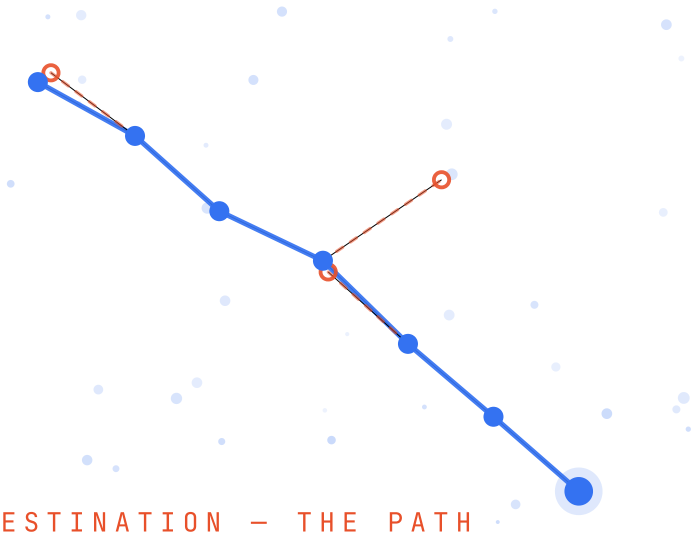
SoTALab

BY ZHIHU



NOT JUST THE DESTINATION — THE PATH

High-Value Math Reasoning Data

Math Reasoning Dataset

Math problems aren't answer problems. For complex, open-ended reasoning, we build process-level data — trainable, evaluable, error-traceable.

PROCESS SUPERVISION

EXPERT-CALIBRATED · MULTI-MODEL REVIEW

POSITION · PROCESS OVER ANSWER

Right answer, wrong reasoning.

A math answer is objective and auto-gradable, so answer-hit rate easily becomes a proxy for a model's reasoning ability. But when training rewards only "did the final result come out right," the model is pushed toward **outcome-driven learning** — to score points, it takes shortcuts.

It may not truly master the derivation; it gets better at **finding shortcuts among problem types, numbers, formulas and answer distributions**. Over time, what the model learns may not be mathematical reasoning itself, **but the statistical correlation between problems and answers**.

So a correct answer doesn't necessarily come from correct reasoning: it may misuse a theorem, drop a condition, skip a key transformation, or land by pattern-matching. As a reward signal the final result is too sparse — it can't tell "stable reasoning" from "local formula-plugging" or "a lucky hit."

> **What the model learns is often not mathematical reasoning itself, but the statistical correlation between problems and answers.**

SHORTCUT · STATISTICAL CORRELATION

PUBLIC SIGNAL · PUBLIC BENCHMARKS

Benchmarks already signal it: even the hardest evals only score the destination.

Hard math benchmarks keep arriving, pushing problems toward research level. But most still grade only the **final answer** — they tell you a model got it wrong, not **where its reasoning broke**.

FrontierMath EPOCH AI	Pushes math problems toward the level of **modern mathematical research**. [1]	research
Geometry3K · GeoQA GEOMETRY	Geometry needs abstract understanding, symbolic reasoning and axiomatic knowledge — reading text, figure and theorem at once. [2]	3 modal
Multimodal Geometry ZERO-SHOT	GPT-4 scores just 60 / 36 / 51% on GPSM4K / PGPS9K / Geometry3K; Gemini-Pro 44 / 30 / 47%; some models still single-digit. [3]	≤60%
MathVista MULTIMODAL	GPT-4V's overall accuracy is still **below human**, often erring on complex figures and rigorous reasoning. [4]	49.9%

A destination score marks only **right or wrong** — not **which step broke**. That is exactly the signal process-level data adds.

[1] Epoch AI · FrontierMath [2] GeoQA · arXiv:2105.04165 [3] Geometry MLLM · arXiv:2412.00846 [4] MathVista · arXiv:2310.02255

EVIDENCE · PROCESS OVER OUTCOME

Process matters more than the answer.

Not just SoTALab's view — the field keeps converging on process supervision.

