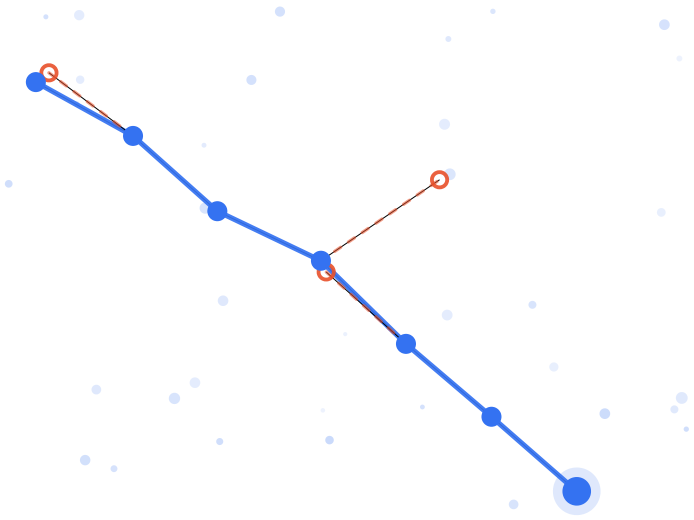
SoTALab

BY ZHIHU



不只看终点，更看路径

高价值数学路径训练数据

Math Reasoning Dataset

数学题不是答案题——
为复杂、开放的数学推理，
构造可训练、可评测、可追错的过程级数据。

立场 · PROCESS OVER ANSWER

答案对了，推理却错了。

数学题答案客观、可自动评估，于是答案命中率很容易被当作模型推理能力的代理指标。但当训练只奖励「最后有没有给对结果」，模型就被推向**结果驱动的学习**——为了拿分，它会去抄近道。

它未必真正掌握推导，而更擅长在**题型、数字、公式和答案分布之间找捷径**。久而久之，模型学到的可能并不是数学推理本身，**而是从题目到答案之间的统计关联**。

于是正确答案并不一定来自正确推理：它可能误用定理、漏掉条件、跳过关键变形，甚至靠模式匹配碰巧得到。最终结果作为奖励信号过于稀疏，分不清「稳定推理」「局部套公式」还是「碰巧答对」。

> **模型学到的，往往不是数学推理本身，而是题目到答案之间的统计关联。**

SHORTCUT · STATISTICAL CORRELATION

公开信号 · PUBLIC BENCHMARKS

公开 benchmark 已给出信号：再难的评测，也只看终点。

近年高难数学 benchmark 迅速出现，把题目推到接近**数学研究**的层级。但它们大多仍以**最终答案是否命中**计分——分数能告诉你模型「错了」，却告诉不了你它**错在推理的哪一步**。

FrontierMath EPOCH AI	把数学问题推进到更接近**现代数学研究**的层级。[1]	研究级
Geometry3K · GeoQA GEOMETRY	几何解题需抽象理解、符号推理与公理化知识，并同时读懂文字、图形与定理。[2]	3 模态
多模态几何 ZERO-SHOT	GPT-4 在 GPSM4K / PGPS9K / Geometry3K 仅 60 / 36 / 51%；Gemini-Pro 44 / 30 / 47%；部分模型仍处个位数。[3]	≤60%
MathVista MULTIMODAL	GPT-4V 整体准确率仍**低于人类**，常在复杂图形理解与严格推理上出错。[4]	49.9%

一个终点分只标出**对或错**，标不出**错在第几步**——这正是过程级数据要补上的信号。

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

研究依据 · PROCESS OVER OUTCOME

过程，比答案更重要。

这不是 SOTALAB 的单方面判断 · 学界正持续向过程监督收敛

2022 **Chain-of-Thought** WEI ET AL. [1]

概念验证 · 尚无规模数据，但确立方向：中间推理步骤本身就有价值

写出推理步骤，即可显著提升算术、常识与符号推理——学界开始把「过程」当作可监督的信号。

2023 **Process Supervision** OPENAI · PRM800K [2]

800K 步级人工标注

OpenAI 用80万条逐步反馈训练过程奖励模型——过程监督不是概念，已有大规模标注支撑。

78% MATH 子集解出率

过程监督 PRM 在 MATH 代表性子集上约解出 78%——对每一步给反馈，效果优于只奖励最终答案。

经人类逐步核验的推理链，比只看终点更可靠、更被认可。

2024 **ProcessBench** ERROR LOCALIZATION [3]

3,400 错误定位测例

竞赛 / 奥赛级题目，专门测「最早错在第几步」——评测从「对/错」推进到过程错误定位。

模型答错时，不再只知道「错了」，而能定位到**具体哪一步**开始偏离。

> 从 CoT 到错误定位，学界已达成共识：过程比答案更值得监督。SoTALab 要做的，是把这条路线推到研究级数学——让数学家逐行定标每一步推理。

[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

小实验 · FIVE PROBLEMS, FIVE FAILURES

五道题，暴露五类失败。

SoTALab的数学家们选取了能覆盖**符号积分、微分几何、代数几何、全纯辛几何、数值 PDE** 的五道代表题——每道通常需要对口方向的数学家**数小时乃至数天**才能解出。让多个前沿模型在同一套题上作答，统一过程 checkpoint、五次运行取均分，再由人工复核——不为证明某个模型「不行」，而为回答：**模型到底错在哪里?**

