

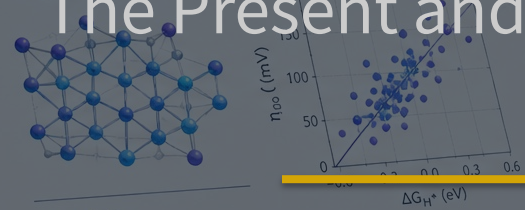
AI-Driven Discovery of Novel Catalysts for Clean Energy Conversion

X. Li, Y. Chen, M. Patel, J. Kim, et al.
Department of Materials Science and Engineering,
Future University

Abstract

We present an AI-driven framework that integrates graph neural networks and high-throughput simulations to discover novel catalysts for electrochemical water splitting. Our model predicts stable materials with high catalytic activity and reveals interpretable structure-property relationships that guide experimental validation. Several predicted materials exhibit superior performance, opening new opportunities for sustainable hydrogen production.

Graphical Abstract



Highlights

- AI framework accurately predicts catalyst activity
- Discovers novel materials with superior performance
- Reveals interpretable structure-property relationships
- Accelerates experimental validation and discovery

AI Scientist 的现在和未来

The Present and Future of AI Scientist

张奇

复旦大学
上海人工智能实验室

Outline / 目录

1

引言：为什么需要 AI Scientist

Introduction: Why AI Scientist?

2

前沿进展：AI Scientist 系统概览

State-of-the-Art AI Scientist Systems

3

学术论文介绍 I：Nature 发表的重要系统

Paper Intro I: Nature Publications

4

学术论文介绍 II：Agent 能力评诂与基准测试

Paper Intro II: Agent Benchmarks

5

WisPaper 功能介绍与 OpenNovelty

WisPaper Platform & OpenNovelty

6

未来展望与总结

Future Outlook & Summary

1

引言：为什么需要 AI Scientist

Introduction: Why AI Scientist?

2024 诺贝尔奖：AI 登上科学最高殿堂

诺贝尔化学奖

Demis Hassabis · John Jumper · David Baker

AlphaFold2 蛋白质结构预测 & 计算蛋白质设计

诺贝尔物理学奖

John Hopfield · Geoffrey Hinton

启发大脑的神经网络基础工作

 关键信号：历史上首次 AI 成果同时斩获两项诺贝尔奖——AI 正在成为科学发现的核心驱动力

人类科研的天花板在哪里？



论文爆炸

每年 500 万+

无法全面追踪



周期漫长

新药 10-15 年

从发现到上市



成本高昂

新药 26 亿\$

资源门槛极高



认知极限

超高维空间

人脑难以发现



重复危机

60-70%

不可重复

科学范式的演进：从实验科学到 AI 驱动的「第五范式」

第一范式

实验科学

千年前

观察归纳

第二范式

理论科学

数百年前

数学建模

第三范式

计算科学

数十年前

模拟仿真

第四范式

数据科学

十年前

大数据挖掘

第五范式

AI 驱动科学

现在

自主发现

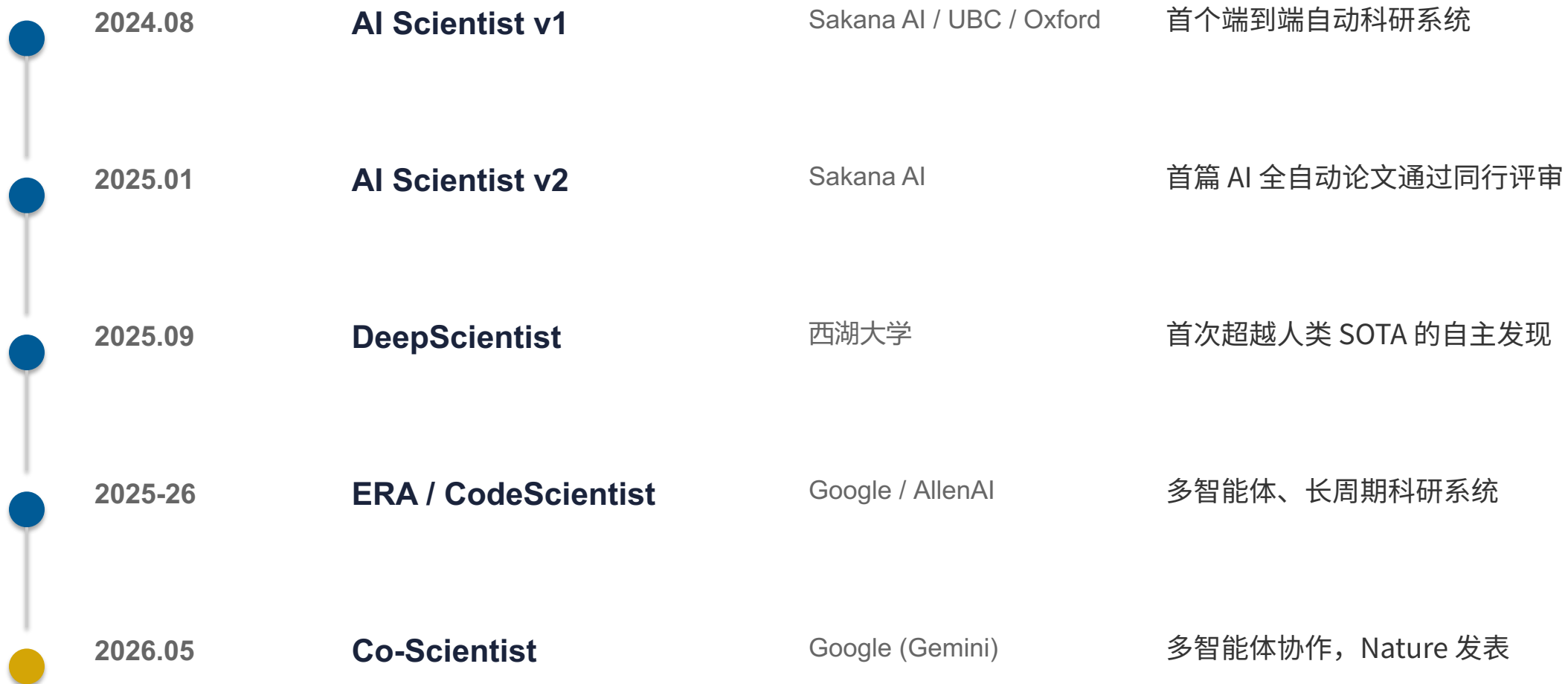
AI Scientist 的终极愿景：从 AI-Assisted → AI-Conducted → AI-Discovered

2

前沿进展：AI Scientist 系统概览

2024-2026 Key Systems Overview

LLM based AI Scientist 关键系统时间线



AI Scientist: 首个全流程自动科研系统

Sakana AI / UBC / Oxford · Nature 2026



技术基座

- Claude Sonnet, GPT-4o, o3, o4-mini
- Template-based: 人类提供研究脚手架
- Template-free: 完全开放式自主探索

核心创新

- Agentic Tree Search 渐进式搜索
- VLM 视觉反馈迭代优化图表
- Automated Reviewer 自动评审门控

ICLR 2025 Workshop — AI 论文首次通过同行评审

历史性时刻：零的突破

被接收论文

Compositional Regularization: Unexpected Obstacles
in Enhancing Neural Network Generalization

ICLR 2025 Workshop · 3 篇投稿中 1 篇通过

评审分数



均分 6.33/10 超过接收阈值

诚实透明

Sakana AI 接收后主动撤稿，披露 AI 身份

客观看待

Workshop 接收率 60-80% (主会 20-30%)

深远意义

AI 全自动研究首次被人类同行评审接受

2026 年的 AI Scientist: 能力边界

✓ 已经能做到

- 自主生成研究想法并编写/运行实验代码
- 撰写完整学术论文 (LaTeX)
- 在特定领域超越人类 SOTA (DeepScientist)
- 通过 Workshop 级别同行评审 (ICLR 2025)
- 多智能体协作科研 (Co-Scientist)
- 生物医学假说生成并实验验证