2022

Chain-of-Thought WEI ET AL. [1]

> **Proof of concept** · no large-scale data yet, but it set the direction: intermediate reasoning steps have value in themselves.

Writing out reasoning steps measurably improves arithmetic, commonsense and symbolic reasoning — the field starts treating "process" as a supervisable signal.

2023

Process Supervision OPENAI · PRM800K [2]

800K STEP-LEVEL HUMAN ANNOTATIONS

OpenAI trains a process reward model on 800K step-by-step feedbacks — process supervision is no longer a concept; it has large-scale annotation behind it.

78% MATH SUBSET SOLVE RATE

The process-supervised PRM solves ~78% of a representative MATH subset — feedback on every step beats rewarding only the final answer.

A reasoning chain verified step by step is more reliable, and more trusted, than one judged only at the destination.

2024

ProcessBench ERROR LOCALIZATION [3]

3,400 ERROR-LOCALIZATION CASES

Competition / olympiad-level problems built to test "which step it first goes wrong" — evaluation moves from right/wrong to process error localization.

When a model is wrong, you no longer just know it's wrong — you can locate the **exact step** where it begins to drift.

> **From CoT to error localization, the field agrees: supervise the process, not just the answer. SoTALab pushes that to research-level mathematics — mathematicians calibrating every step, line by line.**

[1] Wei et al. · Chain-of-Thought · arXiv:2201.11903 [2] Lightman et al. · Let's Verify Step by Step · arXiv:2305.20050 [3] ProcessBench · arXiv:2412.06559

MINI-EXPERIMENT · FIVE PROBLEMS, FIVE FAILURES

Five problems, five kinds of failure.

SoTALab's mathematicians picked five representative problems spanning **symbolic integration, differential geometry, algebraic geometry, holomorphic symplectic geometry and numerical PDE** — each typically takes a domain mathematician hours or even days to solve. We had several frontier models answer the same set under a unified process checkpoint, averaged over five runs, then human-reviewed — not to prove a model "can't," but to answer: **where exactly does it go wrong?**

PROBLEM · DIRECTION	TYPICAL FAILURE MODE	GPT-5.5 PRO	GEMINI 3.1	CLAUDE OPUS 4.8
Quartic-radical integral ★ SYMBOLIC INTEGRAL	Can't see the hidden structure; never finds the real entry point.	1.6	0.3	6.2
Geodesic on an elliptic cylinder DIFFERENTIAL GEOMETRY	Treats 'straight after unrolling' as globally shortest.	7.8	6.2	7.5
Parametric RUR specialization ALGEBRAIC GEOMETRY	Compresses the key proof into 'specialization works.'	6.6	2.6	6.7
Fixed-point-free symplectic action HOLOMORPHIC SYMPLECTIC	Misses the bridge that closes the theorems into a full proof.	7.6	7.0	7.8
D2D / D2E regularity ★ NUMERICAL PDE	Skips 2nd-derivative integrability — lower error isn't better.	7.1	7.5	7.6

5-problem avg · 5-run /10 GPT-5.5 Pro **6.1** Gemini 3.1 **4.7** Claude Opus 4.8 **7.2**

The five point to three missing abilities: ① **find the entry point** · ② **keep the reasoning chain valid** · ③ **locate & fix process errors**. This brief details the two starred ★ problems.

WORKED SAMPLE · STRUCTURE RECOGNITION

Sample ① | Symbolic computation | Quartic-radical integral

The model can compute — it just can't see the hidden structure

> What's hard isn't the length of the integrand, but that the **entry point is deeply hidden** — the usual substitutions stall fast.

PROBLEM

$$I = \int \frac{dx}{(x-1)\sqrt[4]{x^3+x}}$$

✦ BREAKTHROUGH

> **Hidden identity** — usual substitutions stall; the real entry is spotting $(x+1)^4 - 8(x^3+x) = (x-1)^4$.

> **Normalizing variable** — set $u = \frac{\sqrt[4]{8x^3+8x}}{x+1}$, collapsing the quartic radical into a rational integral.

> **Genuinely elementary** — reduces to $I = -2^{-1/4}\int \frac{u^2}{1-u^4}\,du$, which has an elementary antiderivative.

