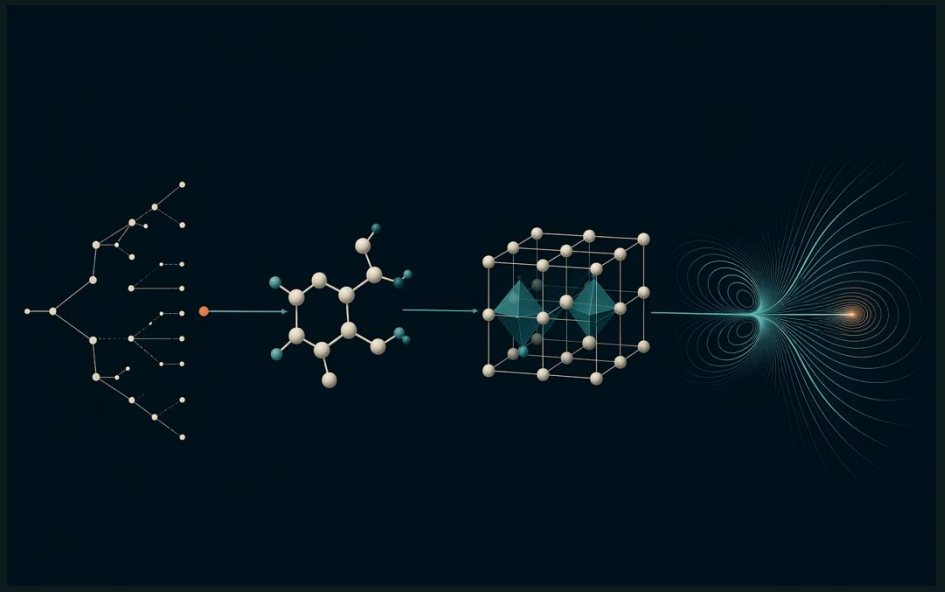
AI FOR SCIENCE · 2023—2026

见证未来

AI 加速中的人类科技演化



徐兴成
上海人工智能实验室

从预测下一个 token，到提出、验证并理解新的科学候选。

集智俱乐部·未来学读书会第九期
2026年11月15日

陶哲轩：从 2023 年预言到研究级成果

关键前提仍是工具集成与正确使用

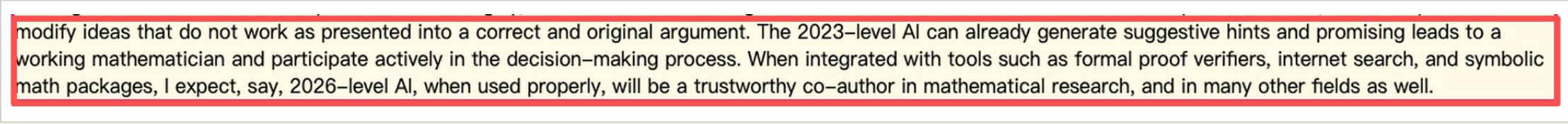
2023 · 预言

当 AI 与形式证明验证器、搜索和符号数学工具集成并正确使用时，2026 年 AI 将成为**值得信赖的合著者**。

2026 · 复盘

AI 已经能够突破**研究级重大成果**。

Lean、搜索、符号计算与实验环境，让生成结果进入可验证的研究流程。


modify ideas that do not work as presented into a correct and original argument. The 2023–level AI can already generate suggestive hints and promising leads to a working mathematician and participate actively in the decision–making process. When integrated with tools such as formal proof verifiers, internet search, and symbolic math packages, I expect, say, 2026–level AI, when used properly, will be a trustworthy co–author in mathematical research, and in many other fields as well.

工具集成是“可信赖合著者”的前提。

科学加速依赖模型、工具、验证与反馈循环的协同

四种能力共同缩短“提出—验证—反馈—再提出”的周期

01 → 02 → 03 → 04

模型

压缩知识、推理与生成候选

Harness

组织上下文、工具、状态与行动

验证环境

编译器、物理约束、实验与临床终点

反馈循环

成功、失败与不确定性进入下一次尝试

候选更多 → 筛选更快 → 实验更多 → 发现周期更短

演讲结构：从模型能力到科学发现

四个部分，逐步回答科学发现如何被加速

01	大模型如何训练	参数、对齐、推理、环境
02	Agent Harness 是什么	Context、接口、Loop、RSI
03	AI如何加速数学研究	从个人实践到批量成果
04	AI如何加速科学发现	物理、生命、材料

主线：模型能力如何变成可靠的科学发现?

PART 01

01

大模型如何训练

从“预测下一个 token”，到“在环境中完成工作”。

DATA

MODEL

TOOLS

DISCOVERY

训练全景：后续阶段不断加入新的学习信号

预训练始终是底座；变化发生在它之上

01	02	03	04	05
预训练	**指令微调**	**偏好对齐**	**推理 RL**	**智能体 RL**
下一 token 数据 + 参数 + 算力	人类示范 任务格式	奖励模型 RLHF / DPO	可验证奖励 测试时计算	工具 + 环境 长时程轨迹