题目 · 方向	典型失败模式	GPT-5.5 PRO	GEMINI 3.1	CLAUDE OPUS 4.8
四次根式积分 ★ SYMBOLIC INTEGRAL	看不见隐藏结构，找不到真正的突破口。	1.6	0.3	6.2
椭圆柱面测地线 DIFFERENTIAL GEOMETRY	把「展开后是直线」等同于全局最短。	7.8	6.2	7.5
参数 RUR 特化 ALGEBRAIC GEOMETRY	把关键证明压缩成「specialization works」。	6.6	2.6	6.7
无固定点辛作用 HOLOMORPHIC SYMPLECTIC	漏掉把定理连成闭合证明链的桥梁。	7.6	7.0	7.8
D2D / D2E 正则性 ★ NUMERICAL PDE	跳过二阶导可积性——低误差不一定更好。	7.1	7.5	7.6

五题均分 · 5-run /10　GPT-5.5 Pro **6.1**　Gemini 3.1 **4.7**　Claude Opus 4.8 **7.2**

五道题共同指向三种缺失能力：① **发现解题入口** · ② **维护推理链合法性** · ③ **定位并修正过程错误**。本册详解 ★ 标注的两道。

评测样本 · 结构识别

样本 ① 符号计算 四次根式积分

模型不是不会算，而是看不见隐藏结构

> 真正困难的不是积分式很长，而是**解法入口非常隐蔽**——常规代换会很快卡住。

PROBLEM · 题目

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

✦ 突破口 BREAKTHROUGH

> **隐藏恒等式**常规代换很快卡住，真正入口是看出 $(x+1)^4 - 8(x^3+x) = (x-1)^4$。

> **归一化变量**引入 $u = \frac{\sqrt[4]{8x^3+8x}}{x+1}$，四次根式被压成有理积分。

> **确属初等**归约为 $I = -2^{-1/4}\int \frac{u^2}{1-u^4}\,du$，存在初等原函数。

⚠ 模型如何失手

- **GPT-5.5 Pro** 解析对根式，却误判「无初等原函数」，推给特殊函数 / 数值。
- **Gemini 3.1** 看错根式次数，方向整体偏离。
- **Claude Opus 4.8** 找到恒等式，但末步系数仍需人工复核。

✓ 过程 CHECKPOINT · CHECKPOINTS · 5-RUN 均分

过程 CHECKPOINT	分值	GPT-5.5 PRO	GEMINI 3.1	CLAUDE OPUS 4.8
正确解析根式	1	1.0	0.0	1.0
识别普通代换不足	1	0.6	0.3	1.0
选择保结构代换 u	2	0.0	0.0	1.2
使用隐藏四次恒等式	2	0.0	0.0	2.0
化为 u 的有理积分	2	0.0	0.0	0.7
求得并回代初等原函数	2	0.0	0.0	0.3
Total	10	1.6	0.3	6.2

一个含糊的低分 → **原子化 checkpoint 把失分精确定位到「误判无初等原函数」这一步**

评测样本 · 数值 PDE 与算子学习

样本 ⑤ | 数值 PDE | D2D / D2E 正则性

低误差，不一定代表数学上更好

两类神经算子策略表面上只是简单计算，真正的数学问题在**边界附近**的正则性。

SETUP · 设定

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

✦ 正则性分析 ANALYSIS

一阶导无界 $\partial_y \hat{u}_{\text{D2D}} = \frac{1}{\alpha} y^{1/\alpha - 1}$，当 $y \to 0$ 无界。

H^2 判定 $\int_0^1 y^{2/\alpha - 4} dy$ 收敛 $\Leftrightarrow \alpha < \frac{2}{3}$；但 $\alpha > 1$，故 **D2D ∈ H^2**，D2E 光滑。

低误差陷阱 L^2 误差 D2D 3.46 / D2E 3.51 / GeoFNO 6.52——D2D 仅低 0.05pp，却不属于 H^2。

⚠ 模型如何失手

- **GPT-5.5 Pro** 算对逆映射与误差，却跳过二阶导可积性检查。
- **Gemini 3.1** 计算多数完成，排序解释仍需人工复核。
- **Claude Opus 4.8** 结论最完整，少量步骤仍需复核。

✓ 过程 CHECKPOINT · CHECKPOINTS · 5-RUN 均分

过程 CHECKPOINT	分值	GPT-5.5 PRO	GEMINI 3.1	CLAUDE OPUS 4.8
计算逆映射与 u_D2D	2	1.8	1.8	1.8
分析一阶导无界	2	1.6	1.6	1.7
二阶导检验 H^2 可积性	2.5	1.2	1.8	1.9
计算两种相对误差下降	2	1.8	1.7	1.8
解释标量误差 vs 正则性排序	1.5	0.7	0.6	0.4
Total	**10**	**7.1**	**7.5**	**7.6**