⚠ HOW MODELS FAIL

- **GPT-5.5 Pro** parses the radical correctly but wrongly judges "no elementary antiderivative," punting to special functions / numerics.
- **Gemini 3.1** misreads the radical's degree; the whole direction drifts off.
- **Claude Opus 4.8** finds the identity, but final-step coefficients still need human review.

✓ PROCESS CHECKPOINTS · 5-RUN AVG

PROCESS CHECKPOINT	PTS	GPT-5.5 PRO	GEMINI 3.1	CLAUDE OPUS 4.8
Parse the radical correctly	1	1.0	0.0	1.0
See ordinary substitution fails	1	0.6	0.3	1.0
Choose structure-preserving u	2	0.0	0.0	1.2
Use the hidden quartic identity	2	0.0	0.0	2.0
Reduce to a rational integral in u	2	0.0	0.0	0.7
Back-substitute to elementary form	2	0.0	0.0	0.3
Total	**10**	**1.6**	**0.3**	**6.2**

one vague low score → **atomic checkpoints pin the loss to the exact step: "wrongly judged no elementary antiderivative"**

WORKED SAMPLE · NUMERICAL PDE & OPERATOR LEARNING

Sample ⑤ | Numerical PDE | D2D / D2E regularity

Lower error doesn't necessarily mean better mathematics

> Two neural-operator strategies look like simple computation on the surface; the real math question is regularity **near the boundary**.

SETUP

$$D_\alpha(s,t) = (s, t^\alpha),\ \alpha > 1; \quad \hat{u}_{\text{ref}} = t \Rightarrow \hat{u}_{\text{D2D}} = y^{1/\alpha},\ \hat{u}_{\text{D2E}} = y$$

✦ REGULARITY ANALYSIS

Unbounded 1st derivative — $\partial_y \hat{u}_{\text{D2D}} = \frac{1}{\alpha} y^{1/\alpha - 1}$ is unbounded as $y \to 0$.

H² test — $\int_0^1 y^{2/\alpha - 4} dy$ converges $\Leftrightarrow \alpha < \frac{2}{3}$; but $\alpha > 1$, so **D2D ∈ H²**, while D2E is smooth.

Low-error trap — L^2 error D2D 3.46 / D2E 3.51 / GeoFNO 6.52; D2D is lower by only 0.05pp yet not in H^2.

⚠ HOW MODELS FAIL

- **GPT-5.5 Pro** gets the inverse map and errors right, but skips the second-derivative integrability check.
- **Gemini 3.1** finishes most of the computation; the ranking explanation still needs human review.
- **Claude Opus 4.8** reaches the most complete conclusion; a few steps still need review.

✓ PROCESS CHECKPOINTS · 5-RUN AVG

PROCESS CHECKPOINT	PTS	GPT-5.5 PRO	GEMINI 3.1	CLAUDE OPUS 4.8
Compute the inverse map and u_D2D	2	1.8	1.8	1.8
Analyze the unbounded 1st derivative	2	1.6	1.6	1.7
Check H^2 via the 2nd derivative	2.5	1.2	1.8	1.9
Compute both error reductions	2	1.8	1.7	1.8
Explain scalar-error vs regularity	1.5	0.7	0.6	0.4
Total	**10**	**7.1**	**7.5**	**7.6**

lower scalar error → **function-space regularity**

MATHEMATICIAN NETWORK · WHO BUILDS THE DATA

Pinpointing failure this finely takes a mathematician.

SoTALab fields roughly 50 PhD-level core experts — with research, competition-setting and teaching backgrounds — spanning 7 major areas: algebra · geometry · analysis · numerical computation · probability & statistics · combinatorics · number theory. Every point lost on the previous pages was marked, line by line, by a mathematician.

Research-grade problems	Problems are designed by mathematicians who can read at research level, so “hard” comes from structure, not length.
Line-by-line review	Mathematicians verify the reasoning step by step, marking the first wrong step and its error type — not just the final answer.
Competition background	Core experts have competition-setting and judging experience, and can tell whether a proof “actually holds.”
Cross-domain coverage	Every one of the seven directions has a matched expert — from algebraic geometry to numerical PDE, all can be calibrated.