2023：对话 → 2024—2025：推理 → 2025—2026：智能体

扩展对象从模型内部的 token，走向模型外部的行动轨迹。

预训练：以下一 token 预测学习通用表示

得到的是通用续写底座，而不是天然可靠的助手

给定上文，预测下一个 token

$$\max_{\theta} \sum_{t} \log p_\theta(x_t \mid x_{<t})$$

- 为降低误差，模型压缩语言、知识、代码与部分推理规律。
- 后续训练再把这个底座变成助手、推理者与智能体。

数据

覆盖 · 质量 · 去重 · 配比

参数

表示容量 · 稀疏激活

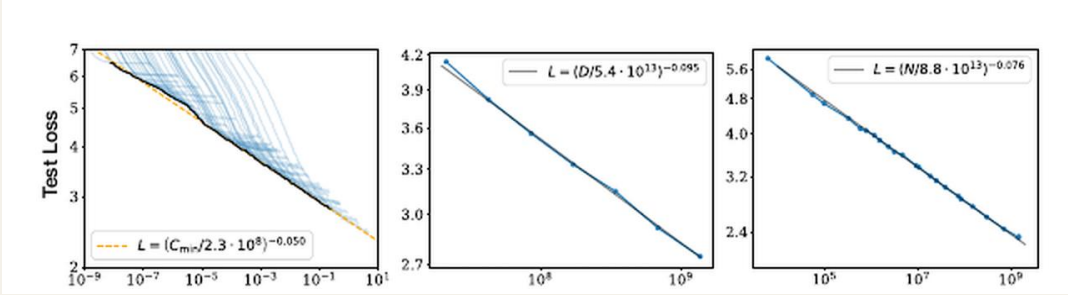
算力

步数 · 精度 · 系统效率

规模定律与涌现：能力如何随规模变化

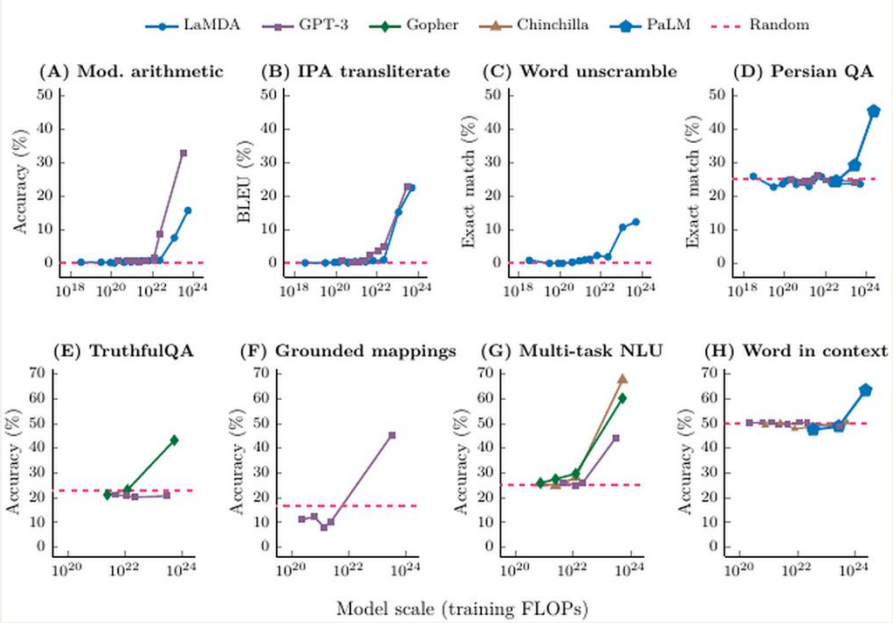
损失可以外推，部分能力却只在规模超过阈值后出现

SCALING LAWS · 可外推的损失



参数、数据与算力需要协同增长；损失曲线可以指导训练配置。

EMERGENCE · 阈值后的新能力



部分能力在阈值前接近随机，超过阈值后显著提升。

可预测的规模效应与不可预测的涌现同时存在。

对齐：让预训练模型遵循指令与人类偏好

2023 年对话式 AI 爆发背后的经典工程链条

01
预训练底座
会续写

→

02
SFT
模仿人类示范

→

03
奖励模型
学习人类偏好

→

04
PPO / RLHF
优化更有用的回答

获得了什么

任务格式、对话语气、指令遵循与部分安全边界。

没有获得什么

“像好答案”不等于逻辑正确；幻觉与能力分布不均仍在。

2023：前沿能力与开放生态同时加速

一条曲线向上突破，一条曲线向外扩散

前沿闭源 · GPT-4

能力与工程可预测性

- 图像 + 文本输入，文本输出。
- 预训练后经 RLHF 对齐。
- 多种专业与学术基准显著跃升。

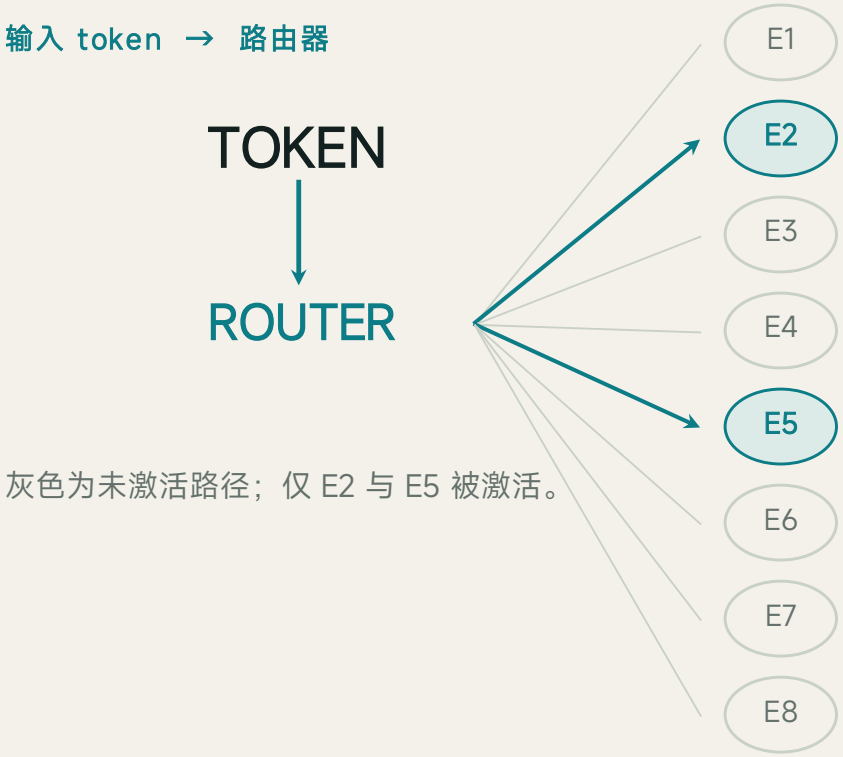
开放权重 · LLAMA 2

生态与复现扩散

- 开放底座与对话微调版本。
- 更多团队可以研究、微调与部署。
- 技术扩散本身也是创新加速度。

DeepSeek-V3：容量扩大，单次计算不必同步放大

MoE 让每个 token 只激活一小群专家



灰色为未激活路径；仅 E2 与 E5 被激活。

5.5%

V3 每个 token 激活的参数比例

671B → 37B

总参数 → 单 token 激活

14.8T

预训练 token

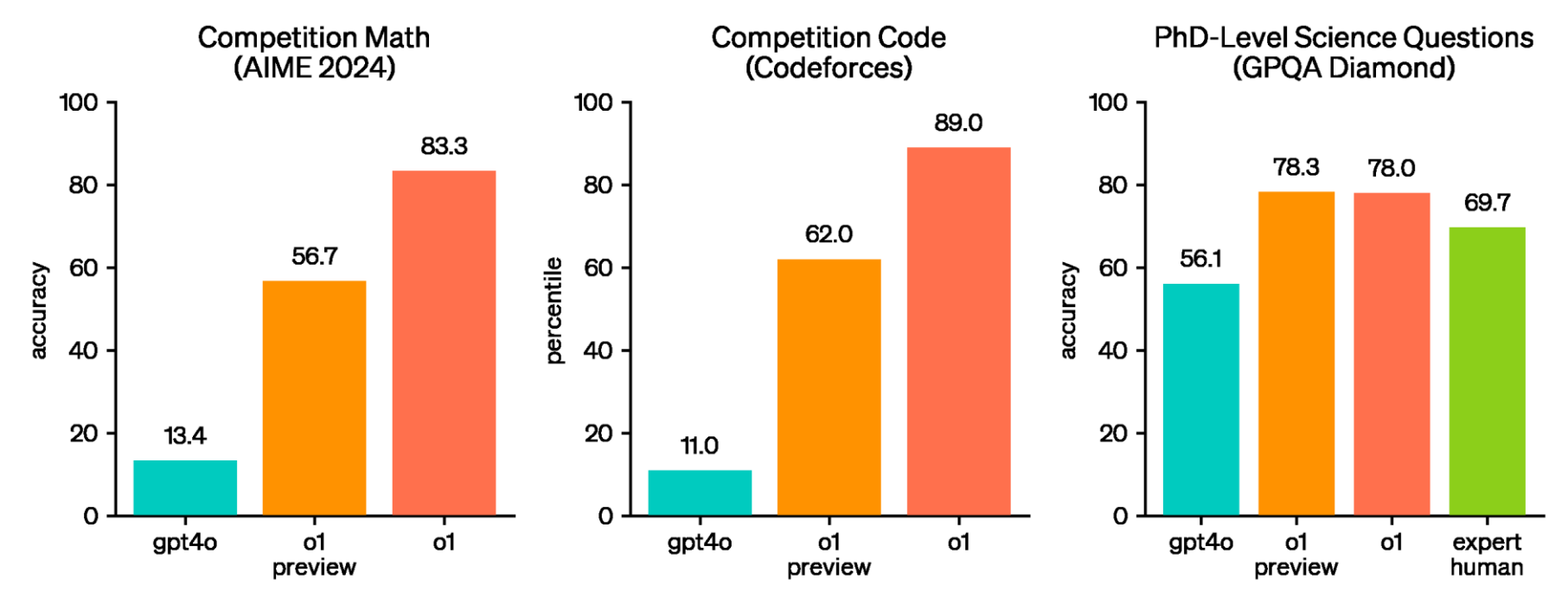
2.788M

完整训练 H800 GPU 小时

MLA、FP8 与算法—框架—硬件协同，共同降低有效训练成本。

o1：高难度推理基准表现显著提升

OpenAI 官方评估：AIME、Codeforces 与 GPQA Diamond，显示 o1 相对 GPT-4o 的改进

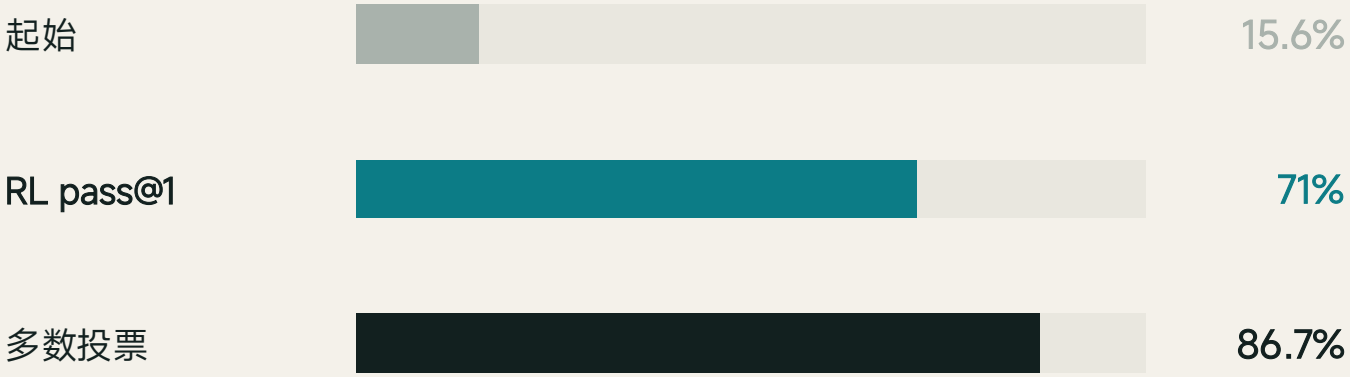


OpenAI 官方评测：AIME、Codeforces 与 GPQA Diamond。

DeepSeek-R1：可验证奖励让模型学会反思

R1-Zero 不先用 SFT 教推理；V3-Base + GRPO + 可验证奖励

AIME 2024 · PASS@1 / 多数投票



15.6 → 71.0

RL 后 pass@1 的变化

可自动判分的环境，把“探索更好的推理”变成训练信号。

R1-Zero 仍有语言混合、可读性差等问题；正式 R1 仍使用冷启动、多阶段 SFT 与 RL。

Kimi K2：训练目标开始指向智能体工作

训练数据不再只有静态问答，而是行动过程

1T / 32B

总参数 / 激活参数

15.5T

预训练 token

0

报告的 loss spike

- **MuonClip**

  提升 token 效率，以 QK-Clip 控制注意力 logit 失稳。

- **智能体数据合成**

  系统化生成工具、任务与多轮调用轨迹。

- **真实 + 合成环境 RL**

  从回答题目扩展到完成任务。

- **RLVR + 自我评价**

  可验证任务与开放任务使用不同奖励信号。

Kimi K3：扩展架构，也扩展轨迹长度

长上下文与长时程工具使用，需要架构和环境共同支撑

2.8T / 104B

总参数 / 激活参数

1M

token 上下文

约 2.5×

论文报告的整体扩展效率提升

基础架构

- Kimi Delta Attention：高效长序列信息流。
- Attention Residuals：跨深度信息选择。
- Stable LatentMoE：896 个路由专家，每 token 激活 16 个。

训练环境

- 长时程编码、搜索与专业知识工作。
- 数百到数千次工具调用、百万累计 token 轨迹。
- 持久 rollout 与可恢复沙箱状态。

K3 后训练：九个专家，再蒸馏回一个统一模型

三个领域 × 三个推理强度 = 九个 RL 专家

01
SFT 冷启动
建立基础智能体能力

02
九个 RL 专家
通用 / 智能体 / 编码
× low / high / max

03
多教师在策略蒸馏
把专长合并为一个模型

04
统一 K3
部署与工具接口一致

先分领域和预算训练，再通过蒸馏整合；推理强度本身也成为训练目标。

智能体训练依赖模拟工作环境，并加速部分 AI 研发

GPU 之外，CPU、内存、调度与状态同步成为新的系统约束

环境扩展

- **Rollout 成为主要计算**

  长时程 RL 的大量计算用于推理与工具交互，不只是反向传播。

- **沙箱保存工作状态**

  文件、进程、工具结果与任务状态都需要持久化。

- **运行时成为瓶颈**

  CPU、内存、调度与状态同步进入训练基础设施。

26%

2026 年 8 月，Anthropic 内部测量中由 Claude 主导的模型 R&D 工作比例

2 月不足 1%。公司自测口径；没有完全自主的工作。

模型 + 工具 + 运行时 + 环境

书生S2：通专融合大模型，开启科学智能体基座时代





四次扩展接力：从回答问题到完成工作

新增信号越来越依赖模型外部世界

扩展轴	典型阶段	新增训练信号 / 工程	代表案例
参数 + 数据 + 算力	预训练规模化	下一 token、数据配比、规模定律	Chinchilla、GPT-4
指令 + 人类偏好	对话助手工程化	SFT、奖励模型、RLHF	InstructGPT、Llama 2
推理 + 测试时计算	推理模型	RLVR、反思、长链推理、采样	o1、DeepSeek-R1
工具 + 环境 + 轨迹	智能体模型	智能体数据、长时程 RL、沙箱	Kimi K2、K3

下一问：模型已经会推理，为什么还需要 Harness?

