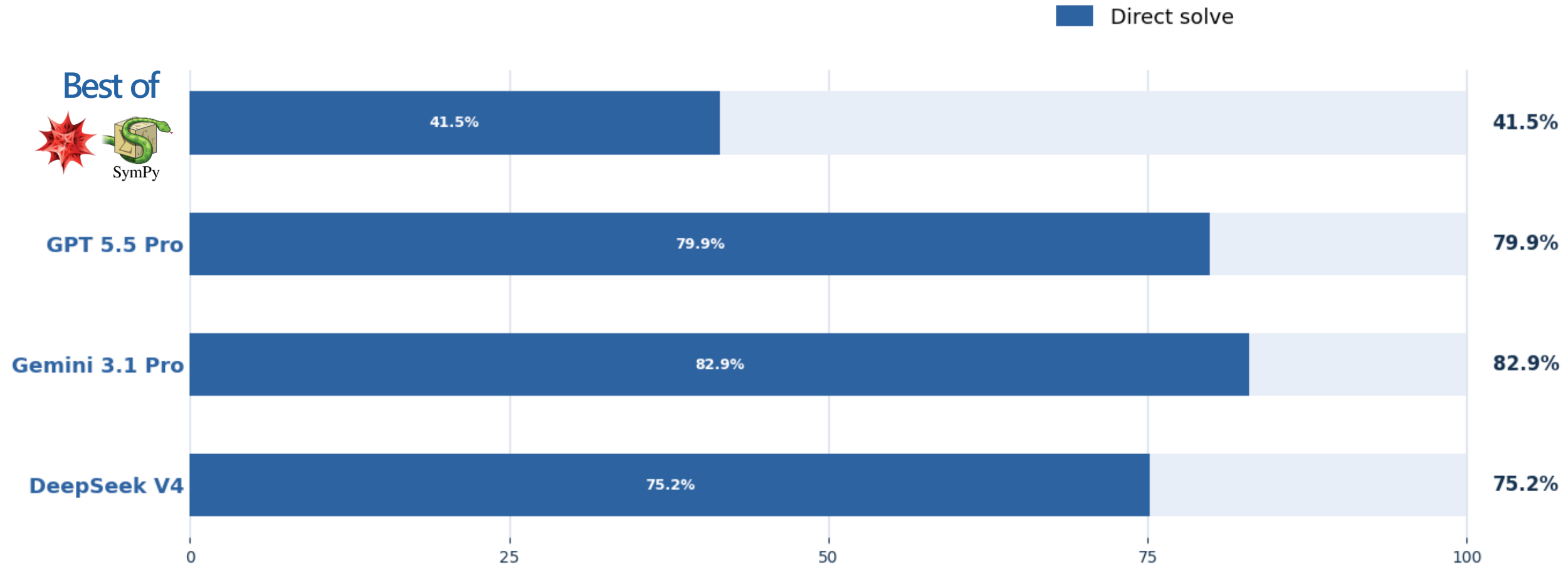
Generate, Verify, Refine

A Closed-Loop LLM–CAS Approach to Symbolic Integration

Stav Belyy · Shalev Zuriel · Tomer Raz · Michael Shalyt · Ido Kaminer | Technion, 2026
Carsten Schneider · Yiman Gao | RISC

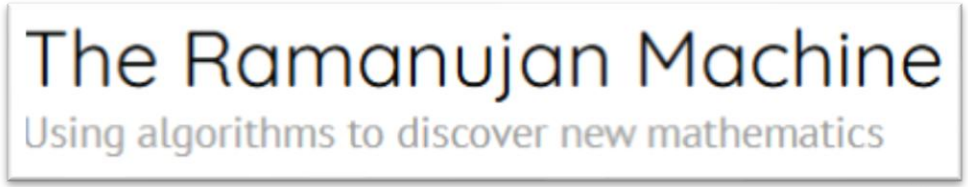
Main result

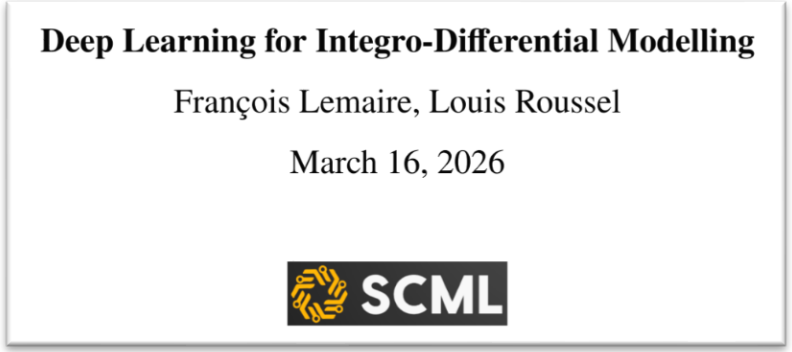
Benchmarking on 1500 integrals, hardest ones among references.



AI-for-math systems

ML systems contribute to mathematical search, proof construction, and algorithm discovery





Tsoukalas, George, et al. "Advancing mathematics research with AI-driven formal proof search." *arXiv preprint arXiv:2605.22763* (2026)

Yang, Kaiyu, et al. "Leandojo: Theorem proving with retrieval-augmented language models." Advances in Neural Information Processing Systems

Raayoni, Gal, et al. "Generating conjectures on fundamental constants with the Ramanujan Machine." Nature 590.7844 (2021): 67-73.

François Lemaire, Louis Roussel: Deep Learning for Integro-Differential Modelling. RISC. Proceedings on Symbolic Computation and Machine Learning, SCML

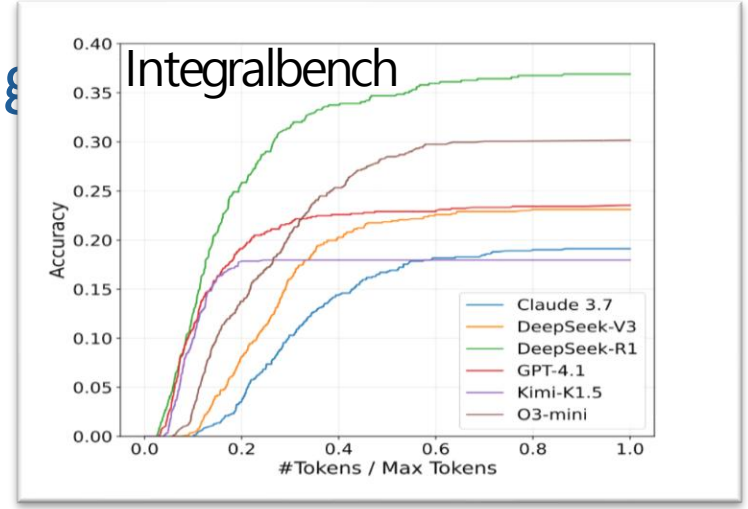
LLM as mathematical assistants

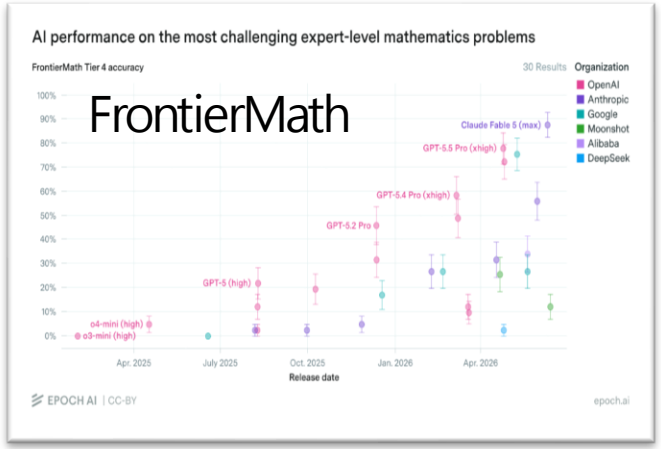
General-purpose LLMs Improving

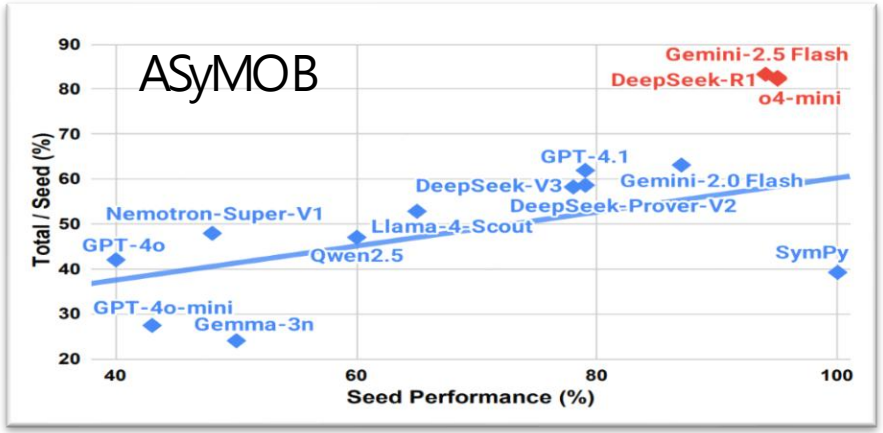
on mathematical tasks,

even without domain-specific

symbolic infrastructure







Tang, Bintao, et al. "Integralbench: Benchmarking llms with definite integral problems." arXiv preprint arXiv:2507.21130 (2025).

Glazer, Elliot, et al. "Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai." arXiv preprint arXiv:2411.04872 (2024).

Shalyt, Michael, Rotem Elimelech, and Ido Kaminer. "Asymob: Algebraic symbolic mathematical operations benchmark." arXiv preprint arXiv:2505.23851 (2025).