✗ 尚未做到

- 通过顶会主会议审稿
- 跨学科的通用科研能力
- 「湿实验」自主操作
- 真正理解科学概念 (vs. 模式匹配)
- 处理长周期复杂科研项目
- 独立提出范式级突破理论

3

学术论文介绍 I: Nature 发表的重要系统

Co-Scientist & ERA (Nature 2026)

Co-Scientist: Multi-Agent Hypothesis Generation (Nature 2026)

Google Research / Google DeepMind / Insilico Medicine / Stanford

Nature, May 2026 | 多智能体协作假说生成系统

► 科学进步的内核在于提出假设，现有 LLM 存在两个核心痛点：

- 幻觉与缺乏严谨性：容易生成看似科学实则无法实现的方案。
- 缺乏计算扩展性：简单地增加模型参数并不等同于深度思考。

Co-Scientist 试图通过结构化思维引擎解决这些问题

► 核心架构：多智能体“科学方法论”

- Generation Agent (生成者)：负责跨学科搜索并提出初步假设。
- Reflection Agent (反思者)：扮演“同行评审员”，批判假设的正确性与新颖性，并使用外挂工具查重。
- Evolution Agent (进化者)：基于辩论反馈，对高质量假设进行迭代优化和交叉杂交。
- Proximity Check Agent (拓扑邻近者)：生成假设的邻近图，相似想法的聚类、去重以及对假设空间的卓越高效探索。
- Meta-review Agent (评价者)：综合评价反馈与辩论规律，迭代优化各代理性能。
- Ranking Agent (排序者)：引入 Elo 锦标赛机制，让不同假设进行“科学辩论”，优胜劣汰。

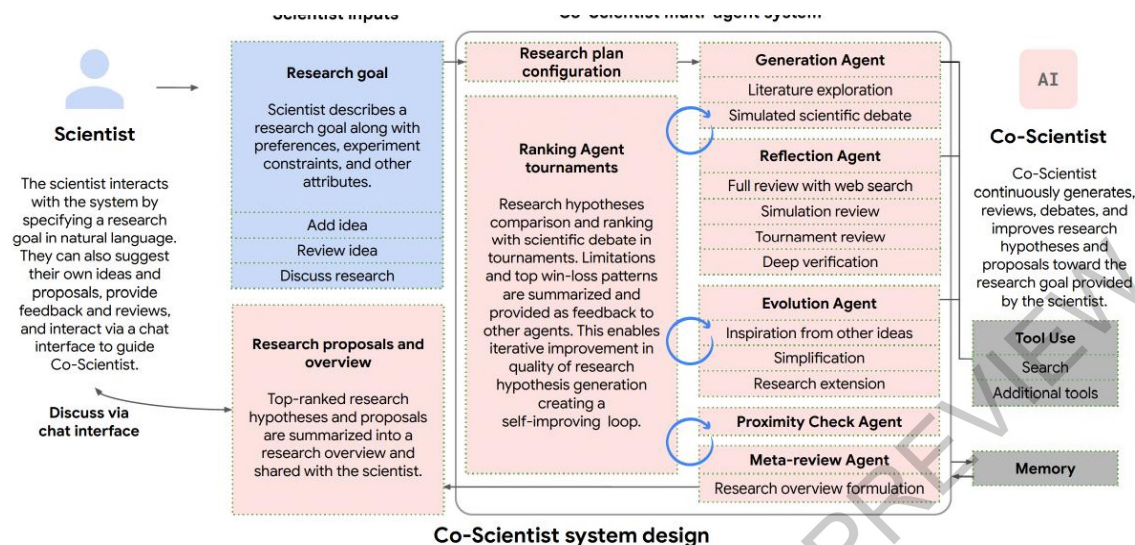


Fig. 1: Co-Scientist system architecture and validation

Co-Scientist: Multi-Agent Hypothesis Generation (Nature 2026)

Google Research / Google DeepMind / Insilico Medicine / Stanford

Nature, May 2026 | 多智能体协作假说生成系统

► 湿实验验证：从硅片到培养皿

- 发现治疗急性髓系白血病 (AML) 的新药物重定位候选
- 发现新的协同组合疗法方案
- 通过体外实验验证 AI 提出的假说

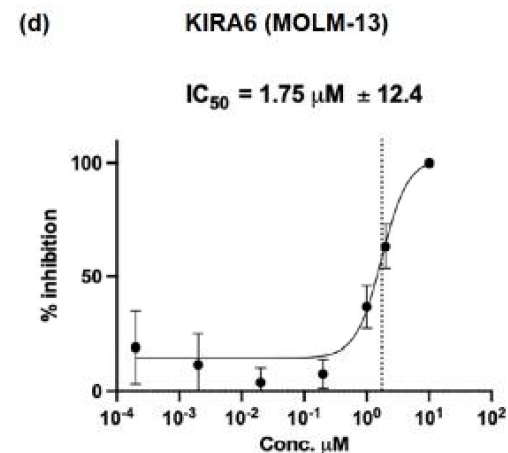


图 3: KIRA6 药物在不同 AML 细胞系中的剂量响应曲线, 验证了 AI 预测的选择性毒性

更令人惊叹的是在**多药联合治疗**上的表现。寻找 2-3 种药物的协同效应是一个指数级增长的搜索空间, 而 Co-Scientist 成功预测了 JNJ-64619178 与 Selinexor 等组合的强协同作用, 这通过传统的人工筛选几乎是不可能的。

Co-Scientist: Multi-Agent Hypothesis Generation (Nature 2026)

Google Research / Google DeepMind / Insilico Medicine / Stanford

Nature, May 2026 | 多智能体协作假说生成系统

► 关键启示:

- **Inductive Bias (归纳偏置):** 将“科学辩论”逻辑嵌入智能体交互, 比单纯提升模型参数更高效。
- **人机协作:** 系统允许科学家在关键节点修改“研究计划配置”, 实现了真正的“人在回路”指导。

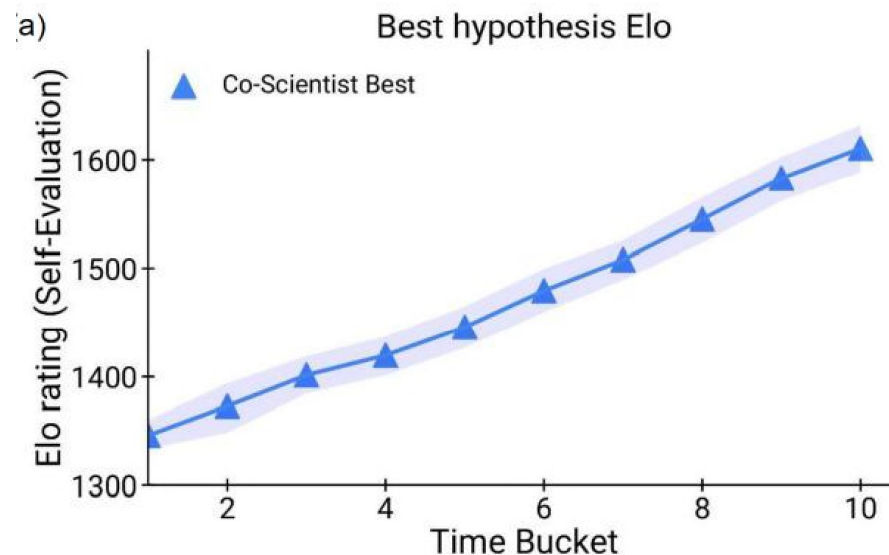
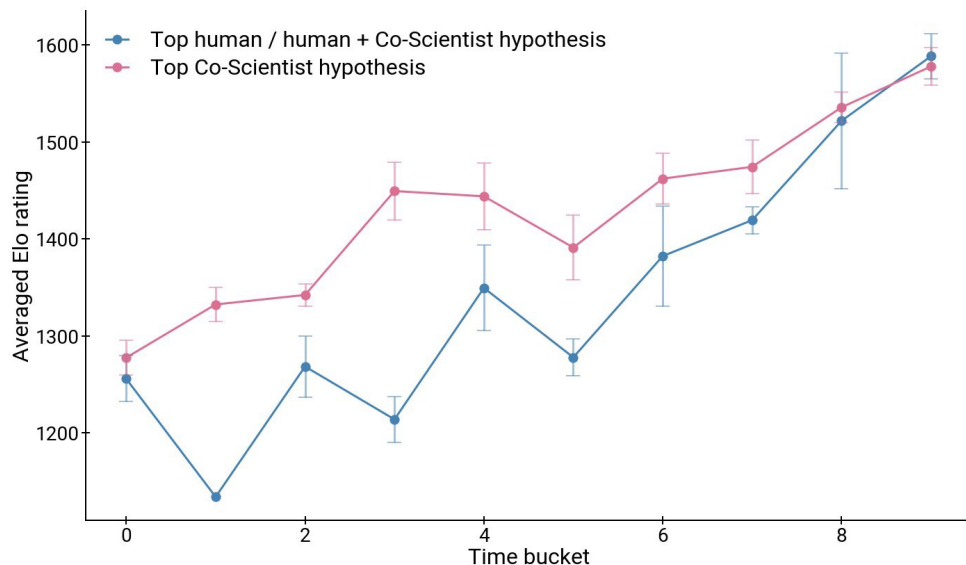