PART 02

02

Agent Harness 是什么

同一个模型，为什么会展现出不同的工作表现？

DATA

MODEL

TOOLS

DISCOVERY

Harness：连接模型、工具、状态与运行环境

用户看到一次回答；系统负责把回答变成一段可靠工作



- **上下文**
  放什么信息、何时压缩、如何记忆。
- **工具接口**
  模型能看、能写、能运行什么。
- **控制循环**
  下一步、重试、停止、并行与恢复。
- **运行边界**
  沙箱、权限、审批与安全策略。

Agent 的核心是环境反馈驱动的循环

模型决定下一步；环境告诉它真实发生了什么



阶段 A · Context：让模型知道该知道什么

从写好指令，到按需取得外部信息

01

Prompt / RAG

检索外部资料，缓解知识截止与私有知识缺失。

02

ReAct

把推理与行动交替，让行动结果进入后续推理。

03

Toolformer

学习何时调用工具、填什么参数、如何使用结果。

用户问题 → 检索 / 工具 → 增强上下文 → 模型回答

模型不再只依赖参数内记忆。

阶段 B · Interface：让能力成为可靠接口

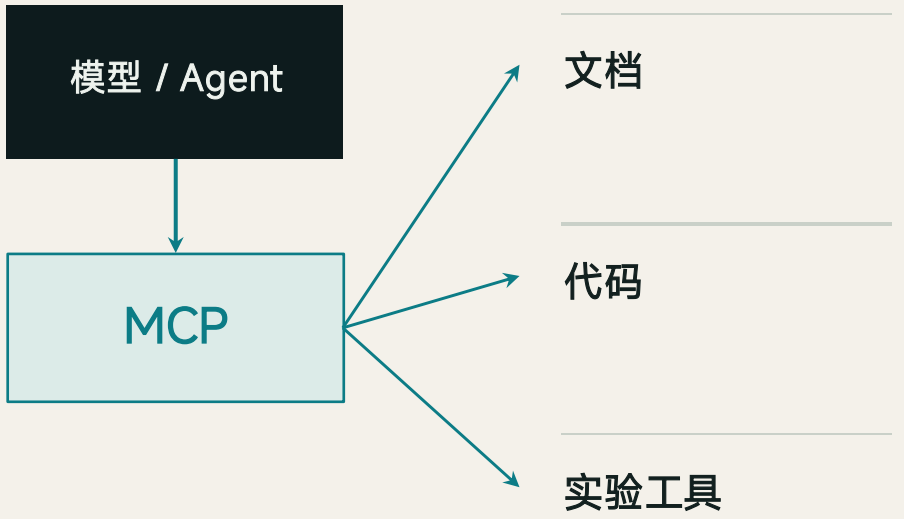
工具名称、参数、返回值与错误信息都要适合模型理解和恢复

AGENT–COMPUTER INTERFACE

- 文件、终端、浏览器与模拟器需要一致的操作边界。
- 错误信息要帮助模型恢复，而不只是给人看。
- 工具设计本身就是模型能力的一部分。

Anthropic 在 SWE-bench 实践中，曾花更多时间优化工具而非总提示词。

MCP · 标准化连接

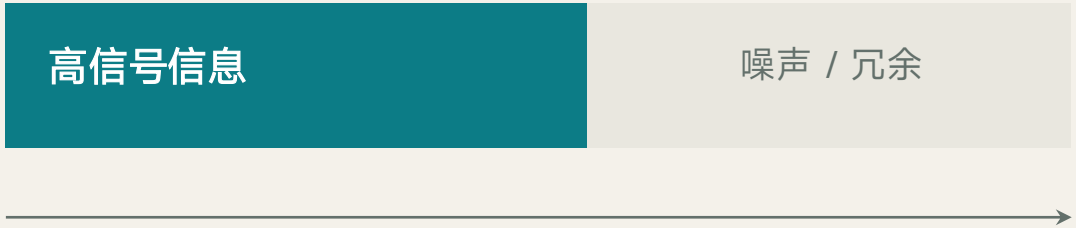


连接标准化，不自动保证工具好用、权限正确或结果可靠。

上下文工程：窗口很大，注意力仍然有限

容量是必要条件；每一步保留最高信号的信息，才是工作条件

CONTEXT BUDGET



上下文变长，recall 与推理精度可能下降（context rot）。

窗口扩大，不等于可以放弃信息管理。

- **即时检索**

  保存路径与链接，需要时再读取。

- **压缩 Compaction**

  保留决策与未决问题，丢弃冗余输出。

- **结构化笔记**

  把长期记忆放到窗口之外。

- **子智能体**

  让不同任务拥有独立、干净的上下文。

阶段 C · Loop：固定流水线与动态决策各有边界

能用简单 Workflow，就不必过度建设 Agent

WORKFLOW · 路径由人预先写死

适合步骤明确、需要稳定复现的任务。

链式处理、路由、固定并行。

AGENT · 路径由模型动态决定

适合步数不可预知的开放问题。

根据环境反馈调整下一步。

01

生成候选

模型提出一个版本

02

评估 / 验证

环境给出可测量反馈

03

达到标准

交付；否则带着反馈再来一轮

多智能体：扩大研究覆盖率，同时显著增加 token 成本

适合高价值、可并行、信息超出单窗口的任务



并行搜索 → 压缩结果 → 主 Agent 综合

+90.2%

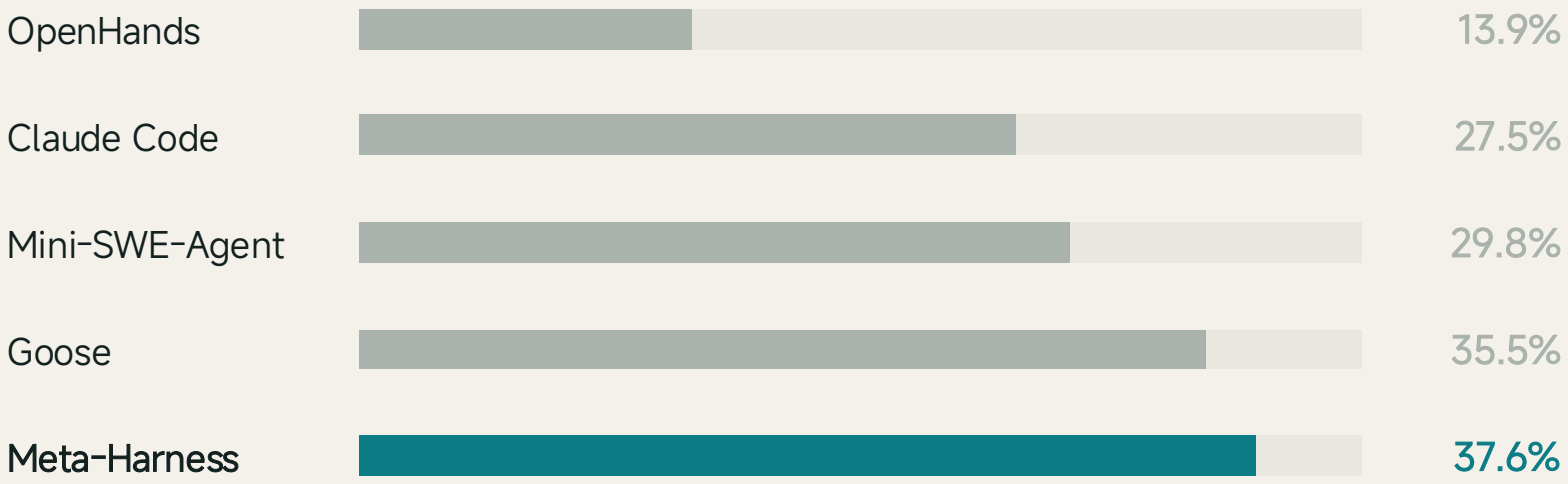
研究评测中，多智能体相对单 Agent Opus 4 的提升

约 15×

多智能体系统相对普通聊天的 token 消耗

固定模型条件下，Harness 显著影响任务表现

TerminalBench-2 · 固定 Claude Haiku 4.5



+10.1 pp

Claude Code 27.5% → Meta-Harness 37.6%

模型权重没变；改变的是信息存取、工具执行与循环策略。

比较模型必须同时披露 Harness、工具、预算与提示词。

Meta-Harness：自动优化模型外围的执行代码

外循环优化模型周围的代码，与训练模型权重是两条不同路径

01 **读取历史**

代码 + 轨迹 + 分数

02 **提出新 Harness**

修改外围执行代码

03 **运行评测**

在任务上比较结果

04 **保存经验**

全部历史进入下一轮

+7.7 pp

在线文本分类，相对 ACE

4×