Verification is becoming the bottleneck

Harder evaluation settings

- Unpublished problems
- Expert-created tasks
- Research-level problems

Proof verification is hard

- Generated proofs can contain gaps
- Requires formal or expert review

The field is shifting from “can it answer?” to “can we trust it?”

Verification is becoming the bottleneck

Harder evaluation settings

- Unpublished problems
- Expert-created tasks
- Research-level problems

FIRST PROOF

FrontierMath

Erdős problems

The Ramanujan Challenge For A

Proof verification is hard

- Generated proofs can contain gaps
- Requires formal or expert review

The field is shifting from “can it answer?” to “can we trust it?”

Why symbolic integration?

Hard

$$\int f(x)\,dx \to F(x)$$

Requires finding the right symbolic path

Verifiable

$$D[F](x) - f(x) \equiv 0$$

Can be verified by differentiation

Central

$$\int f(x)\,dx$$

Canonical CAS, physics and engineering problems

symbolic integration is **hard to solve,** but **structured to verify.**

The Symbolic Integration Gap

CAS

Wolfram 14.3.0 Kernel

File Edit

```
Wolfram 14.3.0 Kernel for Microsoft Windows (64-bit)
Copyright 1988-2025 Wolfram Research, Inc.

In[1]:= FullSimplify[Integrate[x/((1 + x)*(1 - x^3)^(1/3)), x,
        Assumptions -> -1 < x < 1, GenerateConditions -> False], Assumptions -> -1 < x < 1]

                                x
Out[1]= Integrate[-------------------, x, Assumptions -> -1 < x < 1, GenerateConditions -> False]
                                3 1/3
                  (1 + x) (1 - x )
```

VS

LLM

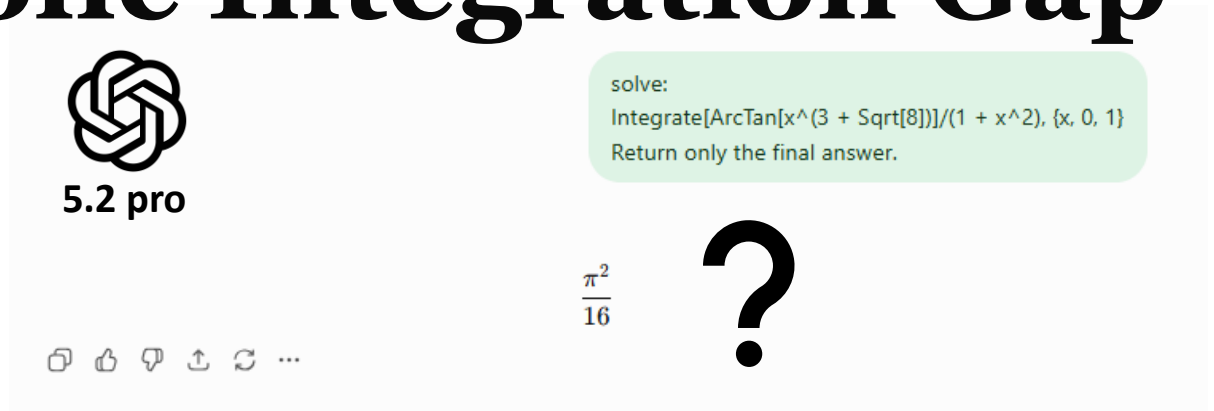
5.5 pro

solve: Integrate[x/((1 + x)*(1 - x^3)^(1/3)),x]
under assumptions -1<x<1
return only final answer, simplified form, with elementary function if possible

$$\frac{\sqrt{3}}{2\sqrt[3]{2}}\arctan\left(\frac{1+2^{1/3}(1-x)}{\sqrt{3}\sqrt[3]{1-x^3}}\right)-\frac{1}{\sqrt{3}}\arctan\left(\frac{1-\frac{2x}{\sqrt[3]{1-x^3}}}{\sqrt{3}}\right)+\frac{1}{2}\ln\left(x+\sqrt[3]{1-x^3}\right)+\frac{\ln\left((1-x)(1+x)^2\right)}{4\sqrt[3]{2}}-\frac{3\ln\left(-1+x+2^{2/3}\sqrt[3]{1-x^3}\right)}{4\sqrt[3]{2}}+C$$