图 2: 测试时计算扩展带来的 Elo 评分线性增长, 展示了系统自我改进的潜力

Co-Scientist 证明了 AI 已具备在复杂生物医学领域进行“原创性推理”初级能力。

ERA: LLM + Tree Search for Expert-Level Scientific Software (Nature 2026)

Google DeepMind

Nature, May 2026 | LLM + 树搜索创建专家级科学软件

► 核心挑战：科学软件开发的“迭代困局”

在许多科学领域，研究的核心往往在于设计能够最大化某种“质量分数”的经验软件（Empirical Software）。

例如，为了预测蛋白质结构、模拟大气流动或分析复杂的基因组轨迹。

然而，现状却极度低效：

- **开发周期长**：构建一个健壮的实验代码库通常需要数年的领域积累。
- **搜索空间受限**：科学家往往基于直觉选择特定方法，难以穷尽所有可能的算法组合。
- **转化能力弱**：即便有优秀的前沿论文产出，将其思想复现并适配到具体任务中也是巨大的工程负担。

ERA 的出现，正是为了将这种“手动试错”转化为“自动化的全局搜索”。

ERA: LLM + Tree Search for Expert-Level Scientific Software (Nature 2026)

► 核心思想：将科研探索建模为树搜索问题

- **代码变异 (Code Mutation)**: 利用 LLM 作为“变异引擎”。不同于传统遗传算法的字符级随机扰动，LLM 可以在语义层面理解代码逻辑，根据外部输入的“研究建议”（如某篇 Nature 论文的摘要）直接重写算法逻辑。
- **树搜索 (Tree Search)**: 采用类似 AlphaZero 的树搜索策略。系统利用 PUCT 公式平衡“开发 (Exploitation, 优化当前得分最高的方案)”与“探索 (Exploration, 尝试全新的算法分支)”。这保证了 AI 不会陷入局部最优解。
- **思想重组 (Idea Recombination)**: 这是 ERA 最具“创造力”的部分。它会分析多个成功的父代节点，提取两者的核心逻辑，自动融合成一个前所未有的混合模型。

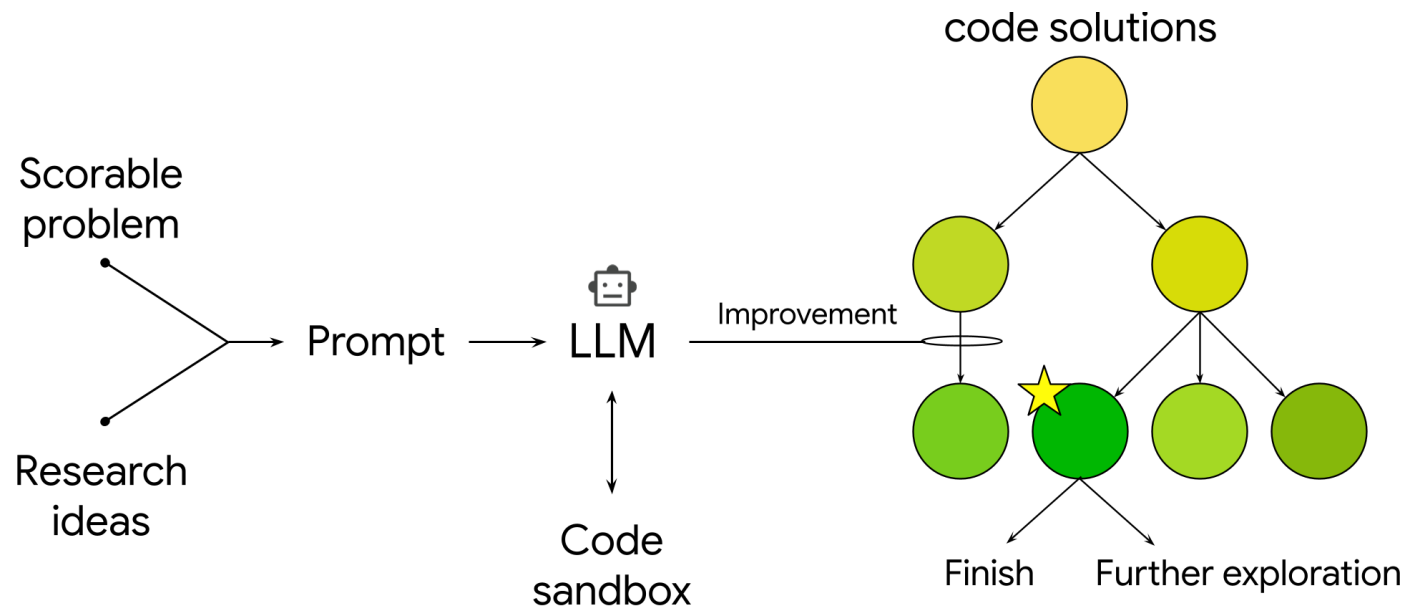


Fig. 1: ERA system architecture

ERA: LLM + Tree Search for Expert-Level Scientific Software (Nature 2026)

1. 基因组学：重塑单细胞单细胞整合排行

在单细胞 RNA 测序 (scRNA-seq) 任务中，目标是消除实验批次效应 (Batch Effects) 同时保留生物学多样性。ERA 基于已有的 9 种基础方法 (如 BBKNN, ComBat 等)，不仅复现了专家代码，还通过“重组”创造了 40 种新方法。

结果：ERA 的方案在 OpenProblems 权威榜单上超越了所有人类开发的 SOTA。特别是 **BBKNN (TS)** 版本，通过将 ComBat 的校准逻辑与 BBKNN 结合，实现了 14% 的性能飞跃。

2. 流行病学：挑战 CDC 集成预测

在对美国 COVID-19 住院数据的预测中，ERA 表现出了极强的鲁棒性。

战果：ERA 产生的 14 个预测策略优于 CDC 的官方 Ensemble 模型。

发现：AI 自动选出的最优解通常是“混搭”：将稳定的传统统计模型 (如 UMass-ar6_pooled) 与灵活的流行病学模型 (如基于再生数 R 的模型) 结合。

3. 时间序列预测：通用库的自动演进

在 GIFT-Eval 测试中，ERA 不仅为每个数据集生成专项代码，还探索出了一个“通用预测库”。

突破点：AI 自动在代码中加入了对节假日的处理逻辑 (Holidays Library) 和分段趋势拟合，这使得代码的泛化能力极强。

ERA: 跨模型性能对比

Model	Method	Batch integration (higher better)	Epidemiology (lower better)
Gemini 2.5 Flash	BoN	0.6306	106.55
	ERA	0.6552	93.07
Mistral Medium	BoN	0.6129	95.73
	ERA	0.6332	87.98
Claude Sonnet 4.6	BoN	0.6502	85.03
	ERA	0.6575	84.56
GPT-5	BoN	0.6740	78.04
	ERA	0.6671	74.55
Gemini 3.1 Pro	BoN	0.6461	92.39
	ERA	0.6641	72.70