上下文 token 更少

+4.7 pp

检索增强数学，五个留出模型平均

RSI：将自我修改纳入可验证的优化循环

改动进入下一轮系统，并用后续任务表现验证，形成递归自我改进

01 · 修正产物

代码、证明或实验方案变得更好

02 · 修改模型权重或Harness

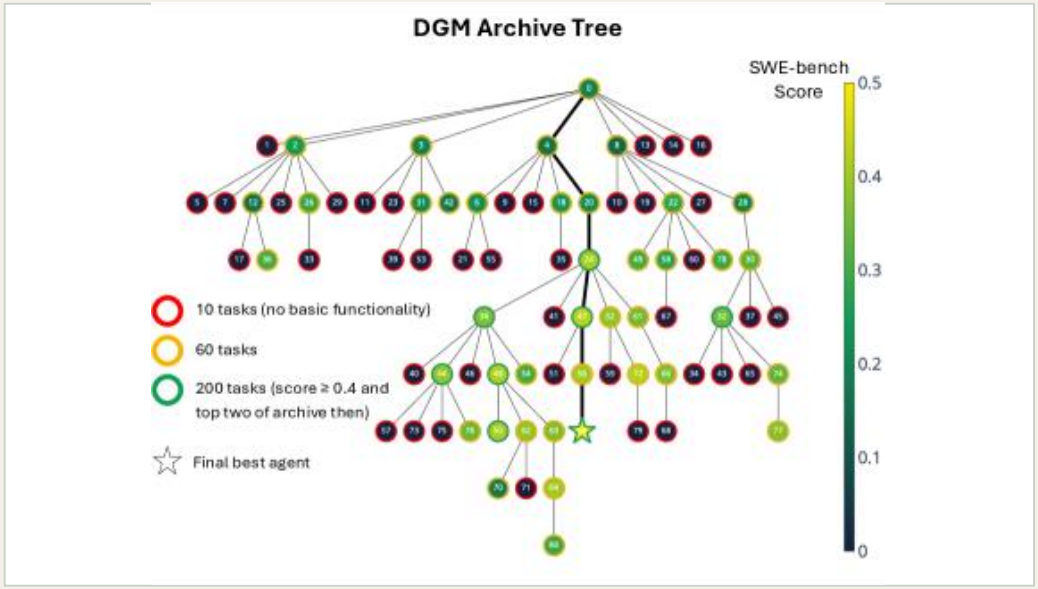
模型权重；工具、上下文与工作流进入优化循环

03 · 改进研究过程

研究代理自身的代码成为被优化对象

关键：经过验证的改动成为下一轮系统的一部分。

DGM · OPEN-ENDED ARCHIVE



每个节点都是一个可继续自我修改的编码代理

20.0% → 50.0%
SWE-bench

14.2% → 30.7%
Polyglot

前沿 RSI：AIDE² 连续 8 天改写研究代理

外环修改研究代理，内环在 AI R&D 任务上评分；接受改写成为下一轮起点

01 **修改代理代码**
搜索策略、上下文与反馈机制

02 **固定预算评测**
机器学习、启发式算法与 Harness 任务

03 **隐藏数据筛选**
仅保留私有评分更优的代理

04 **进入下一轮**
被接受的代理成为新的修改起点

0.703 → 0.778
选择基准上的私有评分

8 天 / 7 次
连续接受的自我改写

55% → 32%
另一组留出任务的奖励投机率

AIDE85 在 4 个外部基准均匹配或超过人类工程代理。

点火测试：能否更好地驱动下一轮 RSI？三组种子的结果仍不确定。

科学智能体 = 科学基础模型 + 领域 Harness + 可验证环境

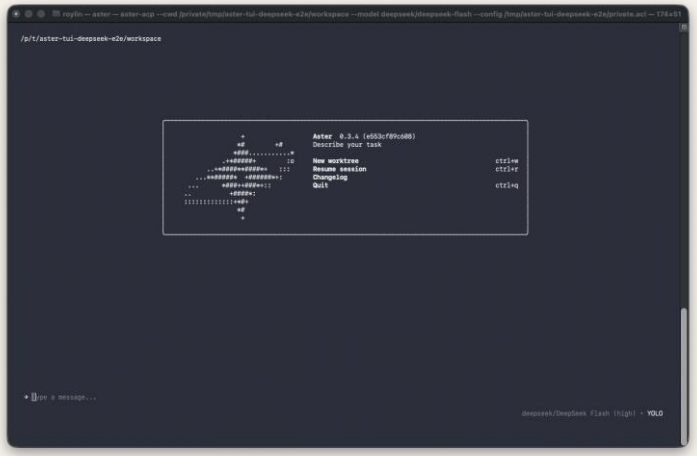
智能体轨迹可以回流，更新提示、工具、记忆，甚至模型权重

Harness 模块	科学研究对应物
Context	论文、数据、实验记录、失败历史
Tools	搜索、符号计算、模拟器、机器人
Loop	假设—计算—实验—再假设
Verifier	Lean、单元测试、DFT、XRD、临床终点

持久投入更集中在：环境接口、状态管理、可观测性、沙箱与权限。

书生·Aster Harness







https://aster-ai-lab.github.io/aster

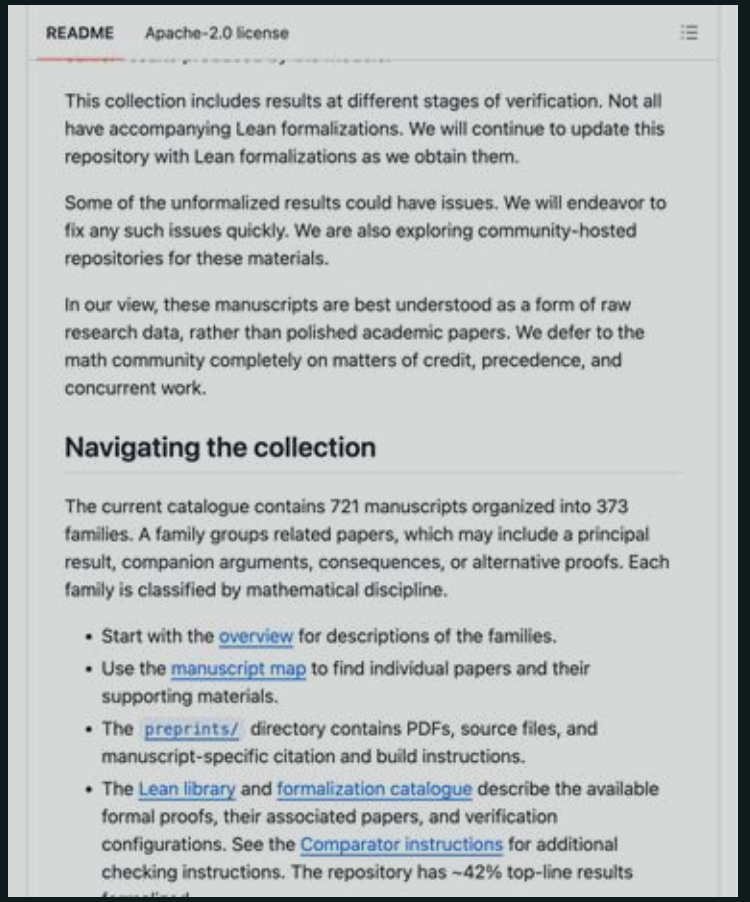
https://github.com/aster-ai-lab/aster

PART 03

03

AI 如何加速数学研究

从一位数学家的四年，到七百份手稿与三项新结果。



AI FOR SCIENCE · 2023—2026

陶哲轩的四年：能力评价与研究工作流的演变

能力判断持续上调；验证、披露与理解同时变得更迫切

2023	2024	2025	2026	此后
迷茫的本科生	平庸研究生	放牧一群猫	初级合著者	Math 2.0
实验、提示、Lean 入门	副驾驶、PFR、ETP	研究级辅助、评测纪律	自主解题、证明过剩	消化、理解、共同体

从辅助证明到批量产出：验证、消化与共同体治理成为中心。

2023 · 从直接求解转向让 AI 提供研究策略

早期模型更适合做建议者、初稿工具与共鸣板

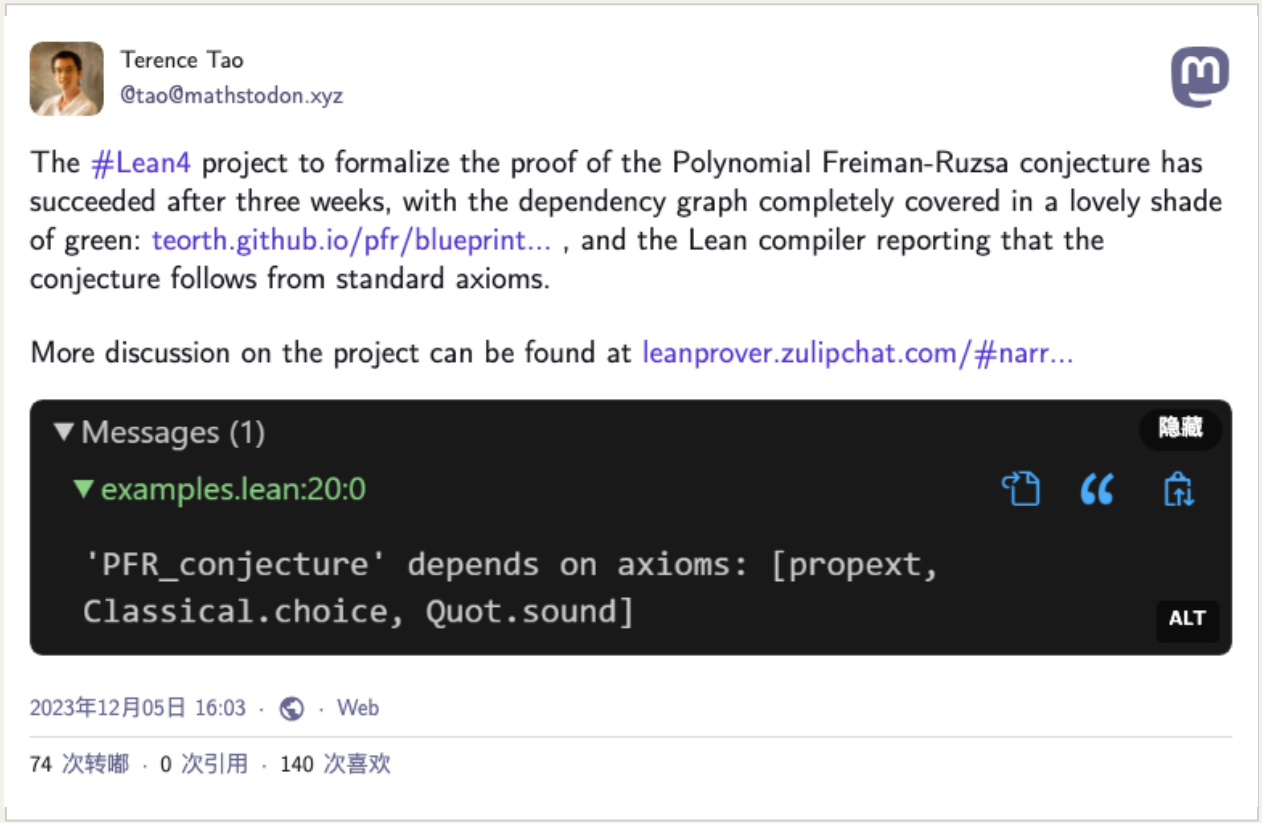
能力定位

- GPT-3.5 工具像“勤奋但有些迷茫的本科生”。
- 直接解研究题容易胡说，给出多种策略却可能命中关键路径。
- GPT-4 已能帮助比较方案、修改表达与探索形式化。

预言的真正结构

- **大模型**
  生成候选
- **搜索**
  寻找依据
- **Lean**
  验证证明
- **可信赖合著者**
  工具集成后的工作目标

PFR 三周形式化：Lean 支持可检查的协作证明



原始帖子与 Lean 依赖图截图。

3 周

PFR 猜想的 Lean 4 形式化完成

- 把证明拆成可认领、可检查的模块。
- 编译器检查每一步，依赖图全绿成为共同进度条。
- 人工数学证明在先；三周完成的是协作形式化。

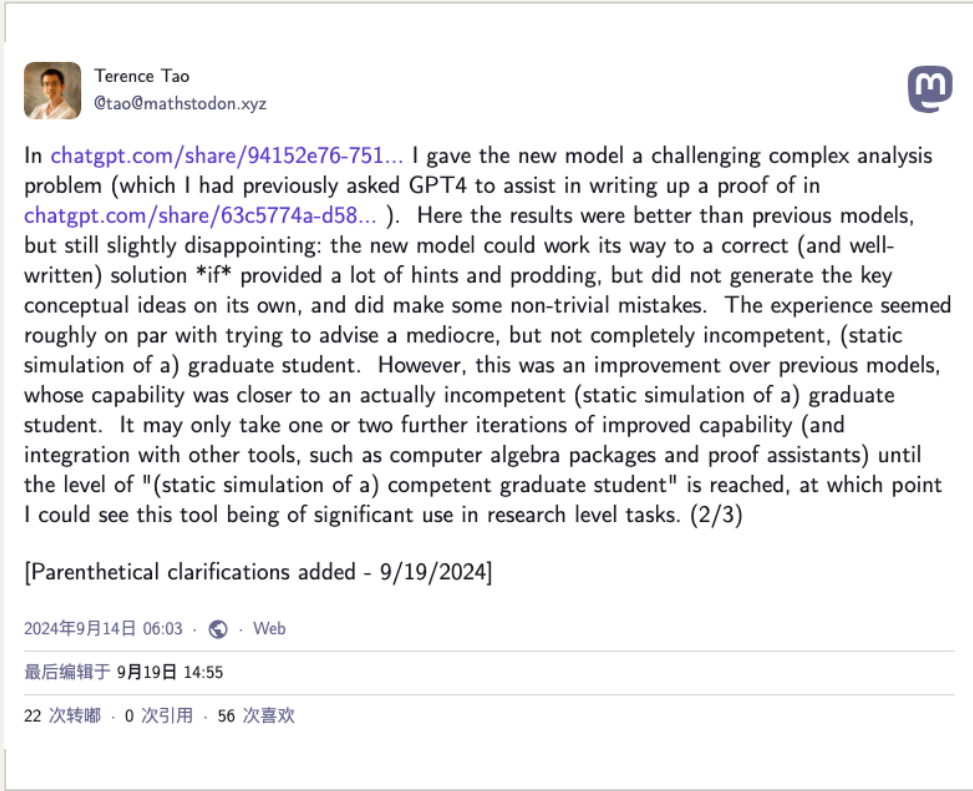
2024 · 能力升级，指导成本仍很高

“平庸研究生”必须放回具体评测语境理解

“平庸但并非完全无能的（静态模拟意义上的）研究生。”

