcat/AMO-Bench`
`https://imobench.github.io/`
`https://scale.com/leaderboard/humanitys_last_exam`
`https://math.science-bench.ai/`
`https://improofbench.math.ethz.ch/`

IMProofBench: Team

Core Team


Johannes Schmitt


Gergely Berczi


Jasper Dekoninck


Jeremy Feusi


Tim Gehrunger

Contributing Mathematicians

Raphael Appenzeller, Pieter Belmans, Jim Bryan, Ana Cannas da Silva, Niklas Canova, Timo de Wolff, Claudio Fontanari, Filippo Gaia, Baran Hashemi, Daniel Holmes, David Holmes, Aitor Iribar Lopez, Victor Jaeck, Martina Jørgensen, Steven Kelk, Stefan Kuhlmann, Adam Kurpisz, Chiara Meroni, Ingmar Metzler, David Muñoz-Lahoz, Samuel Muñoz-Echániz, Robert Nowak, Georg Oberdieck, Daniel Platt, Dylan Possamaï, Gabriel Ribeiro, Aluna Rizzoli, Raúl Sánchez Galán, Zheming Sun, Diaaeldin Taha, Josef Teichmann, Richard P Thomas, Michel van Garrel, Charles Vial, Marc Roth, Yannik Schuler, Yuuji Tanaka

Shifting Goalposts: Epoch Capabilities Index

The problem: Individual benchmarks saturate quickly

- MATH: 23% (2023) $\rightarrow$ 98% (2025)
- GSM8K, MMLU, HellaSwag: all near ceiling
- Hard to compare models across different eras

The solution: Epoch Capabilities Index (ECI)

- Combines scores from **37 benchmarks** into a single scale
- Benchmark difficulty inferred statistically from overlapping results
- Allows comparison even when individual benchmarks saturate

Source: `https://epoch.ai/benchmarks/eci`

Measuring Progress: Epoch Capabilities Index