Table: ERA vs Best-of-N across multiple LLMs on batch integration and epidemiology tasks

ERA: 跨模型性能对比

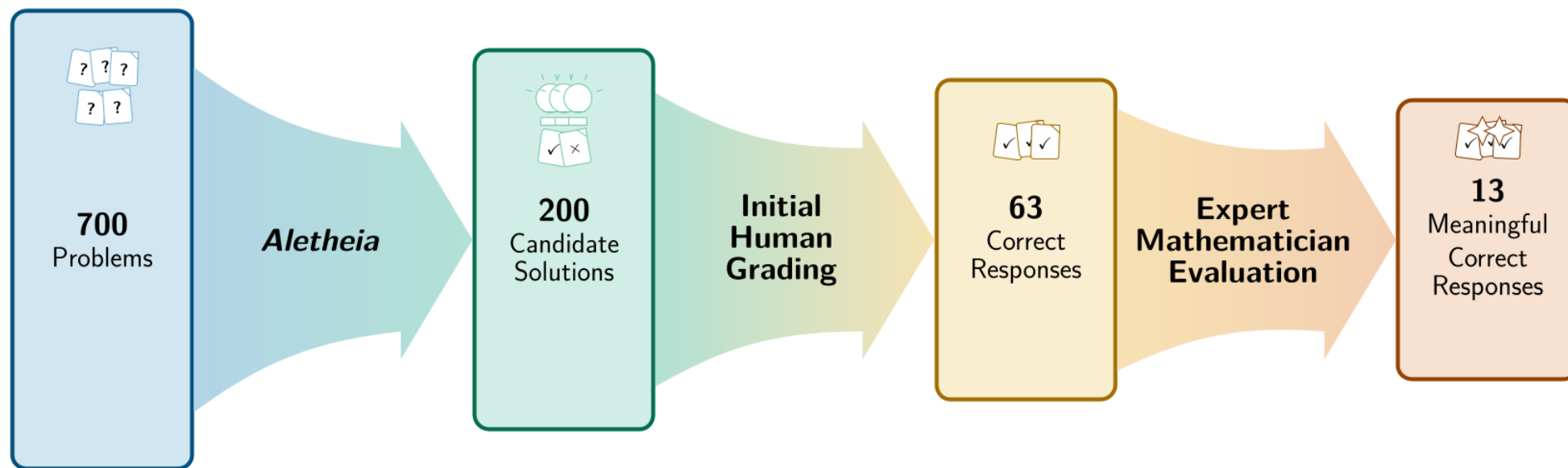
ERA 代表了 AI Agent 在科学领域从“对话式助理”向“工程化助手”的进化。它证明了在具有明确评估标准的任务中，AI 可以极大地降低复杂科学软件的准入门槛。

局限性：尽管 ERA 在经验建模上表现卓越，但它仍然依赖于预定义的“可评分任务（Scorable Task）”。对于需要从零推导理论框架或因果机制的“硬核发现”，ERA 目前主要扮演的是“执行者”和“优化者”的角色。

启示：未来的科学家可能不再需要精通每一行代码，他们的核心价值将转向：**定义高质量的评估指标（Metrics）以及提供精准的领域构想（Ideas）**，而繁琐的实现与调优过程将交给像 ERA 这样的系统去自动完成。

Gemini 挺进数学荒原：Erdős 猜想的半自主发现之路

保罗·埃尔多斯 (Paul Erdős) 留下了海量的猜想，Google DeepMind 团队利用定制的 Gemini Deep Think 智能体 (代号 Aletheia)，Gemini 负责证明、Aletheia 负责验证，对 700 个开放问题发起冲锋。



1.自然语言验证器 (NL Verifier): 在 Gemini 生成初稿后，由特定的验证模块通过逻辑推演排除明显的错误。这让专家面对的候选方案从 700 个锐减至 212 个。

2.领域专家垂直评审: 人类专家介入，通过对比文献和跨学科沟通，剔除那些“理解错题意”的方案。

Gemini 挺进数学荒原：Erdős 猜想的半自主发现之路

在众多成果中，Erdős-1051 最具代表性。问题：如果一个整数序列 (a_n) 增长极快（满足 $\liminf a_n^{1/2^n} > 1$ ），那么级数 $\sum \frac{1}{a_n a_{n+1}}$ 是否一定为无理数？

Aletheia 给出了一个漂亮的证明：

- **逻辑直觉**：通过构造 partial products 和 tails，利用 Mahler 准则证明了该级数无法表示为分母固定的有理数
- **成果转化**：该证明随后由人类数学家使用 Lean 4 进行了形式化验证，并衍生出了一篇全新的学术论文

PROVED (LEAN)

Is it true that if $a_1 < a_2 < \dots$ is a sequence of integers with

$$\liminf a_n^{1/2^n} > 1$$

then

$$\sum_{n=1}^{\infty} \frac{1}{a_n a_{n+1}}$$

is irrational?

#1051: [ErGr80,p.64][Er88c,p.106] irrationality

In [Er88c] Erdős notes this is true if $a_n \rightarrow \infty$ 'rapidly'. In [ErGr80] they further ask 'what the strongest theorem of this type' would be.

This was solved in the affirmative by Aletheia. This was extended by Barreto, Kang, Kim, Kovač, and Zhang [BKZZ26], who essentially give a complete answer: if $\phi = \frac{1+\sqrt{5}}{2}$ is the golden ratio and $1 \leq a_1 < a_2 < \dots$ is a monotonically increasing sequence of integers such that

$$\limsup a_n^{1/\phi^n} = \infty$$

Gemini 挺进数学荒原：Erdős 猜想的半自主发现之路

接下来这个「Erdős-75号灵异事件」，暴露出AI的智商硬伤。

这道题在数学圈臭名昭著，因为它被「诅咒」了。

Erdős-75 [EHS82, Erd95b, Erd95a]

Is there a graph of chromatic number \aleph_1 such that for all $\epsilon > 0$ if n is sufficiently large and H is a subgraph on n vertices then H contains an independent set of size $> n^{1-\epsilon}$?

1995年，埃尔德什在写下这个猜想时，犯了一个低级逻辑错误，这道题题干就是错的，是个伪命题。

魔幻的一幕发生了：Aletheia接手后，不仅没发现题目有问题，反而凭借其狂暴的算力和自我博弈机制，硬生生地输出了一份长达几十页、逻辑闭环的「完美证明」。

Gemini 挺进数学荒原：Erdős 猜想的半自主发现之路

Category	Count	Percentage
Fundamentally Flawed	137	68.5%
Technically Correct	63	31.5%
<i>Meaningfully Correct (subset)</i>	<i>13</i>	<i>6.5%</i>
Total Candidates	200	100.0%

但DeepMind自己也承认，700题里面，智能体过滤了500题，剩下200题中的的68.5%都是学术垃圾。

700个问题里只解答出13个，转化率还不到2%。

Gemini 挺进数学荒原：Erdős 猜想的半自主发现之路

提出了一个极具警示性的概念：**潜意识抄袭 (Subconscious Plagiarism)**

在解决 Erdős-1089 等问题时，AI 生成的证明与 1981 年的一篇东欧冷门数学期刊上的论文高度重合。由于该文献在 AI 的预训练数据集中，即使 AI 在推理日志中没有表现出搜索该行为，它也可能在生成过程中“无意识”地复现了训练数据。

- **风险**：如果人类研究者直接署名发布此类成果，将面临严重的学术诚信挑战。
- **反思**：数学论文的署名权应始终属于人类，因为人类需要承担起“核实文献出处”这一法律与学术责任。

Erdős-1089 [Erd75]

Let $g_d(n)$ be minimal such that every collection of $g_d(n)$ points in \mathbb{R}^d determines at least n many distinct distances. Estimate $g_d(n)$. In particular, does

$$\lim_{d \rightarrow \infty} \frac{g_d(n)}{d^{n-1}}$$

exist?