- 能在大量提示下得到正确解答。
- 仍不会独立提出关键概念。
- 陶哲轩后来为学生分级可能造成的简化印象公开澄清。

同年 6 月：“它会很棒的副驾驶……方向盘还在我们手里。”



Terence Tao
@tao@mathstodon.xyz

In chatgpt.com/share/94152e76-751... I gave the new model a challenging complex analysis problem (which I had previously asked GPT4 to assist in writing up a proof of in chatgpt.com/share/63c5774a-d58...). Here the results were better than previous models, but still slightly disappointing: the new model could work its way to a correct (and well-written) solution *if* provided a lot of hints and prodding, but did not generate the key conceptual ideas on its own, and did make some non-trivial mistakes. The experience seemed roughly on par with trying to advise a mediocre, but not completely incompetent, (static simulation of a) graduate student. However, this was an improvement over previous models, whose capability was closer to an actually incompetent (static simulation of a) graduate student. It may only take one or two further iterations of improved capability (and integration with other tools, such as computer algebra packages and proof assistants) until the level of "(static simulation of a) competent graduate student" is reached, at which point I could see this tool being of significant use in research level tasks. (2/3)

[Parenthetical clarifications added - 9/19/2024]

2024年9月14日 06:03 · Web

最后编辑于 9月19日 14:55

22 次转嘟 · 0 次引用 · 56 次喜欢

等式理论计划：数学开始像大科学工程

大问题被拆成可验证的小任务，用项目图谱管理依赖

4,694

最简 magmas 等式定律

22,028,942

定律之间的蕴含关系

Lean

全部形式化的共同目标

- **多种参与者**

  专业数学家、公众、自动定理证明器与 AI 工具共同参与。

- **新型组织方式**

  任务、依赖和验证状态可以公开展示与协作。

- **数据成为资源**

  项目数据也服务后续 AI 数学工具评测。

冷静复盘：早期现代 LLM 并非主要功臣；传统自动定理证明器与人类协作同样关键。

2025 · 局部任务提速，整体研究管理仍然复杂

能力不是一条平滑斜线，而是一张锯齿状地图

一个具体代码任务

约 1 小时 → 约 10 分钟

人工完成

生成、测试、修改、报告

o3-mini 辅助研究中的代码任务；AI 纠错，人也纠正 AI。

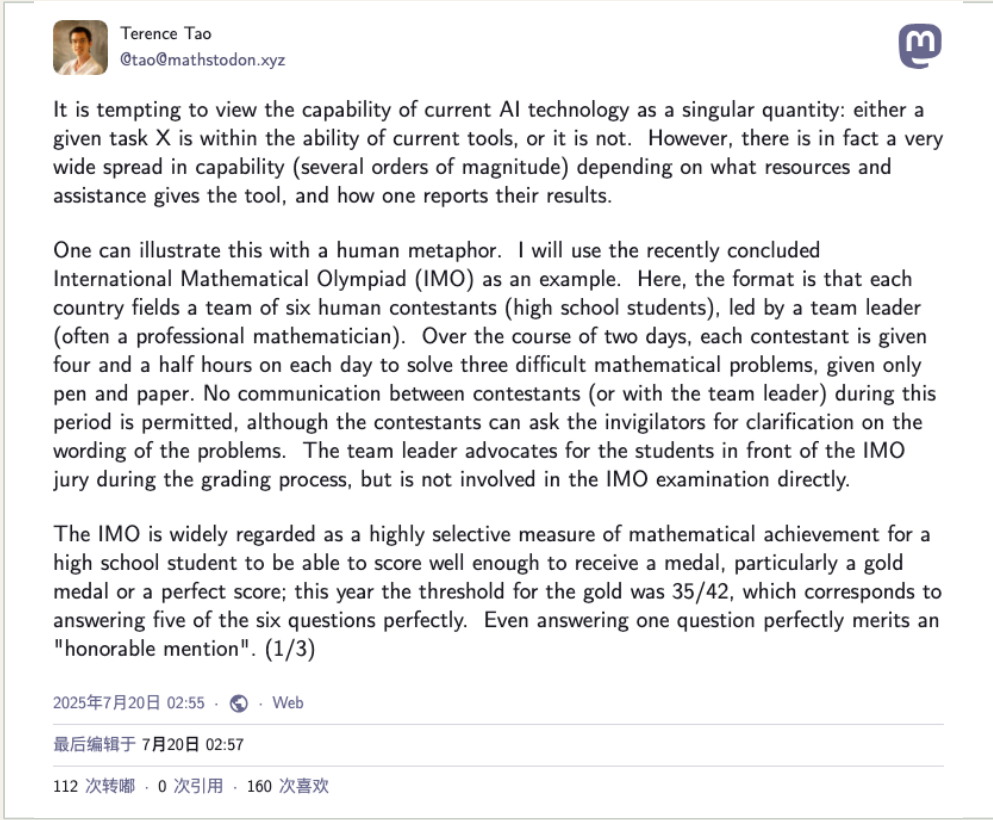
整体研究体验

“它就像放牧一群猫……有时候比自己单干还累。”

任务、尺度、提示与验证成本不同，两种体验并不矛盾。

IMO 金牌：必须同时披露资源、工具与评分口径

- 算力、工具、形式化协助、尝试次数与评分口径，必须一起披露。
- 陶哲轩对未在赛前披露方法的自报竞赛成绩不予置评。
- 还应报告成功率、每题成本、平均解题成本与失败次数。



Terence Tao
@tao@mathstodon.xyz

It is tempting to view the capability of current AI technology as a singular quantity: either a given task X is within the ability of current tools, or it is not. However, there is in fact a very wide spread in capability (several orders of magnitude) depending on what resources and assistance gives the tool, and how one reports their results.

One can illustrate this with a human metaphor. I will use the recently concluded International Mathematical Olympiad (IMO) as an example. Here, the format is that each country fields a team of six human contestants (high school students), led by a team leader (often a professional mathematician). Over the course of two days, each contestant is given four and a half hours on each day to solve three difficult mathematical problems, given only pen and paper. No communication between contestants (or with the team leader) during this period is permitted, although the contestants can ask the invigilators for clarification on the wording of the problems. The team leader advocates for the students in front of the IMO jury during the grading process, but is not involved in the IMO examination directly.

The IMO is widely regarded as a highly selective measure of mathematical achievement for a high school student to be able to score well enough to receive a medal, particularly a gold medal or a perfect score; this year the threshold for the gold was 35/42, which corresponds to answering five of the six questions perfectly. Even answering one question perfectly merits an "honorable mention". (1/3)

2025年7月20日 02:55 · Web

最后编辑于 7月20日 02:57

112 次转嘟 · 0 次引用 · 160 次喜欢

陶哲轩关于 IMO 评测规则的原始讨论。

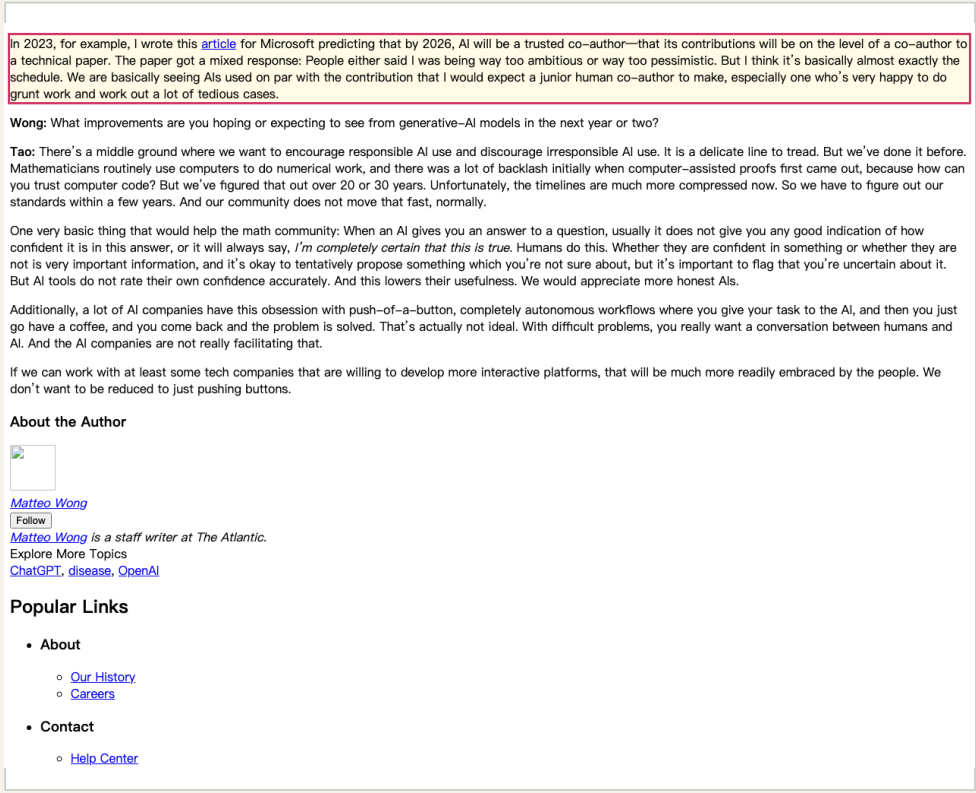
陶哲轩：AI 贡献接近初级人类合著者

贡献增加，仍不等于能力边界消失

“AI 的使用水平，大致相当于一位初级人类合著者能做出的贡献。”

- 能探索大量长尾问题，摘下其中较容易的一部分。
- 仍不能准确评估自己的置信度。
- 人类不想被降格为只按按钮的人。

另一场合的限定：“有用的助手，而非同侪。”



In 2023, for example, I wrote this article for Microsoft predicting that by 2026, AI will be a trusted co–author—that its contributions will be on the level of a co–author to a technical paper. The paper got a mixed response: People either said I was being way too ambitious or way too pessimistic. But I think it's basically almost exactly the schedule. We are basically seeing AIs used on par with the contribution that I would expect a junior human co–author to make, especially one who's very happy to do grunt work and work out a lot of tedious cases.