The Symbolic Integration Gap

CAS



VS

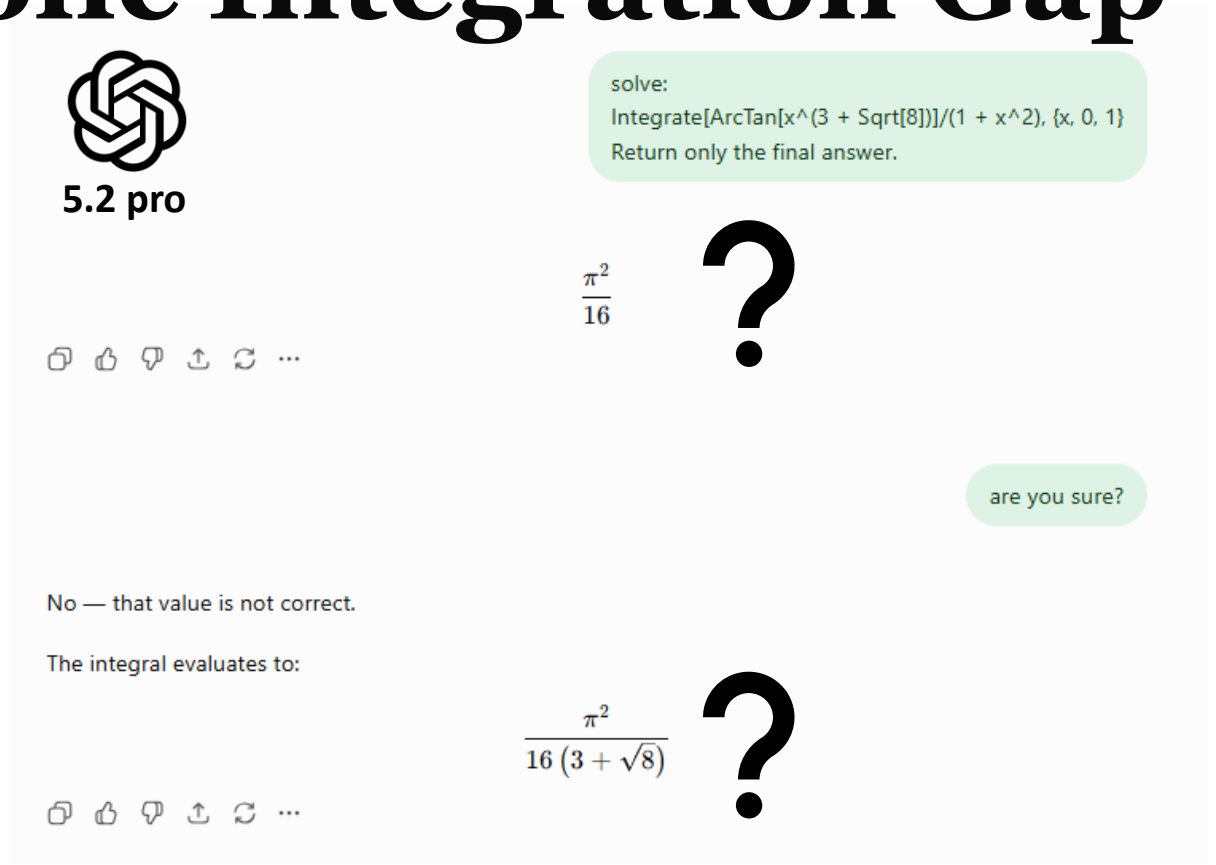
LLM

The Symbolic Integration Gap

CAS

VS

LLM

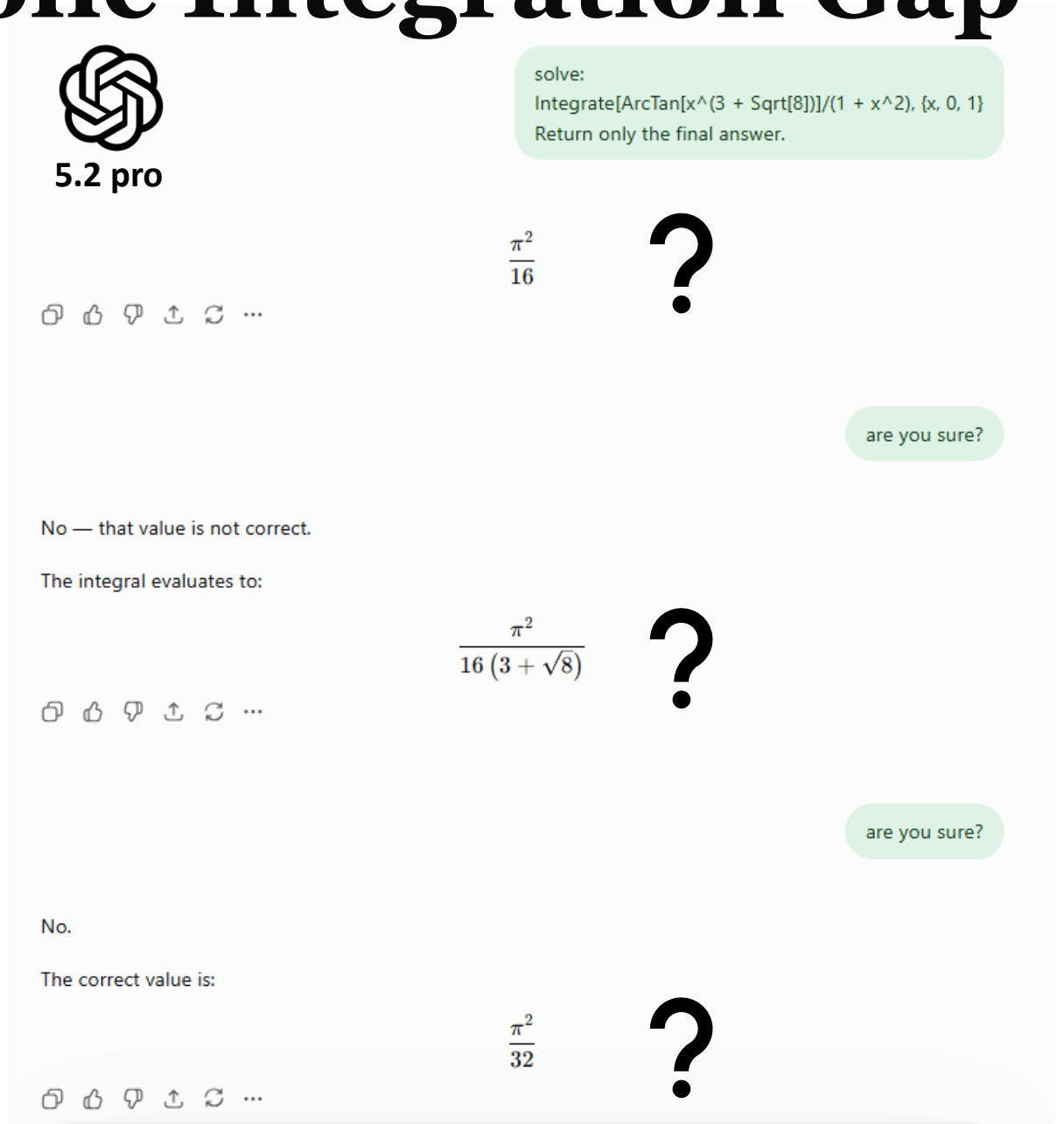


The Symbolic Integration Gap

CAS

VS

LLM



The Symbolic Integration Gap

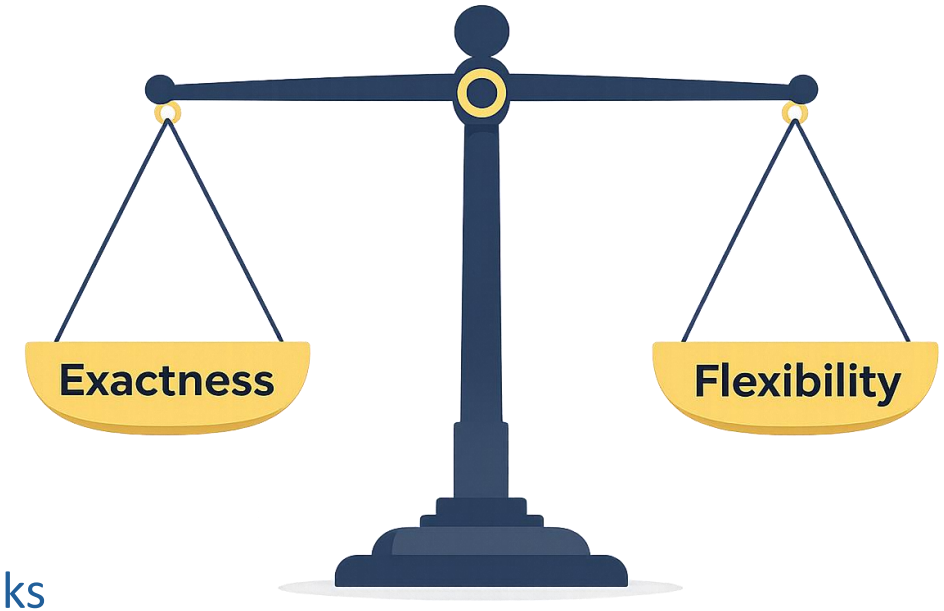
CAS

- Deterministic and exact
- Rule-bound search
- Limited

VS

LLM

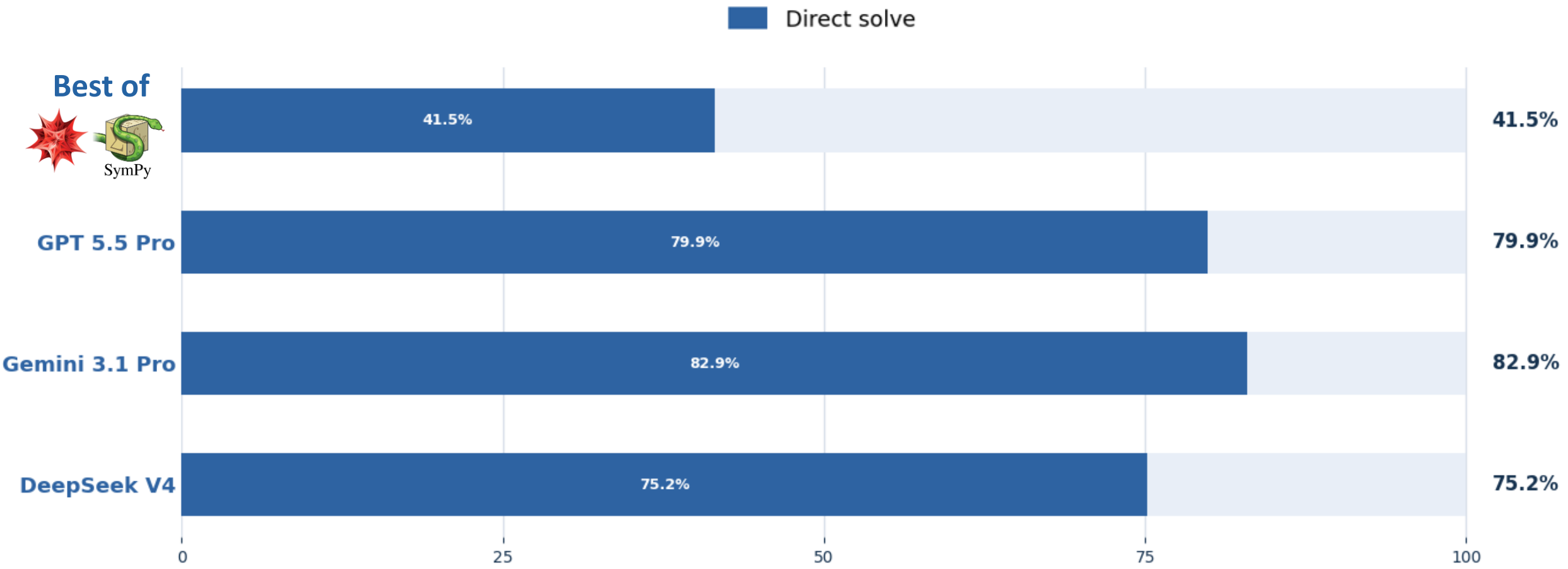
- Creative; Suggests substitutions and tricks
- Produces candidates, not verified results
- Can make mistakes



Opportunity: combine LLM flexibility with CAS exactness.

Main result

Benchmarking on 1500 integrals, hardest ones among references.

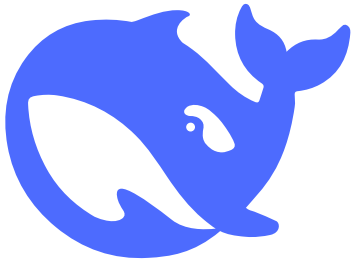


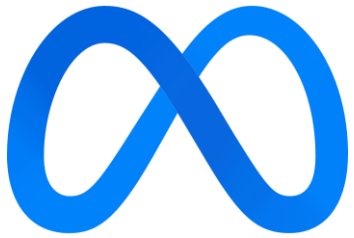
Generate → Verify → Refine

GVR is a closed loop: LLMs propose, CAS verifies, and the system decides what to try next.

Generate





solve:
Integrate[ArcTan[x^(3 + Sqrt[8])]/(1 + x^2), {x, 0, 1}
Return only the final answer.

$$\frac{\pi^2}{16}$$?

are you sure?

No — that value is not correct.

The integral evaluates to:

$$\frac{\pi^2}{16\left(3+\sqrt{8}\right)}$$?

are you sure?

No.

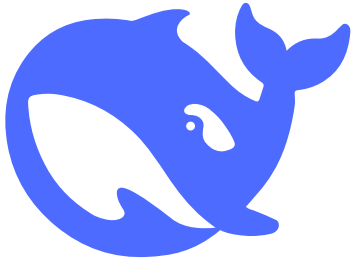
The correct value is:

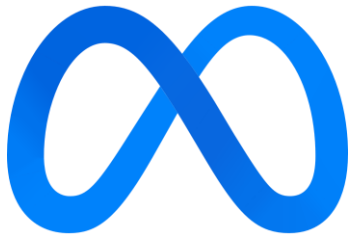
$$\frac{\pi^2}{32}$$?

solve: Integrate[x/((1 + x)*(1 - x^3)^(1/3)),x]
under assumptions -1<x<1
return only final answer, simplified form, with elementary function if possible

$$\frac{\sqrt{3}}{2\sqrt[3]{2}}\arctan\left(\frac{1+2^{1/3}(1-x)}{\sqrt{3}\sqrt[3]{1-x^3}}\right)-\frac{1}{\sqrt{3}}\arctan\left(\frac{1-\frac{2x}{\sqrt[3]{1-x^3}}}{\sqrt{3}}\right)+\frac{1}{2}\ln\left(x+\sqrt[3]{1-x^3}\right)+\frac{\ln\left((1-x)(1+x)^2\right)}{4\sqrt[3]{2}}-\frac{3\ln\left(-1+x+2^{2/3}\sqrt[3]{1-x^3}\right)}{4\sqrt[3]{2}}+C$$

Generate





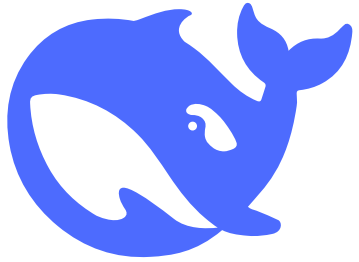
$$\int x^7 \cos(x^4)\, dx$$

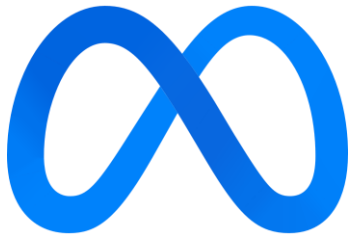
$$\int_1^0 x^2 \log(1+x)$$

$$\int e^x \sin(x\,)\, dx$$

Generate





$$\int x^7 \cos(x^4)\, dx = \frac{1}{4} x^4 \sin(x^4)$$

$$\int_1^0 x^2 \log(1+x) = \frac{1}{3}\log(2) - \frac{5}{9}$$

$$\int e^x \sin(x)\, dx = \frac{1}{2} e^x (\sin(x) - \cos(x))$$

Verification

Independent verification pipeline

model outputs are treated as mathematical claims and accepted only after independent validation.