- **现状**：AI 已经能熟练处理“学生练习题级别”的数学挑战，并能高效辅助专家缩小搜索空间
- **局限性**：AI 在理解 Erdős 独特的命名约定和识别隐蔽文献方面仍显笨拙
- **未来**：数学发现的未来不在于 AI 的全自动化，而在于如何利用 AI 的超强联想力来识别那些被历史尘封的连接点

4

学术论文介绍 II: Agent 能力评价与基准测试

Fudan NLP Group Research Papers

CL-BENCH: A Benchmark for Context Learning



评测基准分类学依据

将复杂上下文解决任务拆解为两个维度：

Context维度

+

Problem维度

→



Context中知识存在的类型
衡量模型能否正确调用知识

如何利用Context中的知识
衡量模型能否解决该类问题

Context分类

规则型
程序型
案例型
概念型

Problem分类

定位记忆
理解解释
应用执行
分析推理
创造发现

难度递进



CL-BENCH: A Benchmark for Context Learning

Domain Knowledge Reasoning



- Finance
- Healthcare
- Humanities,
- Legal Advisory
- Lifestyle
- Management
- Science

Rule System Application



- Game Mechanics
- Mathematical Formalism
- Programming Syntax
- Legal & Regulatory
- Technical Standards

Procedural Task Execution

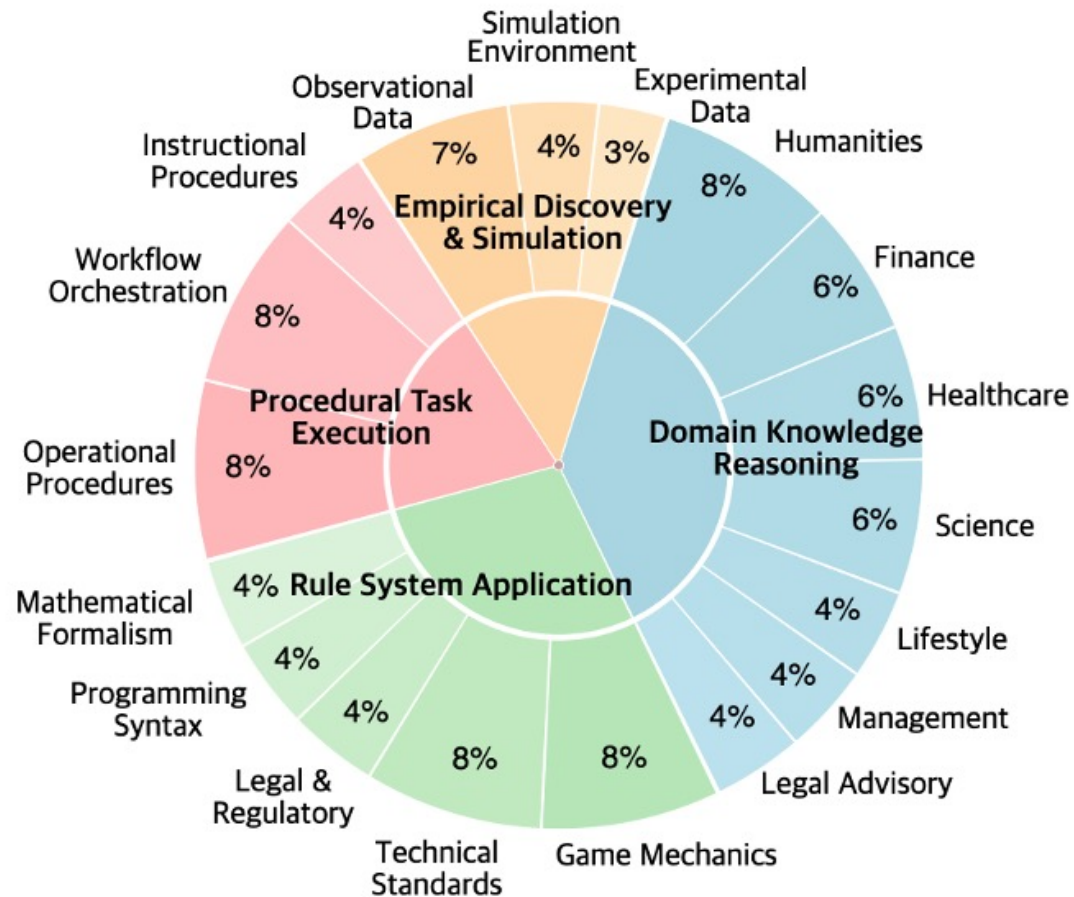


- Instructional Procedures
- Operational Procedures
- Workflow Orchestration

Empirical Discovery & Simulation



- Experimental Data
- Observational Data
- Simulation Environment



500个复杂上下文场景、1899个任务、3.16万项验证标准

解决每个任务要求模型必须从上下文中学习到模型**预训练中不存在的新知识**，并正确使用

CL-BENCH: 主实验结果 — 前沿模型平均仅 17.2%

Model Names	Overall (%)	Domain Knowledge Reasoning (%)	Rule System Application (%)	Procedural Task Execution (%)	Empirical Discovery & Simulation (%)
GPT 5.1 (High)	23.7 ± 0.5	25.3 ± 1.3	23.7 ± 1.3	23.8 ± 1.4	18.1 ± 3.1
Claude Opus 4.5 Thinking	21.1 ± 1.4	23.7 ± 1.2	19.0 ± 1.5	22.6 ± 1.5	15.1 ± 2.3
GPT 5.2 (High)	18.1 ± 0.8	18.6 ± 0.9	17.2 ± 1.3	21.4 ± 1.1	11.7 ± 1.8
o3 (High)	17.8 ± 0.2	18.0 ± 1.4	17.6 ± 1.1	19.5 ± 0.4	13.7 ± 0.8
Kimi K2 Thinking	17.6 ± 0.6	18.7 ± 0.6	17.0 ± 1.5	18.8 ± 0.7	12.6 ± 4.0
HY 2.0 Thinking	17.2 ± 0.6	18.0 ± 1.0	17.3 ± 0.5	19.4 ± 1.1	8.9 ± 0.3
Gemini 3 Pro (High)	15.8 ± 0.3	15.5 ± 1.1	17.7 ± 1.7	16.4 ± 1.6	10.1 ± 3.1
Qwen 3 Max Thinking	14.1 ± 0.1	13.5 ± 0.5	15.6 ± 1.0	15.2 ± 1.4	9.0 ± 1.0
Doubao 1.6 Thinking	13.4 ± 0.1	13.7 ± 0.1	14.2 ± 1.4	13.9 ± 1.5	9.4 ± 0.3
DeepSeek V3.2 Thinking	13.2 ± 0.4	13.6 ± 0.6	13.8 ± 0.6	14.2 ± 0.1	8.0 ± 1.5

Table 2: Task Solving Rate (%) across context categories for 10 frontier LMs

模型在 CL-bench 上的任务解决率 所有模型均在推理模式下进行评估, 结果报告为三次运行的平均值 ± 标准差 (%)。

CCTU: Benchmarking Tool Use under Complex Constraints

Junjie Ye, Guoqiang Zhang, Wenjie Fu, Tao Gui, Qi Zhang, Xuanjing Huang

Fudan University | 复杂约束下的工具使用基准测试

► 动机: LLM 工具使用能力缺乏系统性评价

► 12 类约束分类体系, 4 个维度:

- 资源约束 (Resource)
- 行为约束 (Behavior)
- 工具集约束 (Toolset)
- 响应约束 (Response)