> **Marking a model’s error to the exact step takes precisely this — a team that can solve these problems itself.**

DATA STANDARD · HOW WE BUILD DATA

Difficulty is the floor, not the goal.

Hard problems are easy to write; the value is in making every data record **trainable, verifiable, scoreable and diagnosable** — every claim independently checkable, every wrong step precisely located. A SoTALab math data record must satisfy all five:

Difficult	Designed by PhD-trained core experts, at the level of a graduate qualifying exam and research reading.
Trainable	Carries the full working method — judgment, materials, tool use — so the model learns the **process**, not just the answer.
Verifiable	Comes with a rigorous reference solution and key intermediate results that can be independently confirmed.
Scoreable	Broken into atomic, true/false **checkpoint** lines covering key numbers, judgments, evidence and form.
Diagnosable	A low score maps to the **specific missing step**, not a vague “wrong.”

SCHEMA · DATA STRUCTURE

What a process-level math data record looks like.

It records not just **whether** a model reaches the final result, but **how** it gets there — making explicit the process judgments a human mathematician relies on when solving and reviewing.

problem	The problem itself, with a clear line of inquiry and research-level adversarial difficulty. core
golden_response	An expert-written rigorous derivation (the Golden Response), as the top-quality anchor. core
answer	The final result, used only as an end-of-chain check, not the sole signal.
hint · comment	Breakthrough hints and expert notes, marking the entry point and pitfalls.
classification	Domain / direction / difficulty tags, indexable by ability dimension.
model_error_output	Real erroneous outputs from multiple models, as contrastive negatives.
error_location · error_type	First wrong step + error type (misread / theorem / extrapolation / missed check / back-substitution). core
step_checkpoint	Step-weighted process checkpoints, decomposing every technical bridge into a machine-checkable test. core
human_review	Human review record, keeping the process annotation trustworthy and replayable.

SYNTHESIS · THE THESIS

From an answer bank to a process dataset.

Built around the real bottlenecks of math reasoning, SoTALab makes more than a "question bank" — training samples that are **trainable, evaluable, error-traceable and replayable**:

Algebra · Analysis

Train abstract structure and proof rigor; keep the reasoning chain valid.

Geometry

Train spatial modeling, figure understanding and local–global transfer.

Computation · Combinatorics · Number Theory

Train structure recognition, valid manipulation and result self-checking.

Numerical PDE · Scientific Computing

Train the math behind error metrics, not just the scalar magnitude.

MATH RESEARCHERS	Guarantee research-level rigor in problems and solutions; design difficulty that is genuinely "hard in structure."
COMPETITION SETTERS	Guarantee proofs are judgeable and steps decomposable; write technical bridges as executable checkpoints.
MATH EDUCATORS	Guarantee decomposition granularity and teaching transferability; annotate entry points, pitfalls and checks.
SENIOR AI ENGINEERS	Guarantee the rubric runs stably — multi-model, multi-run, unified scoring, human-review loop.

What we care about isn't a model's total score but **why it loses points** — misreading the problem, misusing a theorem, over-extrapolating a local result, missing a denominator condition, or a final back-substitution error. Every failure maps back to the Schema's `error_location · error_type · step_checkpoint`.

Ordinary QA data

Focuses on knowledge recall. Tells the model what the answer is — right or wrong at the destination, but nothing about where the reasoning broke.

Process-level math data

Focuses on structured reasoning. Tells the model **why it's done this way, where it's easy to err, and how to check a step holds.**

The essential difference is that ordinary QA optimizes recall, while process-level math data optimizes a supervisable, diagnosable, replayable reasoning process — exactly the capability SoTALab's ever-expanding process-level math dataset is built to deliver.

Process over Final Answer

In the future, AI won't replace mathematicians — it must first learn to be examined like one.

SoTALab

BY ZHIHU

Tell us your training / evaluation direction, and we'll quickly assemble a matched process-level math data & evaluation pipeline.

Request the full sample · book a walkthrough →

contact@sotalab.ai · WeChat / Feishu