1 Candidate answer

$$\int x^7 \cos(x^4)\, dx = \frac{1}{4} x^4 \sin(x^4)$$

$$\int_1^0 x^2 \log(1+x) = \frac{1}{3}\log(2) - \frac{5}{9}$$

$$\int e^x \sin(x)\, dx = \frac{1}{2} e^x (\sin(x) - \cos(x))$$

2 Symbolic check

$$D[F](x) - f(x) \stackrel{?}{=} 0$$

3 Numerical validation

$$f(x) \stackrel{?}{=} \tilde{f}(x)$$

$$-0.18432 \neq -0.32450$$

How refinement works

After verification, the system routes failed cases to refine strategy.

Feedback refinement

$$F_0(x) \xrightarrow{verify} D[F_0](x) - f(x) \neq 0 \xrightarrow{feedback} F_1(x)$$

Black-box retry

$$\int x^7 \cos(x^4)\, dx = \frac{1}{4} x^4 \sin(x^4)$$

$$\int_1^0 x^2 \log(1+x) = \frac{1}{3}\log(2) - \frac{5}{9}$$

How refinement works

After verification, the system routes failed cases to refine strategy.

Feedback refinement

$$F_0(x) \xrightarrow{verify} D[F_0](x) - f(x) \neq 0 \xrightarrow{feedback} F_1(x)$$

Black-box retry

Try again!

$$\int x^7 \cos(x^4)\, dx = \frac{1}{4}(x^4 \sin(x^4) + \cos(x^4))$$

$$\int_1^0 x^2 \log(1+x) = \frac{\mathbf{2}}{3}\log(2) - \frac{5}{\mathbf{18}}$$

How refinement works

After verification, the system routes failed cases to refine strategy.

Feedback refinement

$$F_0(x) \xrightarrow{verify} D[F_0](x) - f(x) \neq 0 \xrightarrow{feedback} F_1(x)$$

Black-box retry

Try again!

$$\int x^7 \cos(x^4)\, dx = \frac{1}{4}(x^4 \sin(x^4) + \cos(x^4))$$

$$\int_1^0 x^2 \log(1+x) = \frac{2}{3}\log(2) - \frac{5}{18}$$

Transformation refinement

$$\int f(x)dx \xrightarrow{\text{rewrite / substitution}} \int \tilde{f}(x)dx$$

verify rewrite: $f(x) \equiv \tilde{f}(x)$

Use a substitution

Step-level validation

Solve this!

$$\int x^7 \cos(x^4)\, dx = \frac{1}{4}\int u \cdot \cos(u)du$$

Use integration by parts

$$\int_1^0 x^2 \log(1+x) = \frac{1}{3}\log(2) - \frac{1}{3}\int_1^0 \frac{x^3}{1+x}$$

Feedback refinement: examples

Example A — sign error

$$I(x) = \int f_0(x)\,dx = \int \log\left(\frac{x^2(1 - e^x + x)}{5}\right) dx$$

Candidate: $F_0(x) = -\dfrac{x+9}{x\left(1+log^2\left(\frac{x^2(1-e^x+x)}{5}\right)\right)}$

Cas verifier

FAIL – Derivative check failed

$\dfrac{d}{dx}F_0(x) \neq f_0(x)$

Feedback LLM

New candidate: $F_1(x) = +\dfrac{x+9}{x\left(1+log^2\left(\frac{x^2(1-e^x+x)}{5}\right)\right)}$

PASS

Transformation refinement - example

$$I_1(x) = \int x^3 \log(2+x)^3 \log(3+x)\, dx$$

CAS failed

LLM wrong

Sug g ested rewrite:

$$I_2(x) = \int ((x+2)^3 - 6(x+2)^2 + 12(x+2) - 8)\log(2+x)^3 \log(3+x)\, dx$$

Cas verifier

CAS check
$I_1(x) \equiv I_2(x)$

Solve $I_2(x)$

LLM solved

Evaluation set: hard symbolic integration

Mostly from known benchmarks plus graduate / university-level integral sources.

~1,500 integrals so far

$\int f(x)\,dx$ $\int_a^b f(x)\,dx$

indefinite and definite symbolic integration tasks

Benchmark-derived sources

- IntegralBench
- U-MATH integral-calculus track
- ASyMOB
- CUMath
- MathOdyssey
- UGMathBench

Expert / classical sources

- Vălean, (Almost) Impossible Integrals
- Vălean, More (Almost) Impossible Integrals
- graduate / university integral collections
- MIT Integration Bee

Tang, Bintao, et al. "Integralbench: Benchmarking llms with definite integral problems." *arXiv preprint arXiv:2507.21130* (2025).

Chernyshev, Konstantin, et al. "U-math: A university-level benchmark for evaluating mathematical skills in llms." arXiv preprint arXiv:2412.03205 (2024).

Shalyt, Michael, Rotem Elimelech, and Ido Kaminer. "Asymob: Algebraic symbolic mathematical operations benchmark." arXiv preprint arXiv:2505.23851 (2025).

Tran, Quyen, and Sutthirut Charoenphon. "CUMath: A Benchmark and Evaluation Framework for LLMs on Mathematical Reasoning in Undergraduate Computational Math." (2025).

Fang, Meng, et al. "Mathodyssey: Benchmarking mathematical problem-solving skills in large language models using odyssey math data." Scientific data 12.1 (2025): 1392.

Xu, Xin, et al. "Ugmathbench: A diverse and dynamic benchmark for undergraduate-level mathematical reasoning with large language models." arXiv preprint arXiv:2501.13766 (2025).

Vălean, Cornel Ioan. (Almost) impossible integrals, sums, and series. Springer, 2019.

Vălean, Cornel Ioan. "More (almost) impossible integrals, sums, and series." Cham: Springer (2023).

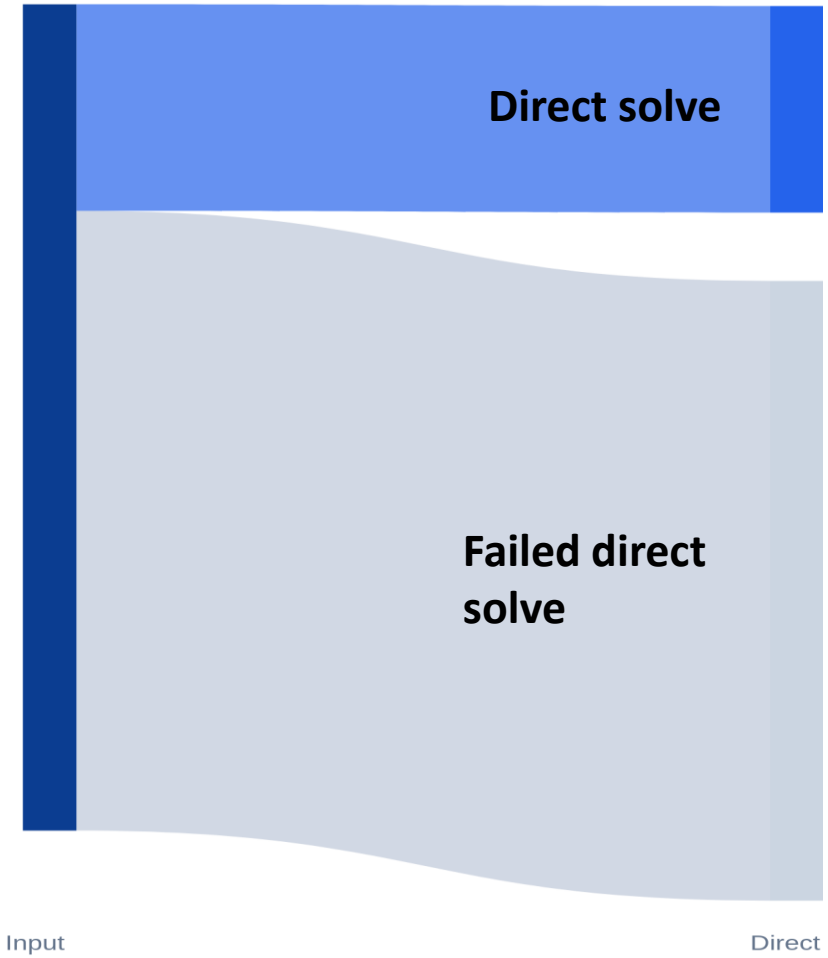
https://math.mit.edu/~yyao1/integrationbee.html

How results are reported

We separate direct success from two refinement gains, then report final verified success.