更低的标量误差 → **函数空间正则性**

数学家网络 · WHO BUILDS THE DATA

能拆到这个颗粒度，因为定标的是数学家。

SoTALab 内约 **50** 位 **PhD 级核心专家**，具备研究、竞赛命题与教研背景，覆盖 **7** **大领域**——代数 · 几何 · 分析 · 数值计算 · 概率统计 · 组合 · 数论。前面每一处失分，都由数学家逐行审稿标定。

研究级命题	由具备研究阅读能力的数学家设计题目，确保「难」来自结构、而非堆长。
逐行审稿	数学家逐步核验推理，标定最早出错的步骤与错误类型，而非只看最终答案。
竞赛 / 奥赛背景	核心专家具竞赛命题与评判经验，能判断一条证明「是否真的成立」。
跨领域覆盖	七大数学方向均有对口专家，从代数几何到数值 PDE 都能定标。

数据标准 · HOW WE BUILD DATA

难，只是我们的底线，不是目标。

难题很容易写，价值在于让每条数据**可训练、可验证、可评分、可诊断**——每个论断都能被独立检查，错一步就能精确定位是哪一步失败。一条 SoTALab 数学数据须同时满足五性：

难 DIFFICULT	由具备博士级训练的核心专家，按研究生资格考试与研究阅读层级设计。
可训练 TRAINABLE	携带完整工作方法——判断、材料与工具使用，让模型学到**过程**而非只学答案。
可验证 VERIFIABLE	配有严谨的参考解与关键中间结果，可被独立确认。
可评分 SCOREABLE	拆成原子化、可判真假的 **checkpoint** 行，覆盖关键数字、判断、证据与格式。
可诊断 DIAGNOSTIC	低分映射到**具体缺失的步骤**，而非一个含糊的「答错」。

字段说明 · SCHEMA

一条过程级数学数据，长什么样。

不只记录模型**是否**给出最终结果，更刻画它**如何**到达结果——SoTALab的每道题把人类数学家在解题与审稿中依赖的过程判断显性化。

problem	题目本身，具备明确探究方向与研究级对抗性。 核心
golden_response	专家撰写的严谨推导过程（Golden Response），作为最高质量锚点。 核心
answer	最终结果，仅作推理链的末端校验，不作唯一信号。
hint · comment	突破口提示与专家点评，标注解题入口与易错处。
classification	领域 / 方向 / 难度标签，支持按能力维度索引。
model_error_output	多模型的真实错误输出，作为对照负样本。
error_location · error_type	最早出错的步骤位置 + 错误类型（题意 / 定理 / 外推 / 漏检 / 回代）。 核心
step_checkpoint	分步加权的过程 checkpoint，把每个技术桥梁拆成可机检判据。 核心
human_review	人工复核记录，保证过程标注可信、可复盘。

总收束 · THE THESIS

从答案数据集走向过程数据集。

围绕数学推理中的真实瓶颈，SoTALab 构造的不是「题库」，而是**可训练、可评测、可追错、可复盘**的数学推理训练样本:

代数 · 分析
训练抽象结构与证明严谨性，维护推理链合法性。

几何
训练空间建模、图形理解与局部—全局转换。

计算 · 组合 · 数论
训练结构识别、合法变形与结果自检。

数值 PDE · 科学计算
训练理解误差指标背后的数学意义，而非只看标量高低。

数学研究者	保证题目与解法的研究级专业性，设计真正「难在结构」的对抗性。
竞赛命题专家	保证证明可判、步骤可拆，把技术桥梁写成可执行的 checkpoint。
数学教研专家	保证拆解颗粒度与教学可迁移性，标注入口、易错处与检查方法。
AI高级工程师	保证 Rubric 可稳定执行——多模型、多次运行、统一评分、人工复核闭环。

我们关心的不只是模型总分，而是**它为什么失分**——题意解析错误、定理调用错误、局部结论外推、分母条件漏检，还是最后回代错误。每一种失分，都映射回 Schema 里的 error_location · error_type · step_checkpoint。

普通问答数据

更关注知识召回。告诉模型**答案是什么**——终点对或错，但说不清推理在哪一步断裂。

过程型数学数据

更关注结构化推理。告诉模型**为什么这样做、哪里容易错、怎样检查这一步是否成立**。

> 高质量数学数据与普通问答数据的本质区别，在于后者优化召回，而前者优化**可监督、可诊断、可复盘的推理过程**——这正是 SoTALab 持续扩展的过程级数学数据集所要交付的能力。

Process over Final Answer

未来，AI 不是替代数学家，而是需要先学会像数学家一样被检验。

SoTALab

BY ZHIHU

告知我们您的训练 / 评测方向，我们将快速组建对口的过程级数学数据与评测管线。

索取完整样本 · 预约 walkthrough →

contact@sotalab.ai · 微信 / 飞书联系对接人