Wong: What improvements are you hoping or expecting to see from generative–AI models in the next year or two?

Tao: There's a middle ground where we want to encourage responsible AI use and discourage irresponsible AI use. It is a delicate line to tread. But we've done it before. Mathematicians routinely use computers to do numerical work, and there was a lot of backlash initially when computer–assisted proofs first came out, because how can you trust computer code? But we've figured that out over 20 or 30 years. Unfortunately, the timelines are much more compressed now. So we have to figure out our standards within a few years. And our community does not move that fast, normally.

One very basic thing that would help the math community: When an AI gives you an answer to a question, usually it does not give you any good indication of how confident it is in this answer, or it will always say, *I'm completely certain that this is true.* Humans do this. Whether they are confident in something or whether they are not is very important information, and it's okay to tentatively propose something which you're not sure about, but it's important to flag that you're uncertain about it. But AI tools do not rate their own confidence accurately. And this lowers their usefulness. We would appreciate more honest AIs.

Additionally, a lot of AI companies have this obsession with push–of–a–button, completely autonomous workflows where you give your task to the AI, and then you just go have a coffee, and you come back and the problem is solved. That's actually not ideal. With difficult problems, you really want a conversation between humans and AI. And the AI companies are not really facilitating that.

If we can work with at least some tech companies that are willing to develop more interactive platforms, that will be much more readily embraced by the people. We don't want to be reduced to just pushing buttons.

About the Author

Matteo Wong

Follow

Matteo Wong is a staff writer at The Atlantic.

Explore More Topics

ChatGPT, disease, OpenAI

Popular Links

- About
  - Our History
  - Careers
- Contact
  - Help Center

OpenAI 700+ 手稿：成果产出规模化，理解与消化成为瓶颈

手稿是原始研究数据，不是已经打磨完成的学术论文

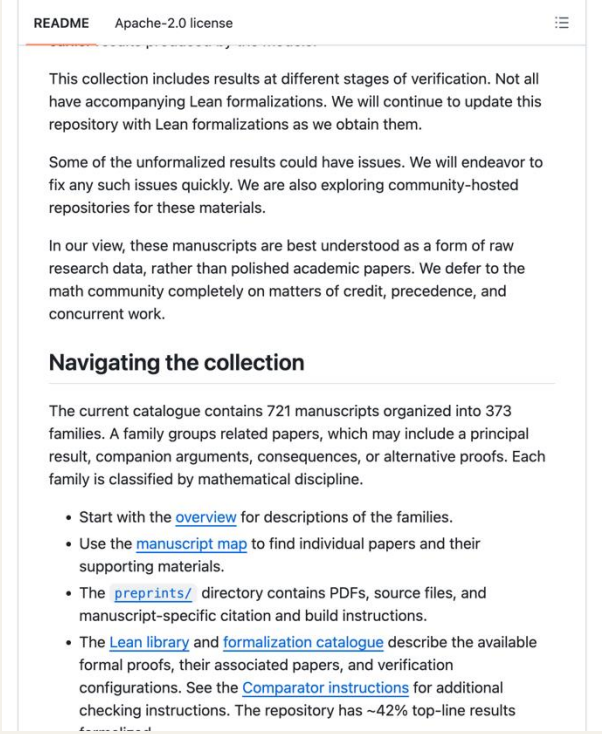
721

份数学手稿

373 个结果族

≈ 4,000 个开放问题被尝试

≈ 42% 顶级结果已形式化



README Apache-2.0 license

This collection includes results at different stages of verification. Not all have accompanying Lean formalizations. We will continue to update this repository with Lean formalizations as we obtain them.

Some of the unformalized results could have issues. We will endeavor to fix any such issues quickly. We are also exploring community-hosted repositories for these materials.

In our view, these manuscripts are best understood as a form of raw research data, rather than polished academic papers. We defer to the math community completely on matters of credit, precedence, and concurrent work.

Navigating the collection

The current catalogue contains 721 manuscripts organized into 373 families. A family groups related papers, which may include a principal result, companion arguments, consequences, or alternative proofs. Each family is classified by mathematical discipline.

- Start with the overview for descriptions of the families.
- Use the manuscript map to find individual papers and their supporting materials.
- The `preprints/` directory contains PDFs, source files, and manuscript-specific citation and build instructions.
- The Lean library and formalization catalogue describe the available formal proofs, their associated papers, and verification configurations. See the Comparator instructions for additional checking instructions. The repository has ~42% top-line results

OpenAI 官方仓库：原始手稿、Lean 与版本记录

数学 · 局部平滑猜想：一个辅助 PDE 贯通三个猜想

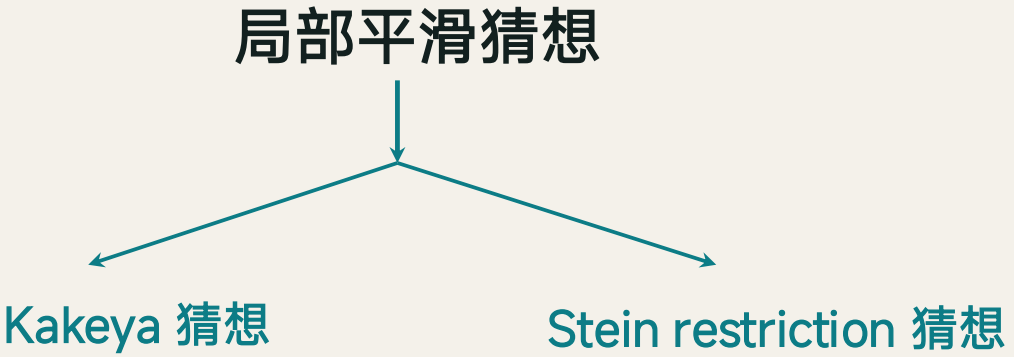
模型给出所有维度的证明；Kakeya 与 Stein restriction 随之成立

LOCAL SMOOTHING

所有维度

最极端的波聚焦必须短暂；平均意义下的聚焦，远小于最坏时刻。

新的辅助偏微分方程，是这条证明路径的核心。



10.18 AI 输出 → 10.19 Lean 形式化

发布策略：公开原始输出与 Lean，让研究者自行组织解释与拓展。

物理 · QCD：四维交互量子场论的严格存在性

结果同时覆盖 QCD、标准模型的统一补全，以及纯 Yang–Mills 存在性子问题

4D

QCD

规范场 + 费米子
四维相互作用量子场论

纯 Yang–Mills 存在性，是 Clay 问题的子问题。

VERIFICATION PATH

Lean 形式化

杨–米尔斯存在性的形式化于 10 月 16 日完成。

外部专家组

David Gross 牵头。专家组暂不公开 AI 原始产物，而是自行撰写、解释与核查。

不能把子问题等同于全部千禧年问题已经解决。

理论计算机 · OpenAI 报告：指数时间假设为假

自然语言反证与 Lean 形式化均已公开；外部专家组正在复核与解释

FALSE*

EXPONENTIAL TIME HYPOTHESIS

这一假设曾用于数百篇条件复杂性下界论文。证伪重新打开了更快算法的可能性，但实际算法改进仍需进一步研究。

*OpenAI 发布的结果，不作为专家组最终结论。

PUBLIC ARTIFACTS

10.16 自然语言反证

↓

10.17 Lean 形式化

公开复核与解释

Russell Impagliazzo、Ramamohan Paturi、Ryan Williams 牵头；原始数据公开，并邀请社区参与公开审查。

证明生成、证明验证、证明消化：第三步最慢

形式正确的答案，不自动成为人类共同体的知识

01

证明生成

候选证明、构造与反例
越来越便宜

02

证明验证

Lean / 形式方法
可以自动检查

03

证明消化

解释、联系、简化与迁移
仍然很慢

验证回答

每一步是否遵循规则?

消化回答

为什么这样想？它与过去和未来有什么关系?

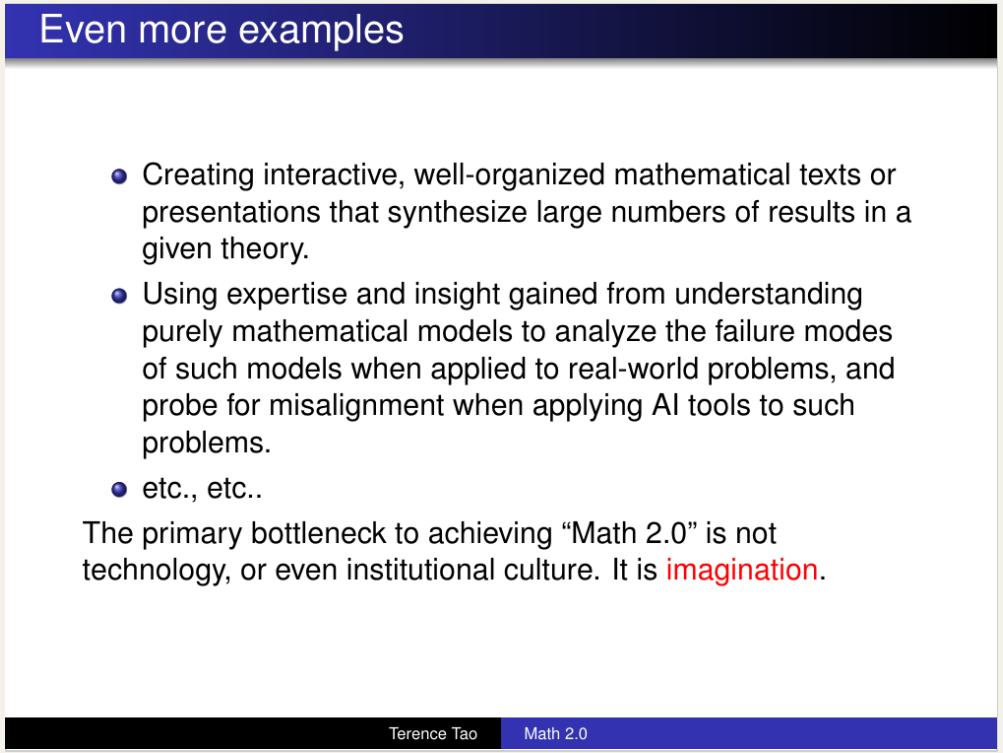
Math 2.0：把验证、消化与组织纳入数学贡献

陶哲轩：AI 应帮助数学共同体理解，而不只是自动解决开放问题

90,000 → 15,000 行

Sendov 原 AI 辅助证明 → 消化后重新形式化

- AI 可以同时加速证明生成与验证；但形式正确的文本，不会自动成为数学理解。
- 数学贡献还应包括验证、消化、组织、教学与提出新问题，信用不能只归给首次生成。
- 研究者需要把 AI 产物发展为可解释、可讲授、可迁移的公共知识，这也是共同体不可外包的责任。