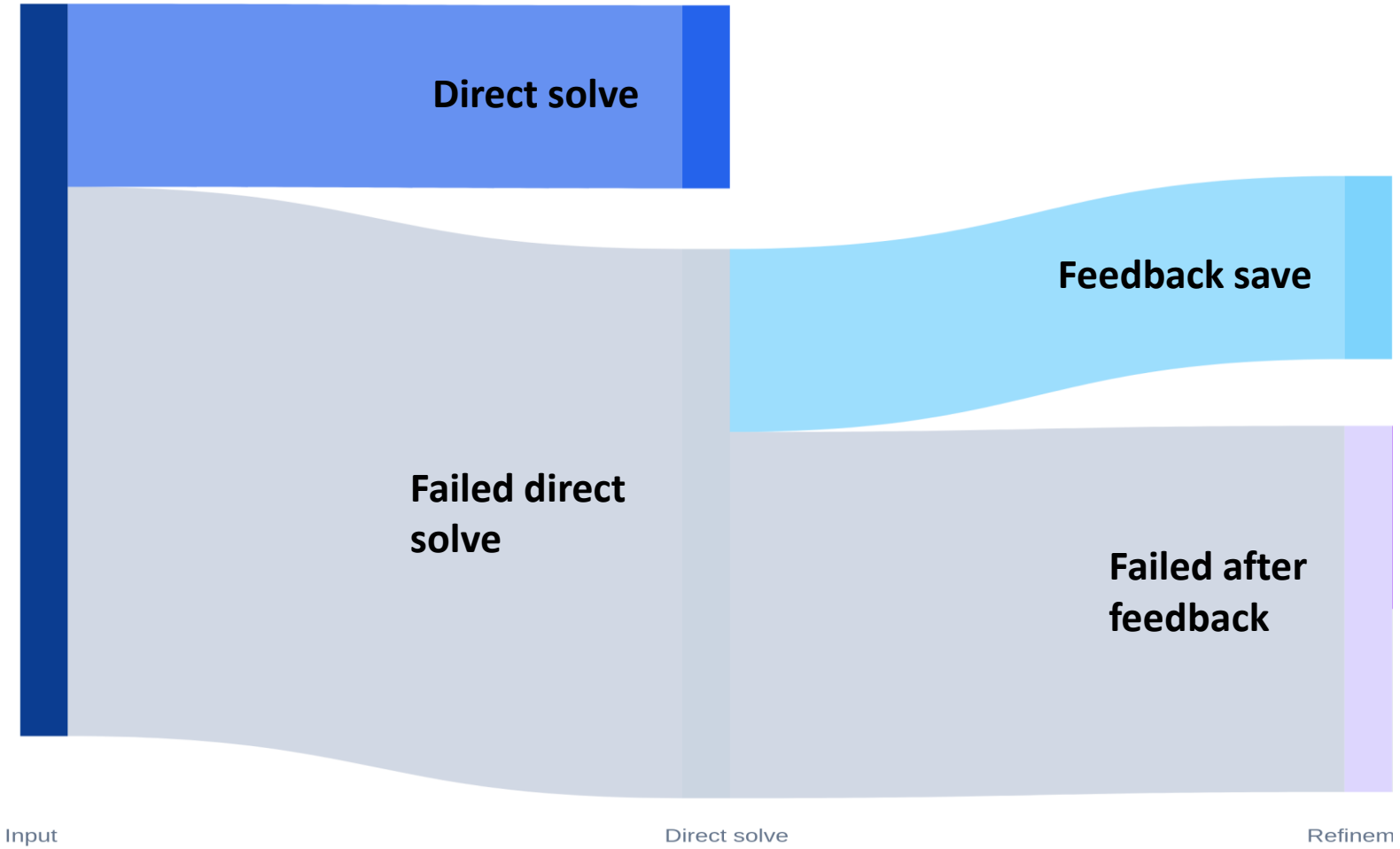
How results are reported

We separate direct success from two refinement gains, then report final verified success.



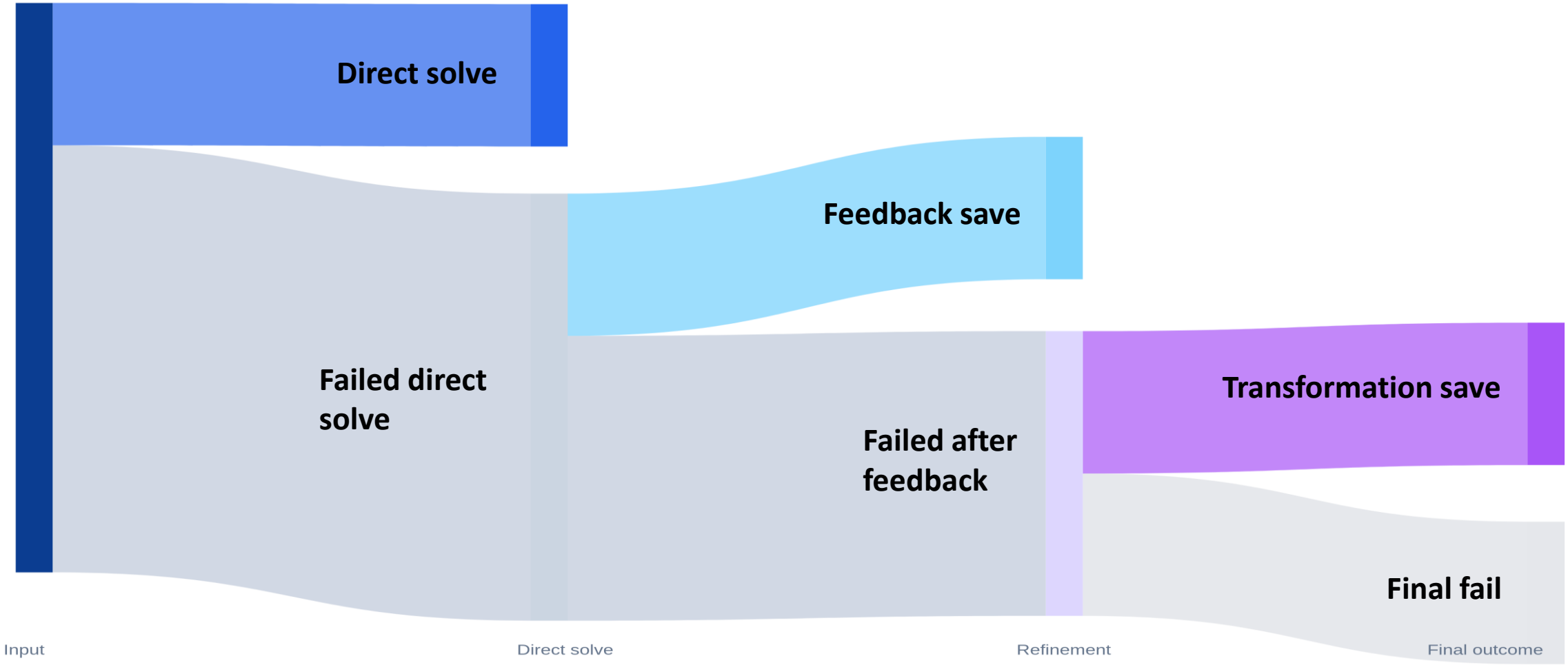
How results are reported

We separate direct success from two refinement gains, then report final verified success.



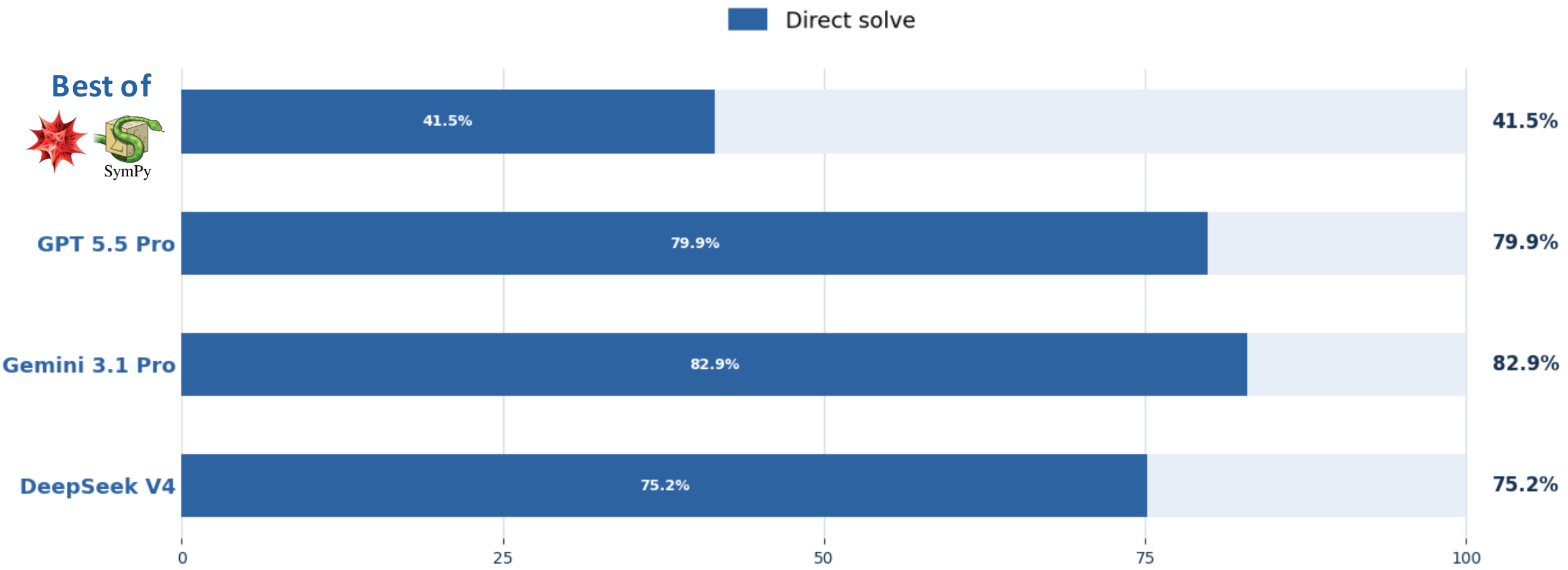
How results are reported

We separate direct success from two refinement gains, then report final verified success.