► 200 个精心构造的测试用例

- 平均 7 种约束类型 / 平均 4,700+ tokens

► 核心发现: 所有模型任务完成率均低于 20%

- 约束违反率超过 50%

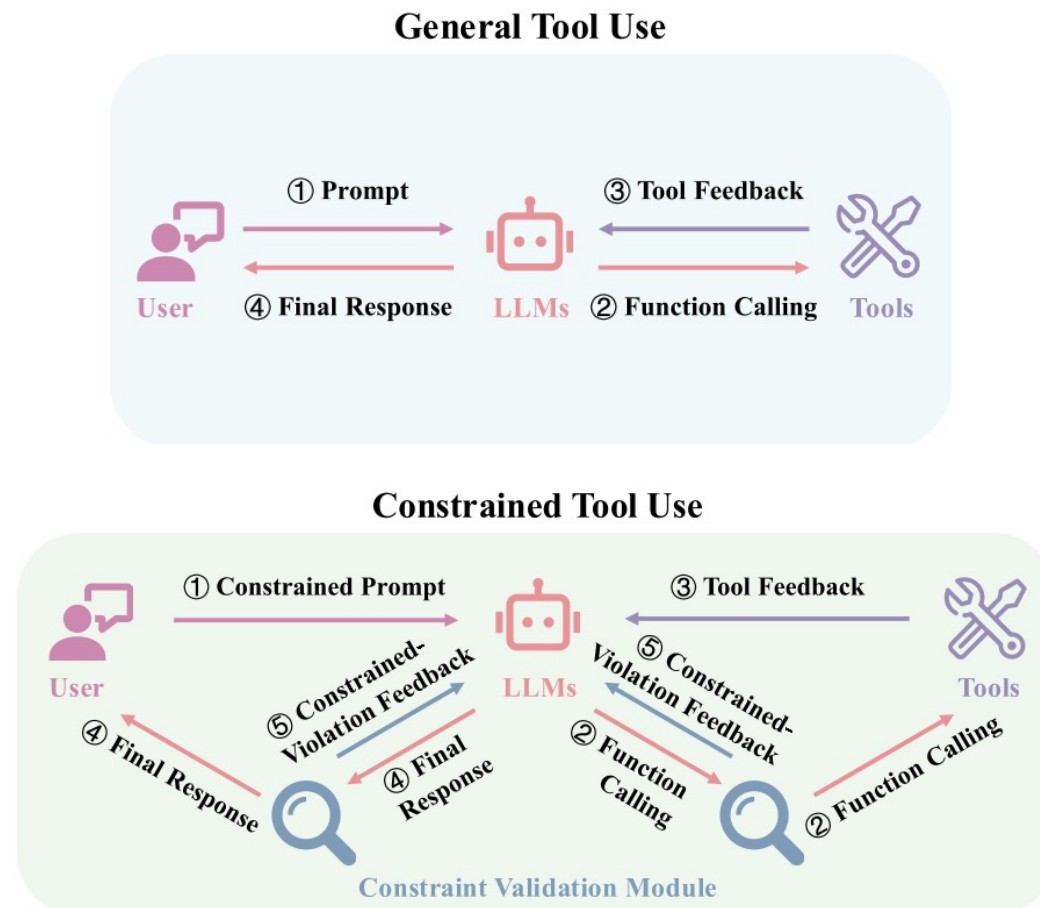


Fig. 1: General vs. Constrained Tool Use

CCTU: 主实验结果 — 所有模型 PSR < 20%

Models	Single-Hop		Parallel Single-Hop		Multi-Hop		Parallel Multi-Hop		Overall	
	SR	PSR	SR	PSR	SR	PSR	SR	PSR	SR	PSR
<i>Thinking Mode</i>										
GPT-5.2	32.67 _{1.89}	24.67 _{3.40}	24.67 _{1.89}	17.33 _{0.94}	25.33 _{2.49}	20.67 _{2.49}	15.33 _{0.94}	10.00 _{1.63}	24.50 _{0.82}	18.17 _{1.03}
GPT-5.1	25.33 _{5.25}	20.00 _{7.12}	20.67 _{2.49}	16.00 _{0.00}	22.67 _{1.89}	20.67 _{1.89}	22.67 _{0.94}	9.33 _{2.49}	22.83 _{2.01}	16.50 _{2.83}
Claude Opus 4.6	34.67 _{4.99}	10.00 _{1.63}	30.67 _{2.49}	13.33 _{0.94}	38.67 _{3.40}	23.33 _{1.89}	32.67 _{1.89}	12.67 _{0.94}	34.17 _{2.25}	14.83 _{1.03}
Seed-2.0-Pro	22.67 _{3.77}	19.33 _{4.99}	20.67 _{5.25}	12.67 _{3.40}	22.67 _{0.94}	18.67 _{0.94}	15.33 _{1.89}	8.67 _{2.49}	20.33 _{2.62}	14.83 _{2.87}
Qwen3.5-Plus	20.67 _{4.99}	5.33 _{2.49}	23.33 _{0.94}	8.00 _{1.63}	32.00 _{2.83}	21.33 _{0.94}	23.33 _{1.89}	8.00 _{1.63}	24.83 _{1.03}	10.67 _{1.55}
Gemini 3 Pro	23.33 _{2.49}	12.00 _{3.27}	28.00 _{1.63}	16.00 _{1.63}	14.67 _{0.94}	11.33 _{2.49}	11.33 _{3.40}	2.67 _{0.94}	19.33 _{1.84}	10.50 _{1.08}
DeepSeek-V3.2	15.33 _{1.89}	6.67 _{2.49}	22.67 _{1.89}	12.00 _{1.63}	26.00 _{4.32}	16.67 _{3.40}	8.00 _{1.63}	0.67 _{0.94}	18.00 _{1.08}	9.00 _{0.41}
OpenAI o3	22.67 _{1.89}	17.33 _{1.89}	7.33 _{0.94}	4.00 _{1.63}	13.33 _{0.94}	10.00 _{1.63}	4.00 _{1.63}	1.33 _{0.94}	11.83 _{1.31}	8.17 _{1.03}
Kimi K2.5	22.67 _{3.77}	4.67 _{2.49}	26.00 _{3.27}	10.67 _{3.77}	20.00 _{2.83}	10.67 _{2.49}	16.67 _{4.11}	4.67 _{2.49}	21.33 _{2.05}	7.67 _{1.65}
<i>Non-Thinking Mode</i>										
GPT-5.2	28.00 _{3.27}	24.00 _{1.63}	19.33 _{0.94}	15.33 _{2.49}	17.33 _{4.11}	14.00 _{4.90}	16.67 _{0.94}	10.67 _{0.94}	20.33 _{0.62}	16.00 _{1.87}
Claude Opus 4.6	38.00 _{1.63}	12.00 _{0.00}	29.33 _{0.94}	13.33 _{0.94}	38.00 _{2.83}	23.33 _{1.89}	32.67 _{0.94}	13.33 _{0.94}	34.50 _{1.08}	15.50 _{0.41}
GPT-5.1	22.67 _{2.49}	19.33 _{0.94}	19.33 _{2.49}	16.67 _{1.89}	16.67 _{2.49}	14.00 _{1.63}	14.00 _{0.00}	6.67 _{3.40}	18.17 _{0.85}	14.17 _{0.62}
Kimi K2.5	19.33 _{3.40}	6.67 _{0.94}	29.33 _{0.94}	14.00 _{0.00}	25.33 _{1.89}	15.33 _{0.94}	16.67 _{0.94}	6.67 _{0.94}	22.67 _{0.85}	10.67 _{0.62}
Gemini 3 Pro	22.67 _{6.60}	12.67 _{4.71}	26.67 _{2.49}	14.67 _{0.94}	16.00 _{3.27}	11.33 _{0.94}	10.67 _{1.89}	2.00 _{1.63}	19.00 _{1.47}	10.17 _{0.85}
Seed-2.0-Pro	20.00 _{2.83}	13.33 _{3.77}	20.00 _{2.83}	10.00 _{1.63}	20.00 _{5.89}	13.33 _{4.99}	12.67 _{0.94}	3.33 _{0.94}	18.17 _{2.87}	10.00 _{1.87}
OpenAI o3	24.00 _{4.90}	18.67 _{2.49}	9.33 _{1.89}	4.67 _{1.89}	10.67 _{2.49}	7.33 _{1.89}	2.00 _{1.63}	1.33 _{0.94}	11.50 _{0.82}	8.00 _{1.08}
Qwen3.5-Plus	20.67 _{2.49}	4.00 _{0.00}	20.00 _{1.63}	6.67 _{2.49}	28.67 _{2.49}	14.67 _{0.94}	16.00 _{1.63}	2.67 _{2.49}	21.33 _{0.85}	7.00 _{1.41}
DeepSeek-V3.2	20.00 _{3.27}	6.67 _{0.94}	17.33 _{0.94}	6.00 _{0.00}	20.67 _{7.54}	12.00 _{3.27}	10.00 _{1.63}	1.33 _{0.94}	17.00 _{2.55}	6.50 _{0.82}