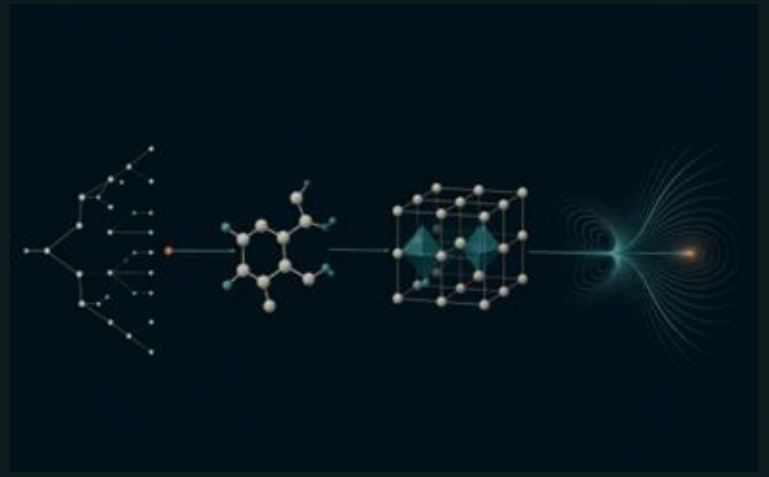


陶哲轩《Math 2.0》演讲：从生成结果走向理解与共同体。

PART 04

04

AI 如何加速科学发现



当科学拥有自己的工具、验证器和实验循环。

科学研究正在四个方向拓展认知边界

AI 帮助人类进入过去难以观测、计算与实验的疆域

向极宏观拓展，向极微观深入；
向极端条件迈进，向极综合交叉发力。

宏

极宏观

宇宙、气候、地球系统与复杂网络

微

极微观

粒子、分子、蛋白质与材料结构

极

极端条件

强场、高温、瞬态、稀有事件与分布外环境

交

综合交叉

数学、计算、实验与领域知识共同约束

不断拓展人类认知边界，让更多“难以观测、难以计算、难以实验”的问题进入可研究范围。

跨学科的共同机制：发现成为可迭代过程

验证器越清晰、反馈越快速，AI 的加速就越真实

01	02	03	04	05
科学表示	生成或近似	独立验证	主动学习	数据回流
公式 / 序列 / 晶体	假设 / 候选 / 模拟	约束 / 计算 / 实验	选择下一次尝试	成功 + 失败 + 不确定性