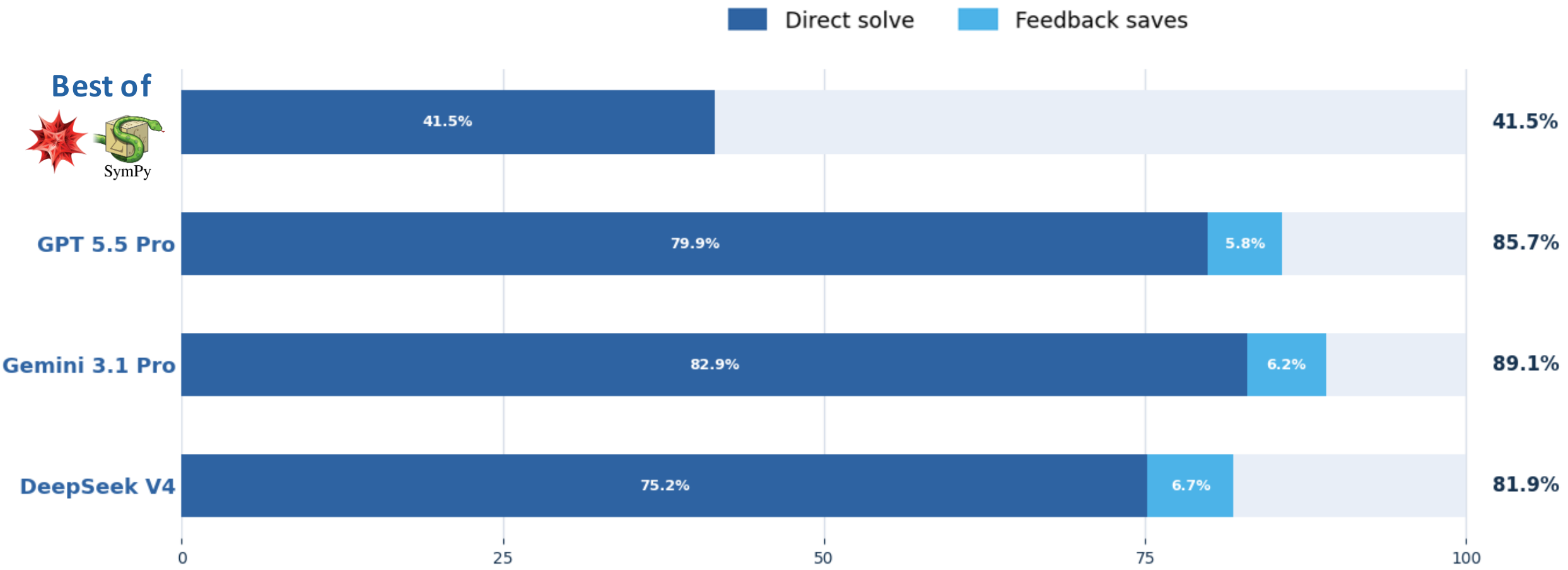
Main result

Frontier models are strong directly; feedback and transformations still add verified solves.



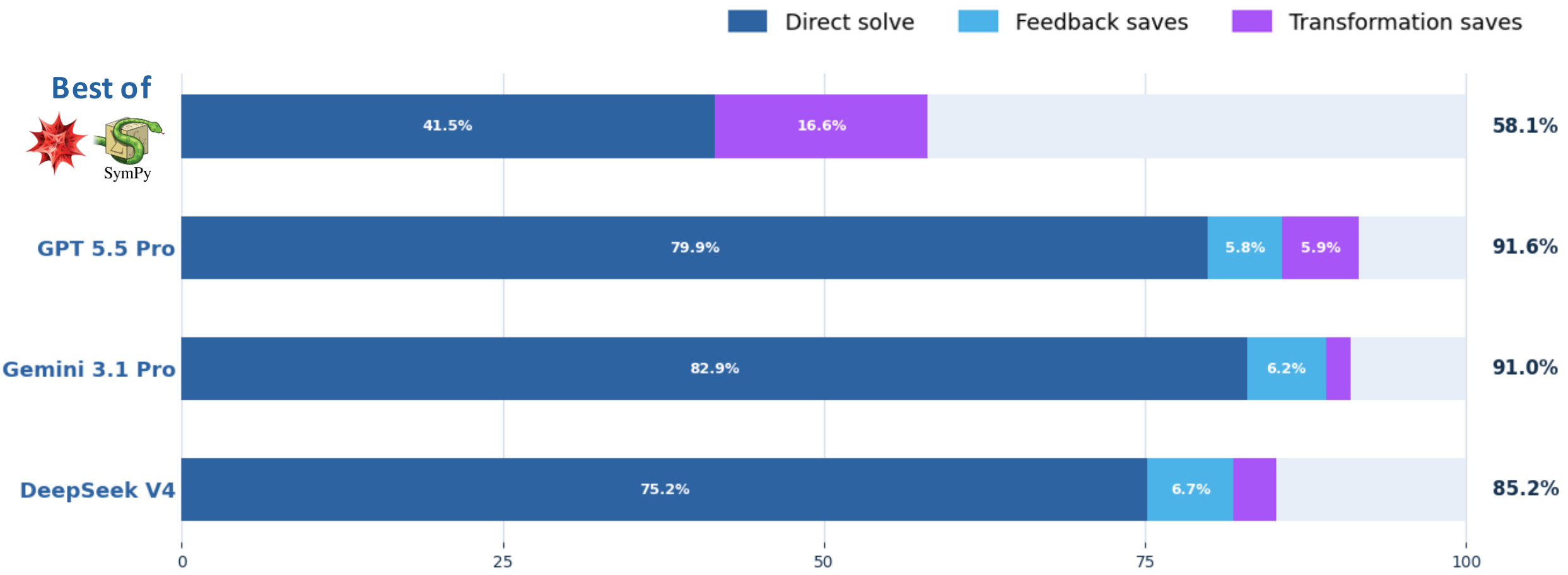
Main result

Frontier models are strong directly; feedback and transformations still add verified solves.



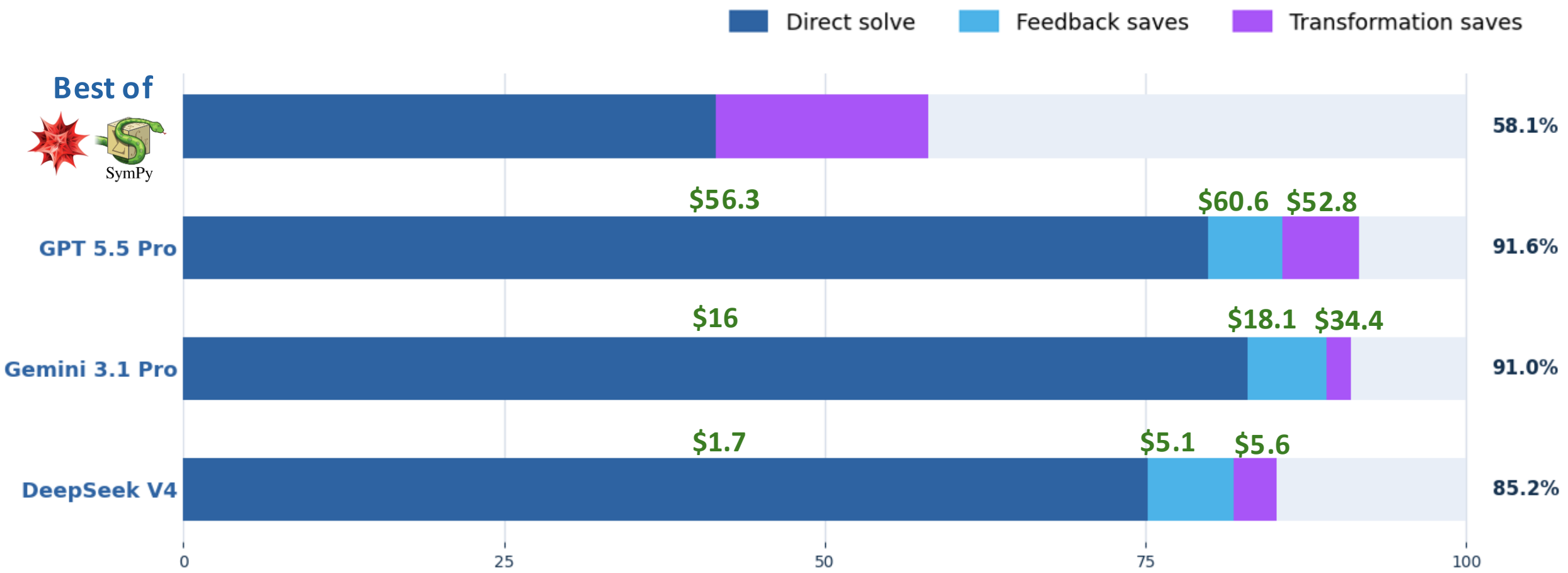
Main result

Frontier models are strong directly; feedback and transformations still add verified solves.