Table 2: Performance of LLMs on CCTU (Solve Rate & Perfect Solve Rate)


SciAgentGym: Multi-Step Scientific Tool-use Benchmark


传统的 LLM 评估（如 GPQA）更像是闭卷考试，考察模型存储了多少科学知识。但在真正的科研中，科学家需要操作分子模拟、查询数据库、编写数据分析代码。


此前工作的局限性在于：


- 1. 缺乏交互反馈：** 模型给出错误指令后无法根据报错修正。
- 2. 工具链逻辑缺失：** 模型知道有哪些工具，却不知道工具之间的 A 依赖 B 的拓扑关系。
- 3. 长程迷失：** 随着推理步数增加，模型极易陷入死循环（Loop）或因一次误操作导致全局溃败。

? **Task:** Identify optimal rod-shaped ligand for binding pocket.
Sub-question: Minimize score $S=(\text{Asphericity} \times 100) + (\text{Eccentricity} \times 10)$.
Candidates: Anthracene, Phenanthrene, Acenaphthylene, Fluorene.

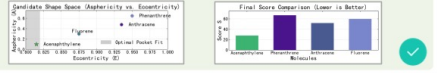
Agent.step 


Action: Chemical Name to SMILES
Tool: Chemical Name to Molecular(['Anthra..])
... **Environment Feedback:** 


Action: Error Encountered-Missing Parameter
Tool: Get 3d Properties({'smiles': 'C1=C..}). 

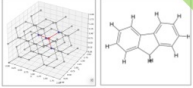
Action: Error Recovery
Tool: Get 3d Properties({'method': 'ETKD..}). 


...
Action: Optimal Candidate Selection
Tool: Visualize Candidate Comparison({'smiles'..}).







... 

Output: Optimal Ligand is Acenaphthylene.
Molecular structure shown as below.
Minimize score $S = 25.94 + 8.96 = 34.9$. 



Environment 

Setup:

-  Dataset
-  Toolkit:53
-  Files:1
-  Code

Execution log:

- Total Steps: 12
- Total Tool Calls: 20
- DB Retrieval...
- Pictures saving...


read 

Figure 1: 代表性科学工具调用交互轨迹

SciAgentGym，这是一个专为 LLM Agent 设计的科学实验室环境。它不考模型“背书”，而是考模型“实操”。

SciAgentGym: Multi-Step Scientific Tool-use Benchmark

- ▶ 1,780 个领域专用工具，覆盖 4 个自然科学学科
 - Physics / Chemistry / Biology / Materials Science

- ▶ SciAgentBench: 259 任务，1,134 子问题
 - 三级难度: L1 (≤ 3 steps) / L2 (4-7) / L3 (≥ 8)

▶ 评估指标

- 成功率 (SR) : 所有子问题回答正确
- SPL = 成功率 \times 路径效率
- (对比专家验证的参考路径长度)

▶ 核心发现:

- GPT-5 成功率从 60.6% (L1) 降至 30.9% (L3)
- 长周期科学工具使用是关键瓶颈

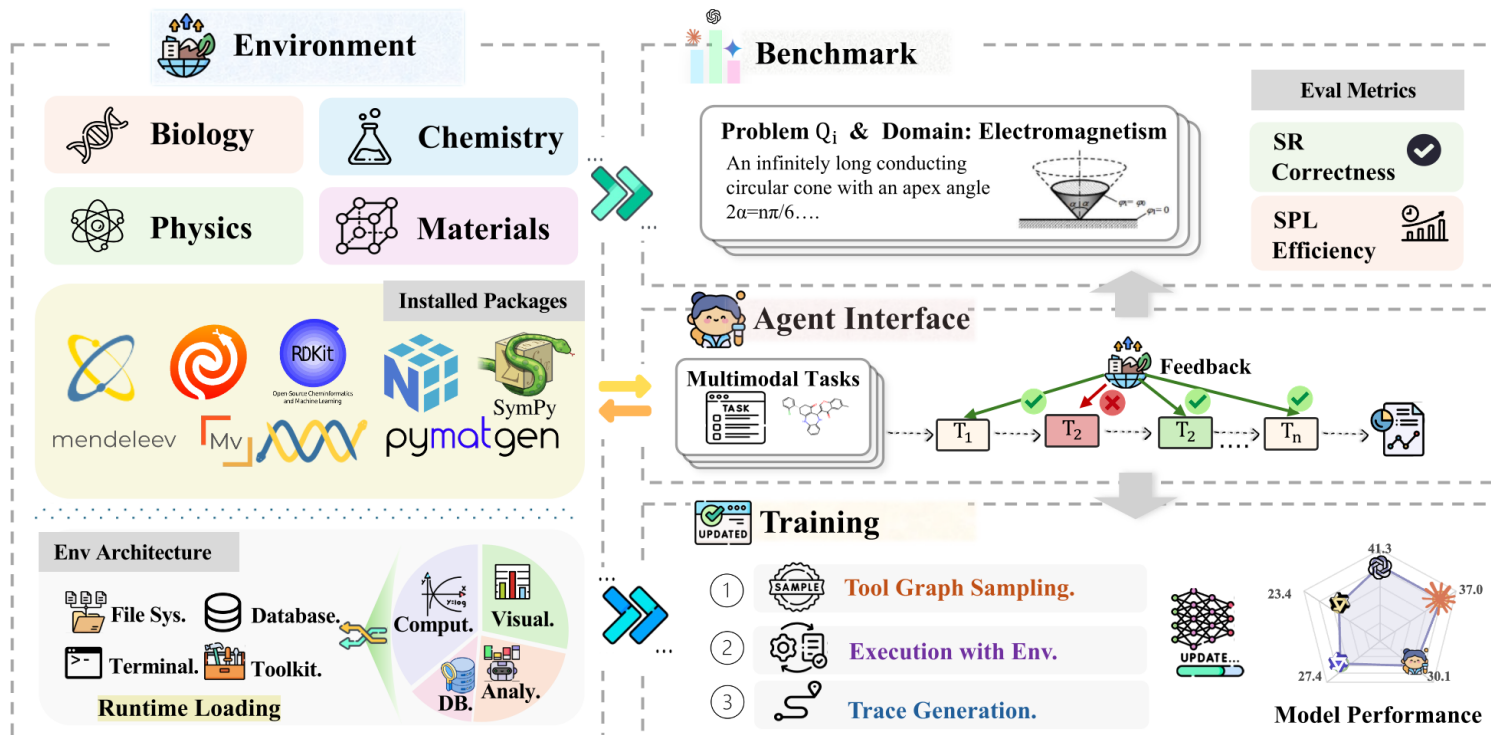


Fig. 2: SciAgentGym Overview

SciAgentGym: Multi-Step Scientific Tool-use Benchmark

核心发现

1 长周期任务仍是关键瓶颈

GPT-5: 60.6%(L1) → 30.9%(L3)