数学

Lean、形式证明与可执行构造

物理 / 材料

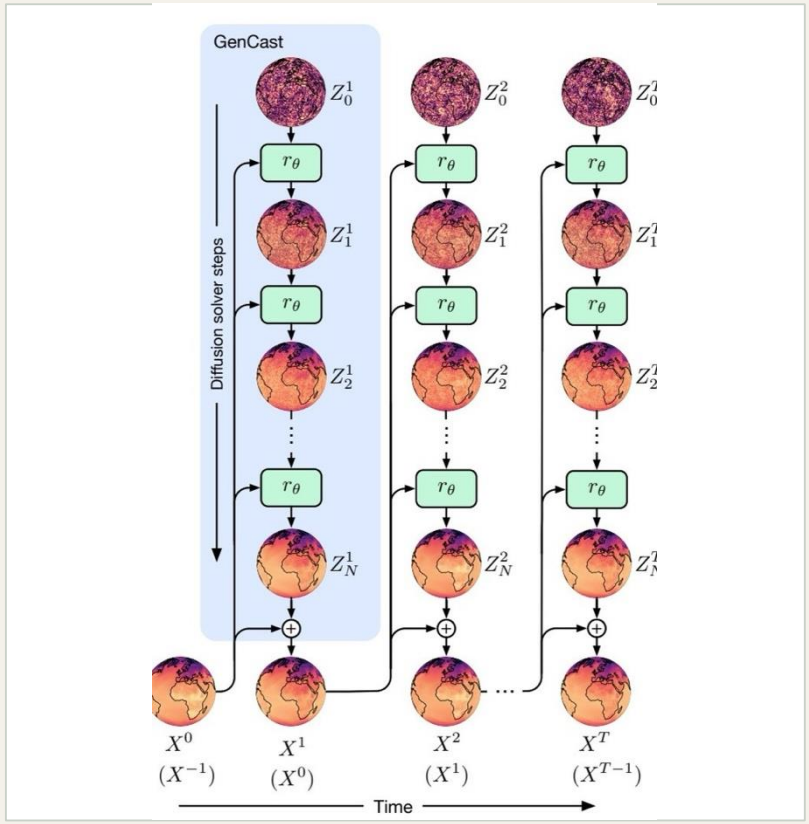
对称性、DFT、XRD 与机器人实验

生物医药

结构、细胞实验与临床终点

GenCast：用扩散模型加速全球集合天气预报

AI 将集合预报转化为扩散采样，在单个 Cloud TPUv5 上 8 分钟生成 15 天预报



GenCast 扩散求解与集合生成流程（原文裁剪）

15 天

全球集合天气预报

0.25°

经纬度分辨率

8 分钟

单个 Cloud TPUv5 生成一次预报

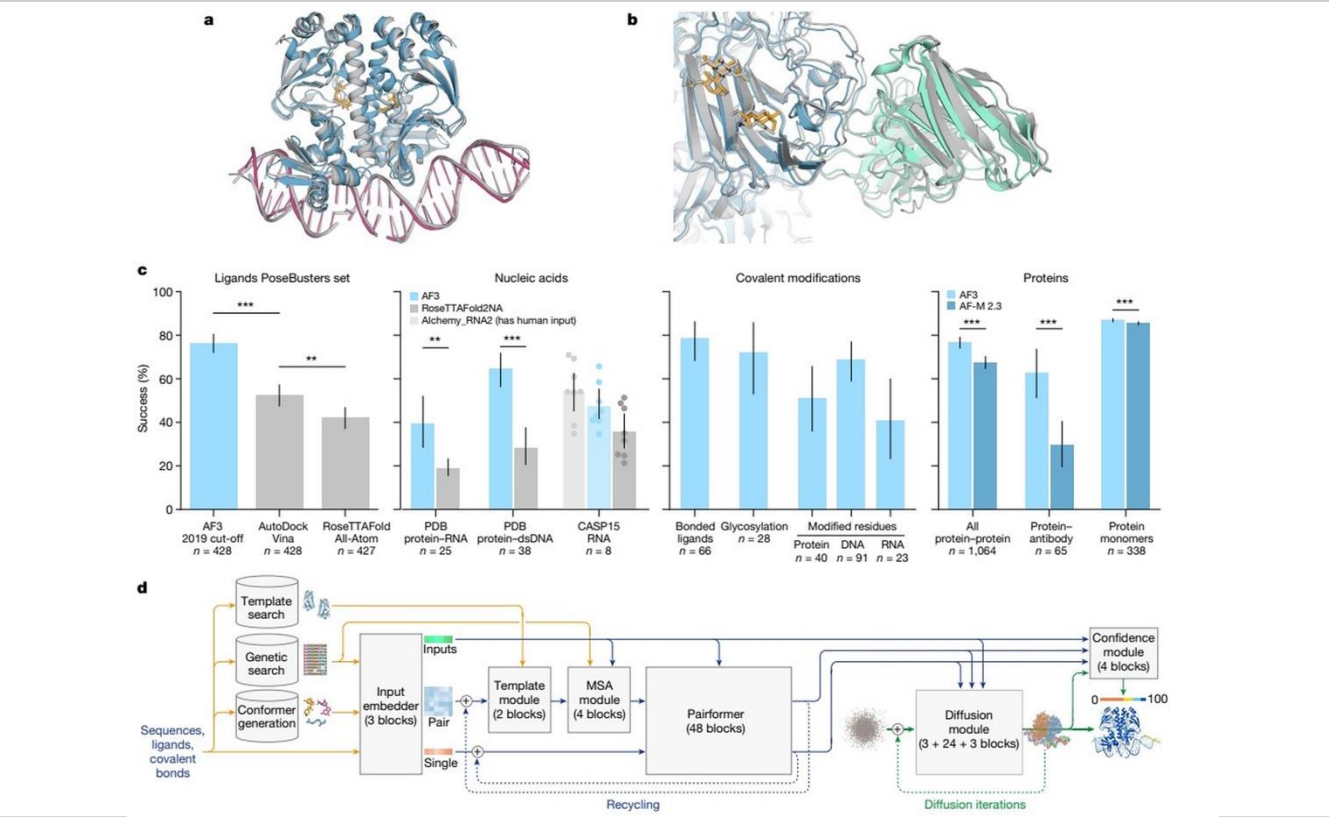
97.4%

1320 个评测目标中优于 ENS 的比例

加速集合生成与风险分析，而不是取代观测和传统初始条件。

AlphaFold 3：统一预测多类型生物分子复合物结构

AI 把结构预测从蛋白单体，推进到蛋白质、核酸与小分子的相互作用

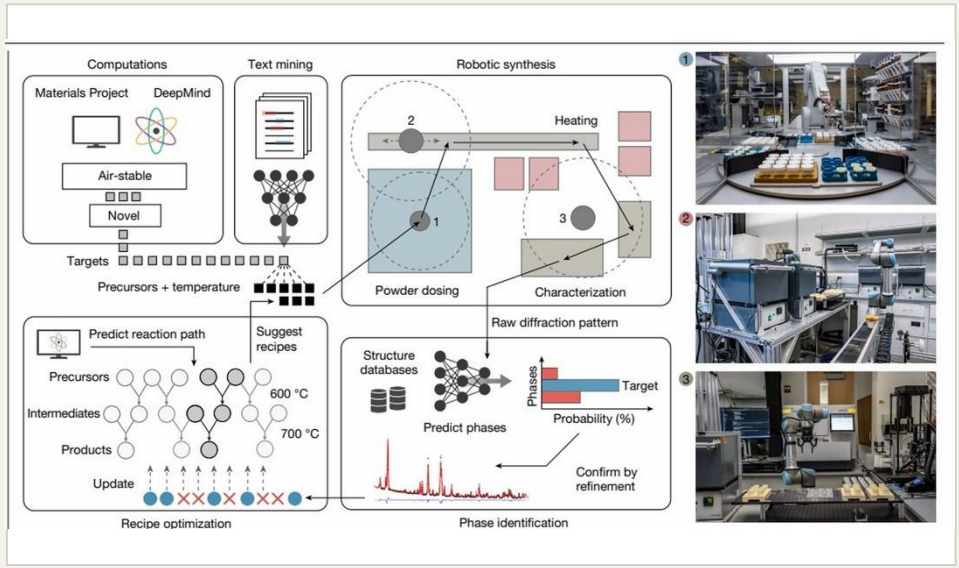


AlphaFold 3 的输入、扩散架构与复合物预测示例。

- 统一处理蛋白质、DNA/RNA、小分子配体、离子与修饰残基。
- 用全原子扩散架构生成复合物结构。
- 多类相互作用预测超过此前专用工具。

静态结构预测不自动等于细胞内动态、药物疗效或临床成功。

GNoME + A-Lab：规模化发现材料，并检验更正与新颖性



计算、文本挖掘、机器人合成、XRD 与主动学习。

2.2M / 380k

GNoME 新晶体候选 / 预测最稳定候选

57 / 36

A-Lab 目标数 / 成功合成数（更正后）

38 万是计算候选，不是已合成材料。旧报道常见 58/41；本页采用更正后的 57/36，新颖性与实用性仍有同行争议。

自动实验室：用主动学习选择下一个实验

每一次迭代减少不确定性，而不只是堆积样品数量

01 **模型提出候选**

目标 + 预期性质

02 **计算筛选**

稳定性 / 可合成性

03 **机器人实验**

合成 + 表征

04 **主动学习**

信息增益最大化

成功、失败与不确定性全部回流

负结果也是科研基础设施。

没有完整轨迹，主动学习会被成功案例带偏。

跨领域比较：加速相同，瓶颈不同

不同领域不能用同一倍数量化科学加速

领域	AI 明显加速	关键验证器	新的瓶颈
数学	候选证明、构造、形式化	Lean、形式逻辑、专家复核	证明消化、问题选择
理论物理	理论搜索、模拟、近似	对称性、守恒律、实验似然	原创问题、可解释性
天气 / 地球	集合预报、模拟采样	观测、回测、物理一致性	初始条件、分布外泛化
生物医药	结构、序列、分子与靶点	湿实验、毒理、临床终点	机制、临床成功率
材料	候选生成、稳定性筛选	DFT、XRD、合成与性能	可合成性、无序、负结果

答案生成越来越快；验证、实验与理解决定真正的科学吞吐量。

科学发现加速后，独立验证与披露必须同步加强

可靠、可迁移的理解，仍然需要共同体制度支撑

01 · 独立验证

引入编译器、物理约束与实验，避免模型自己给自己判分。

02 · 不确定性

报告置信区间、成功率、失败次数与资源消耗。

03 · 披露与复现

披露模型、Harness、工具、数据、提示和人工介入。

04 · 负结果

把失败实验与无效候选变成可共享的数据资产。

05 · 人类可解释

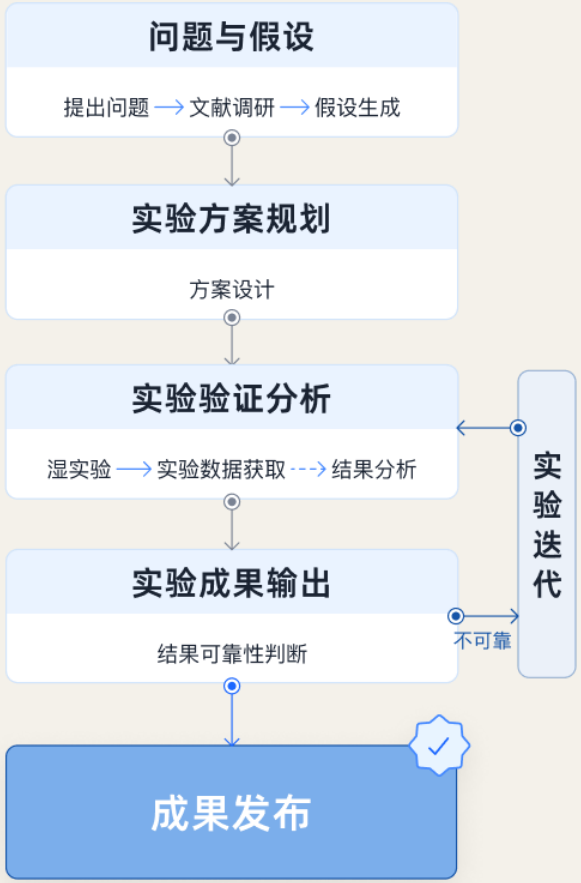
结果需要被专家理解、讲授、归因并放入领域脉络。

06 · 价值对齐

科学目标不是批量制造答案，而是增加可靠理解。

书生·端砚：构建自主进化的科研新范式





Science Harness

可信科研证据链

物理世界验证

书生·端砚：构建自主进化的科研新范式

面向科学家的统一智能入口

1 科学工作台

云端Web工作台

本地桌面工作台

启动研究资源 » 定义科学问题 » AI 联合研究 » 实验验证 » 成果发布

科研Token Plan

2 科学领域

生命科学



领域Skill

SCP

amix

科学数据集

学科方法论

证据研究

物质科学



材料知识　结构生成

第一性原理　AI 性质预测

催化与界面　表征解析

地球科学



观测数据　遥感处理

气候与分析　生态变化

实验与模拟　时空可视化

物理科学

空间科学

空气动力学

3 科学AI引擎

科研 Harness

跨轮次上下文　研究状态保持　多智能体架构　异构资源编排　多角色协同　实验反馈重规划　长期经验沉淀

可信科研证据链

交付对齐　评审智能体独立核验　执行复现

物理世界验证

统一协议　异构编排　执行与恢复　实验与迭代

4 模型基座

科学模型

面向长程复杂任务的智能体模型

世界模型

面向物理环境湿实验的具身操作模型

见证未来：三次正在发生的迁移

从模型能力，到科学工作方式的变化

01

从模型到系统

参数与数据仍重要；工具、状态与环境决定能力能否落地。

02

从答案到循环

单次回答越来越便宜；验证与实验反馈的周期成为杠杆。

03

从产量到理解

证明、论文与候选过剩；消化、解释与问题选择更稀缺。

更强模型 → 更好 Harness → 更严验证 → 更深理解

科学共同体的下一步：从产出结果到维护知识

AI 加速生成与验证；阐释、出版与经典化仍依赖长期的人类合作

01 **生成**
候选证明、理论与实验方案

02 **验证**
Lean、计算、实验与独立复核

03 **阐释**
解释、简化、讲授与建立联系

04 **出版**
归属、引用、同行评议与公开

05 **经典化**
进入教材、工具与公共知识

首发权之外，长期责任更重要

谁把结果讲到可理解，谁把它送入教学与公共知识，谁就在创造长期价值。信用会更多流向阐释、核验与维护。

未来需要更多负责任的阐释者、维护者与组织者。

研究训练也会改变

提出好问题、识别错误、整合证据、解释与教学，将成为更重要的研究技能；独立思考与失败经验仍需被保留。

当科学产出过剩，社会必须重新分配什么？

真正的长期影响发生在信任、职业、资源与公共治理

01 · 信任与信用

从“谁最先生成”转向“谁能核验、解释并维护”。完整的来源、证据链与责任归属会成为科研基础设施。

02 · 职业与教育

初级研究训练将更强调问题选择、批评、表达与合作；若只追求快速答案，技能形成与耐心会被侵蚀。

03 · 资源与权力

算力、数据、工具和实验平台可能集中研究议程。开放模型、共享验证与公共科研基础设施决定机会是否公平。

04 · 共同体与治理

哪些题适合自动化攻克，哪些应保留给人类学习与合作？负结果、研究边界与收益分配都将成为公共议题。

谁决定研究什么？谁有能力核验？谁从发现中受益？

AI 科学最终要回答：它让人的生活有什么不同?

当发现加速成为常态，科学的价值要落实到人的生命、能力与共同未来

真正被扩展的，是人类社会发现、验证与创新的能力。

更多知识 → **更强能力** → **更多选择** → **更大责任**

未知边界继续扩展

技术与生产被重塑

个体拥有新的可能

公共治理必须跟上

如果 AI 科学只增加成果数量，它可能只是加速既有体系；如果它同时降低验证和获取门槛，它才有机会扩大整个社会的科学能力。

最终的衡量标准：是否增进人的健康、尊严、自由与公共福祉。

超智元·SuperMind（微信公众号）

欢迎关注
超智元微信公众号



当 AI 走进实验室：2026 年八篇 Nature、Science 正刊...



Anthropic新研究扎心：你以为AI只是工具，它却在改变...



我梳理了陶哲轩这四年：从"迷茫的本科生"到"严重错位"



AI时代的劳动力市场：从职业替代到任务重组

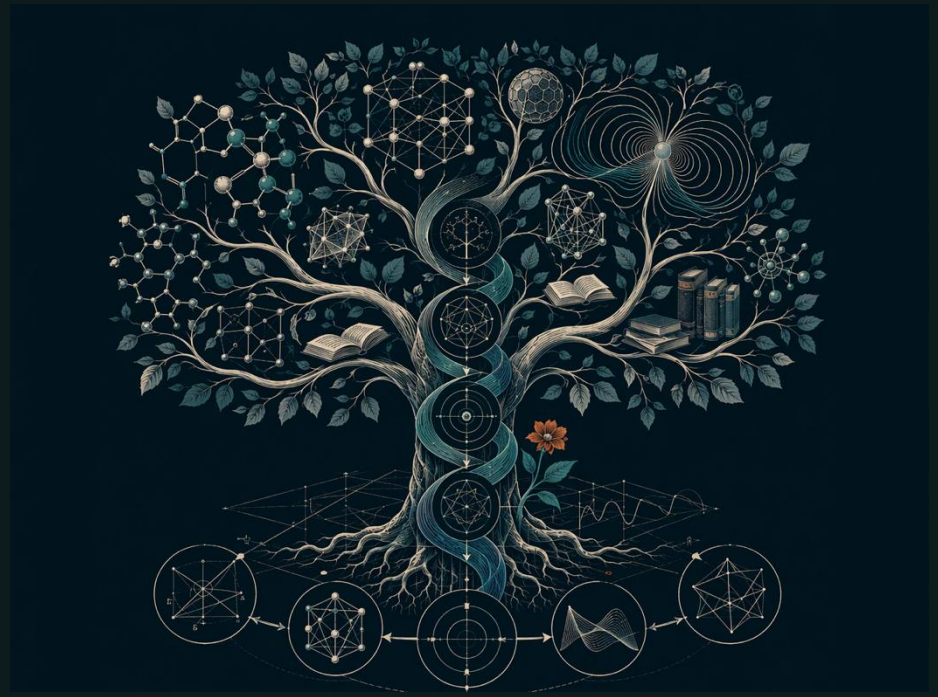
THE NEXT SCIENTIFIC WORKFLOW

未来，建造更好的科学进化系统

什么问题值得问，
什么证据足够可靠，
什么理解值得被传承。

生成可以过剩，理解不能稀缺。

人提出问题，机器扩展搜索；人设计验证，机器加速迭代；人赋予科学以意义。



生成、验证、理解与共享构成科学进化系统。

谢谢！

AI FOR SCIENCE · 2023—2026