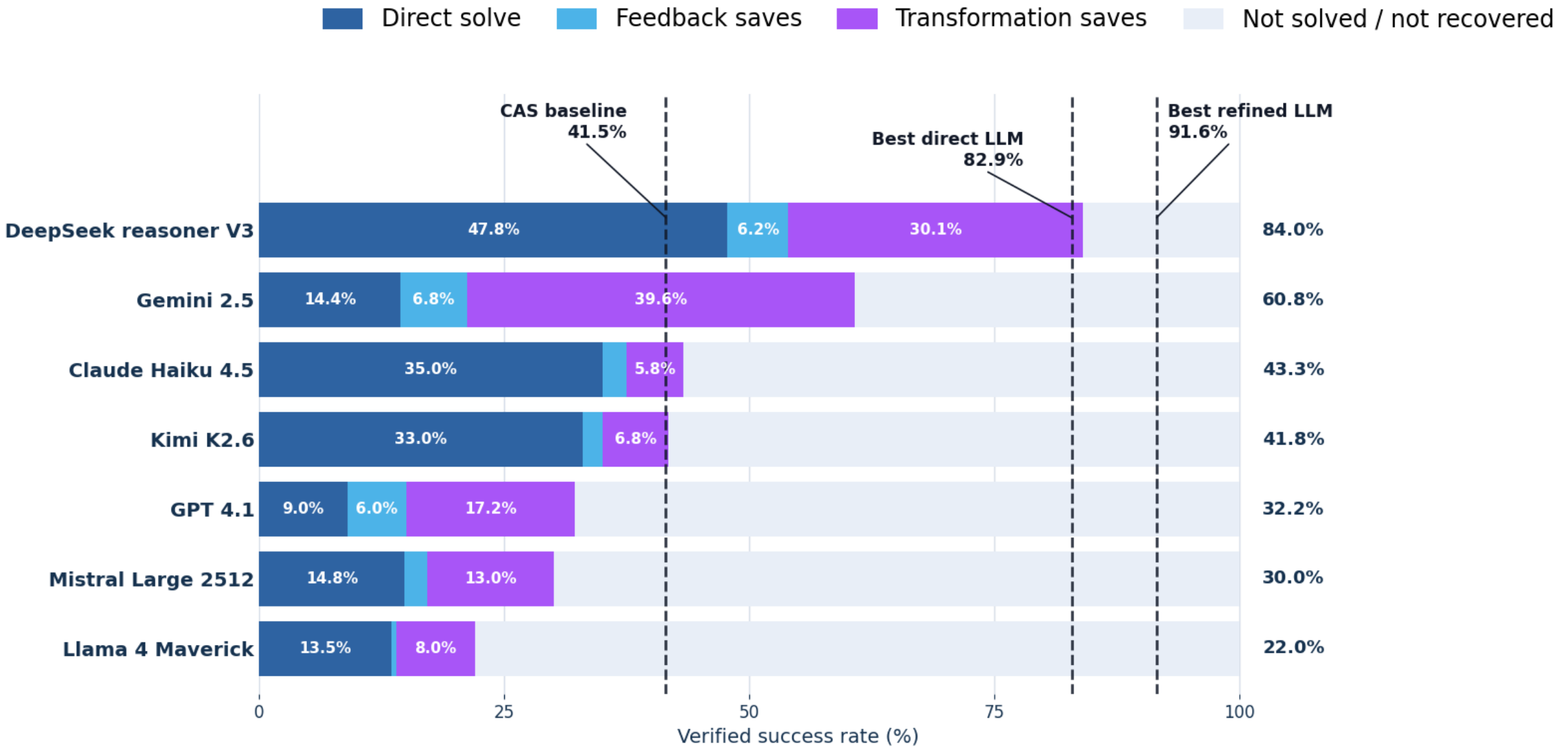
Main result

Frontier models are strong directly; feedback and transformations still add verified solves.



Broader model comparison

Transformation gains are larger for weaker-reasoning models:



Cost estimate per 100 integrals

Usage-derived estimates show large stage and model differences.

Frontier Models

Model	Direct	Direct + feedback	Transform + solve	Full
GPT 5.5 Pro	$56.3	$60.6	$52.81	**$169.8**
Gemini 3.1 Pro	$15.9	$18.12	$34.41	**$52.53**
DeepSeek V4	$1.74	$5.12	$5.67	**$10.79**
DeepSeek V3	$1.06	$2.26	$3.42	**$5.68**

Assumptions per 100 integrals:

Direct = one direct solve + feedback
Direct + feedback= one direct solve + feedback
Transform + solve= one transform + direct solve
Full = direct solve + feedback + transform + solve

Small model Model

Model	Direct	Direct + feedback	Transform + solve	Full
Kimi K2.6	$0.14	$0.28	$0.31	**$0.51**
Claude Haiku 4.5	$0.5	$0.98	$0.99	**$1.97**
Mistral Large	$0.08	$0.17	$0.16	**$0.33**
Llama Maverick	$0.02	$0.04	$0.05	**$0.09**

Vision

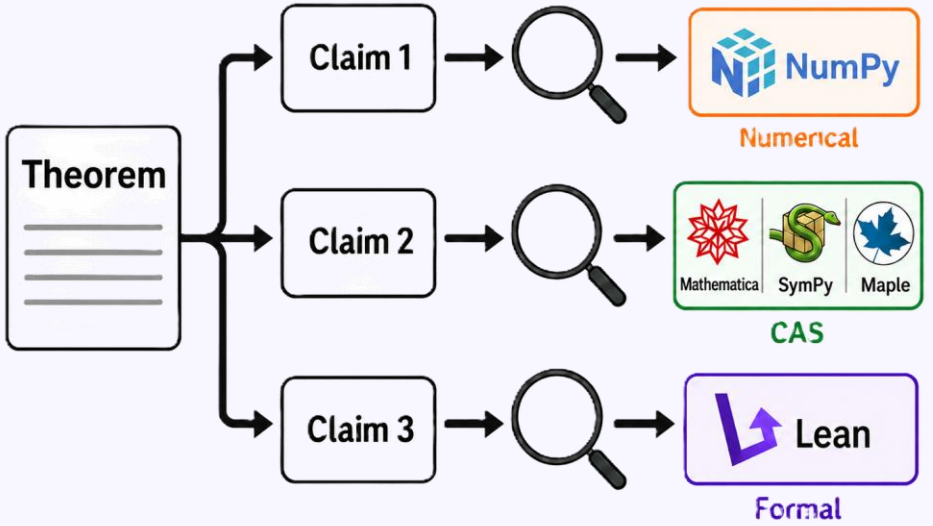
What is the future of CAS?

Hybrid LLM–CAS workflows may evolve in three directions:

- **Intrinsic mastery:** models improve direct symbolic reasoning.
- **Deeper integration:** CAS will co-evolve with LLMs for stronger symbolic support.
- **Autonomous tool creation:** models may begin building internal symbolic procedures.

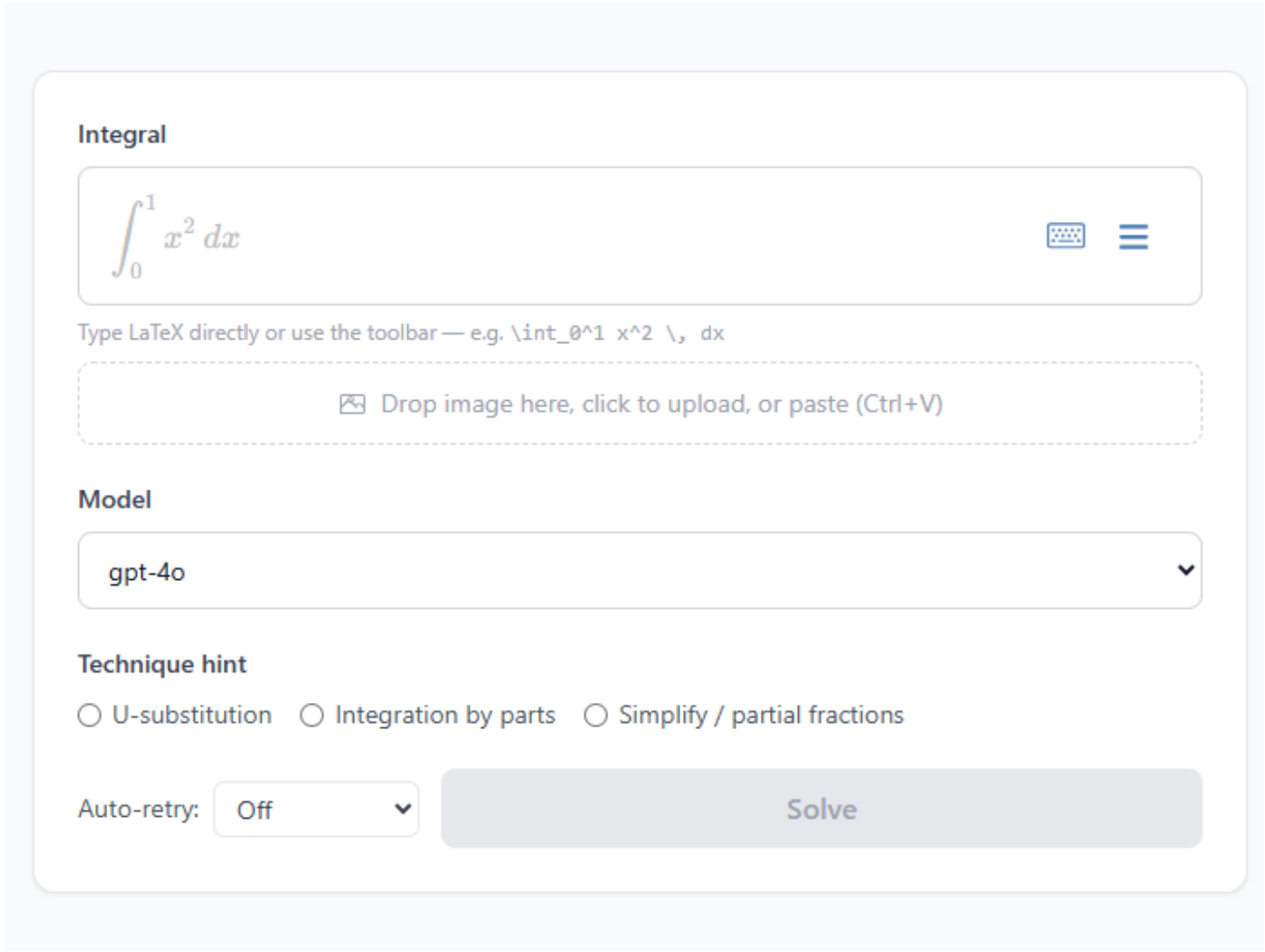
Trust through verification workflows

The next step is not only stronger models, but **trusted workflows** that combine generation with verification.



Shalyt, Michael, Rotem Elimelech, and Ido Kaminer. "Asymob: Algebraic symbolic mathematical operations benchmark." arXiv preprint arXiv:2505.23851 (2025).

Check out our demo!



Check out more from our team at SCDDE

New Directions in Irrationality

Hila Barkan, Elyasheev Leibtag, Rotem Kalisch, Shachar Weinbaum, Michael Shalyt, Ido Kaminer

Efficient Symbolic Evaluation of D-Finite Functions

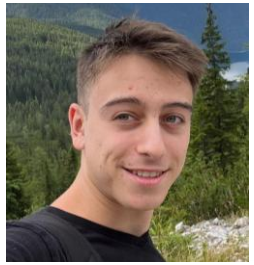
Michael Shalyt Rotem Kalisch Elyasheev Leibtag Ido Kaminer

An Infinite Family of Apéry-like Recurrences for Euler's Constant

Rotem Kalisch Tali Monderer Elyasheev Leibtag Shachar Weinbaum Michael Shalyt Ido Kaminer

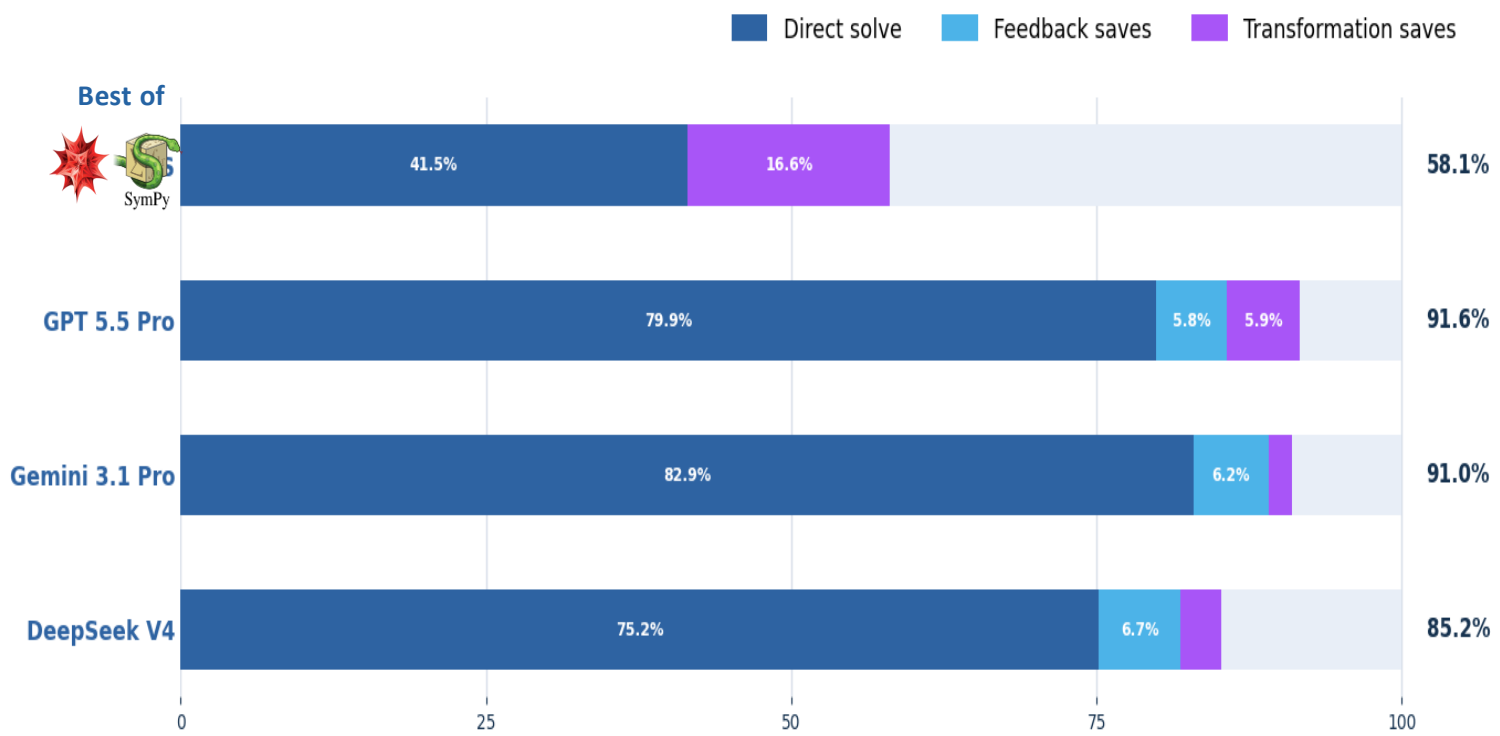
Questions?

demo

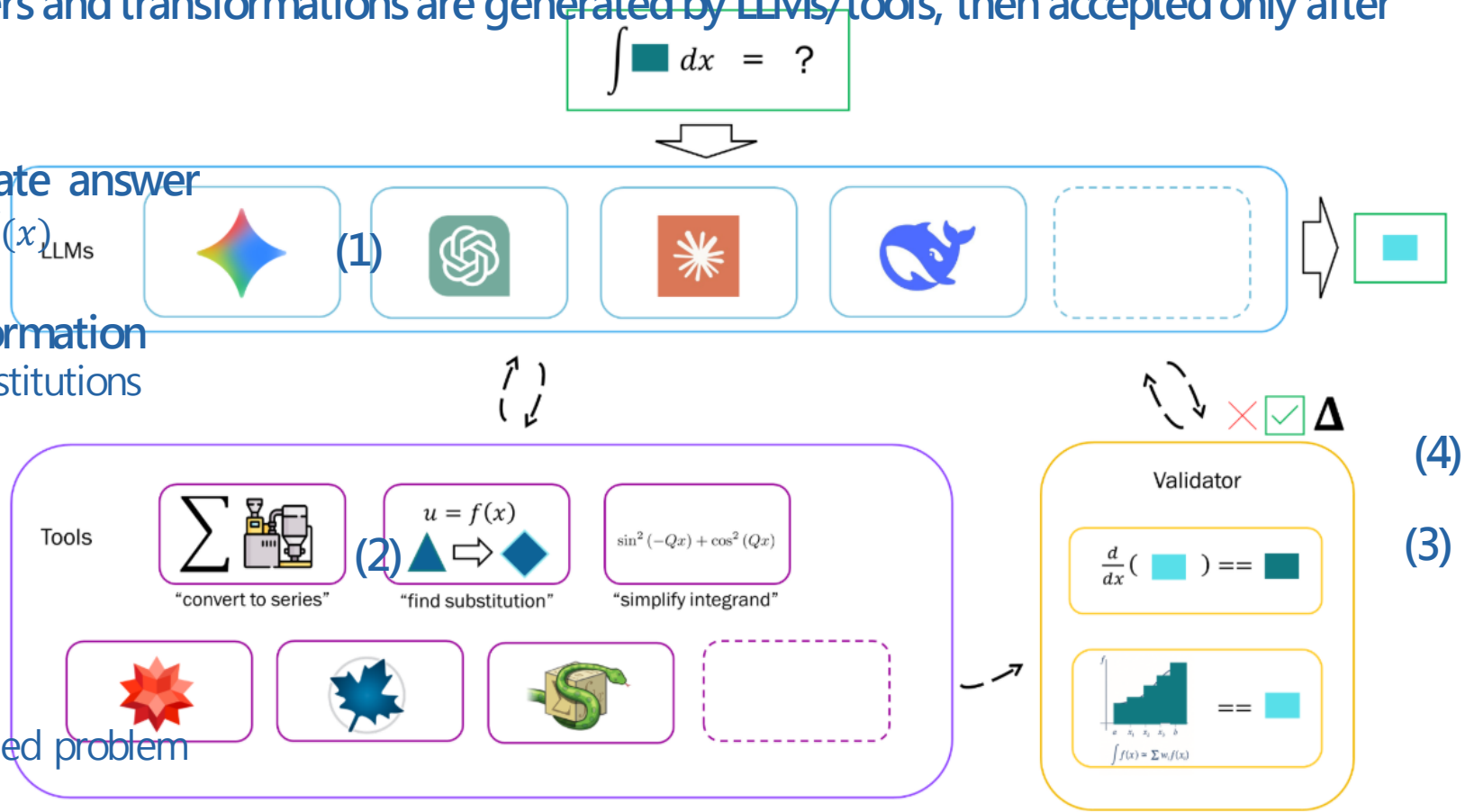

THE RAMANUJAN CHALLENGE FOR AI

bitly



System architecture: generate, verify, refine

Candidate answers and transformations are generated by LLMs/tools, then accepted only after validation.



(1) Generate candidate answer
LLM proposes $F(x)$ / $\tilde{f}(x)$

(2) Generate transformation
propose rewrites / substitutions

(3) Verify
derivative check + numerical check

(4) Refine
feedback or transformed problem returns to solver