平均下降 68%

2 工具增强不可或缺

Claude-Sonnet-4: +13.5%

SciAgent-8B: +6.8%

3 小模型可超越大模型

SciAgent-8B 30.1%

> Qwen3-235B-Inst 23.9%

4 学科间差异显著

生命科学工具依赖最强 +8.4%

材料科学整体最具挑战性

Model	Overall				By Subject (w/ Tools)				By Difficulty (w/ Tools)		
	w/o Tools	w/ Tools	Δ	SPL	Phys.	Chem.	Mat.	Life	L1	L2	L3
Closed-Source Models											
GPT-5	32.3	41.3	+9.0	0.24	46.3	43.8	28.6	32.3	60.6	38.6	30.9
Grok-4-1	30.4	40.3	+9.9	0.25	47.2	38.2	32.4	30.0	52.8	43.2	27.6
Claude-Sonnet-4	22.4	35.9	+13.5	0.19	39.4	39.5	27.0	25.0	55.6	36.4	16.9
Gemini-2.5-Flash	28.5	32.7	+4.2	0.21	38.3	32.4	28.6	17.2	49.5	32.7	22.0
Gemini-2.5-Pro	24.8	32.6	+7.8	0.21	37.3	35.1	26.5	18.8	56.5	33.4	16.1
O3	26.6	32.0	+5.4	0.26	35.5	37.3	32.4	6.5	50.4	31.1	20.2
O4-mini	27.8	31.1	+3.3	0.24	31.2	35.5	30.6	20.0	53.3	35.2	9.7
Gemini-2.5-Pro-Think	28.9	28.8	-0.1	0.19	33.3	28.9	21.2	21.9	51.7	28.9	11.7
GPT-4o	17.1	18.7	+1.6	0.14	21.3	20.5	8.6	16.0	27.3	16.4	7.5
Open-Source Large Models (>30B)											
GLM-4.6V	26.0	30.9	+4.9	0.25	30.9	37.5	22.2	18.8	44.4	31.4	21.6
Qwen3-VL-235B-Think	24.4	28.0	+3.6	0.16	30.6	29.5	22.9	22.6	53.9	26.9	15.1
Qwen3-VL-235B-Inst	23.0	23.9	+0.9	0.16	28.1	26.5	5.0	17.2	50.0	21.8	2.4
Qwen3-VL-32B-Think	24.4	27.9	+3.5	0.17	33.0	31.2	8.8	22.6	49.1	26.2	14.6
Qwen3-VL-32B-Inst	22.8	27.4	+4.6	0.15	31.8	29.3	20.0	16.1	47.1	25.4	15.1
Open-Source Small & Medium Models (≤30B)											
Qwen3-VL-8B-Inst	18.4	23.4	+5.0	0.09	24.0	28.6	7.1	24.1	38.7	20.9	8.2
SciAgent-8B	23.3 ^{+4.9}	30.1 ^{+6.7}	+6.8	0.16	33.0 ^{+9.0}	35.2 ^{+6.6}	9.1 ^{+2.0}	31.0 ^{+6.9}	41.9 ^{+3.2}	27.5 ^{+6.6}	19.6 ^{+11.1}
Qwen3-VL-4B-Inst	17.0	19.7	+2.7	0.10	23.8	20.6	10.3	13.3	40.7	15.2	10.0
SciAgent-4B	17.4 ^{+0.4}	25.2 ^{+5.5}	+7.8	0.13	28.4 ^{+4.6}	28.4 ^{+7.8}	14.7 ^{+4.4}	19.4 ^{+6.1}	43.3 ^{+2.6}	27.5 ^{+12.3}	13.5 ^{+3.5}
Pixtral-12B	7.8	7.2	-0.6	0.07	7.5	6.3	5.9	10.0	12.1	2.9	3.7
Average	23.2	28.1	+4.9	0.17	31.7	30.4	18.9	20.2	46.4	27.2	14.7

Table 3: SciAgentBench 完整评测结果 (SR%)

SciAgentGym: Multi-Step Scientific Tool-use Benchmark

SciForge 四步流程

① 构建工具依赖图

1,780 个工具间基于类型兼容性建边

② ϵ -贪心阶段感知采样

兼顾 workflow 逻辑顺序与复杂依赖探索

③ 前向执行验证

生成黄金轨迹 + 错误恢复增强轨迹

④ 轨迹→问题生成

语义抽象隐藏中间结果，确保非平凡性

消融实验核心结论

- ▶ 通用工具数据 → 负迁移 (-4.6%)
- ▶ 错误恢复轨迹对鲁棒性至关重要
- ▶ 工具使用能力比静态知识更易随数据扩展
- ▶ 科学领域间正向跨域迁移已得到验证

Model	Phys.	Chem.	Life	Mat.	Avg.
Qwen3-VL-8B	24.0	28.6	24.1	7.1	23.0
<i>Ablation: Training Data Composition</i>					
Qwen3-VL-8B-OtherTools	21.1 -2.9	21.0 -7.6	20.1 -4.0	3.7 -3.4	18.5 -4.6
Qwen3-VL-8B-NoError	30.5 +6.5	26.4 -2.2	26.7 +2.6	14.7 +7.6	26.5 +3.5
<i>Ablation: Domain Transfer</i>					
Qwen3-VL-8B-Physics	30.5 +6.5	31.5 +2.9	25.0 +0.9	17.1 +10.0	28.2 +5.2
Qwen3-VL-8B-Chem	24.8 +0.8	35.6 +7.0	31.2 +7.1	20.0 +12.9	28.3 +5.3
Qwen3-VL-8B-Merged	33.0 +9.0	35.2 +6.6	31.0 +6.9	9.1 +2.0	30.0 +7.0

Table 4: 训练数据组成与跨域迁移消融实验

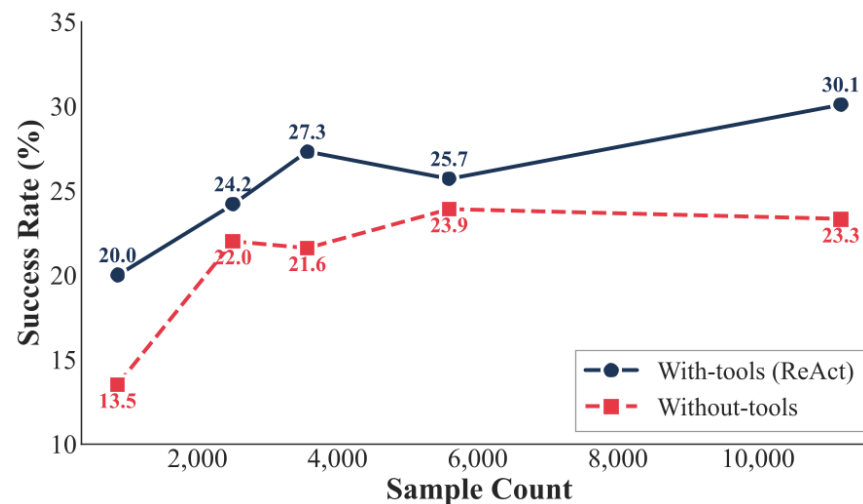


Figure 5: 训练数据量与成功率的扩展关系

SciAgentGym: 工具使用环境对比

Works	Domain	MM	Env	DSM	Traj
<i>Interactive Environments</i>					
AgentBench [11]	Multi-Env	✗	✓	✗	✗
AgentGym [24]	Multi-Env	✗	✓	✗	✓
DiscoveryWorld [9]	Sim Science	✗	✓	✗	✗
MedAgentGym [28]	Biomedical	✗	✓	✗	✓
<i>Tool-Use</i>					
ToolBench (ToolLLM) [16]	General APIs	✗	✗	✓	✓
SciAgent [13]	Sci Reasoning	✗	✗	✓	✓
BFCL [15]	Func Calling	✗	✗	✗	✗
τ -Bench [31]	Retail/Airline	✗	✓	✓	✗
SciAgentGym (Ours)	Multi-Science	✓	✓	✓	✓

Table 1: Comparison of interactive environments and tool-use benchmarks

MM-Doc-R1: Multi-turn RL for Long Document Visual QA

! 核心挑战

传统 RAG 系统依赖单次检索，无法处理长文档中需要跨页面迭代信息收集的复杂多跳查询 (Multi-hop Questions)。

📄 现有方法局限

- BM25 / BGE-M3: 仅初始查询检索，无迭代能力
- ColPALI / ColQwen: 视觉嵌入但缺乏逐步推理
- MDocAgent / M3DocRAG: 主要针对单跳问题

💡 我们的方案

设计视觉感知智能体 workflow + 多轮 RL (SPO) 训练迭代信息发现能力

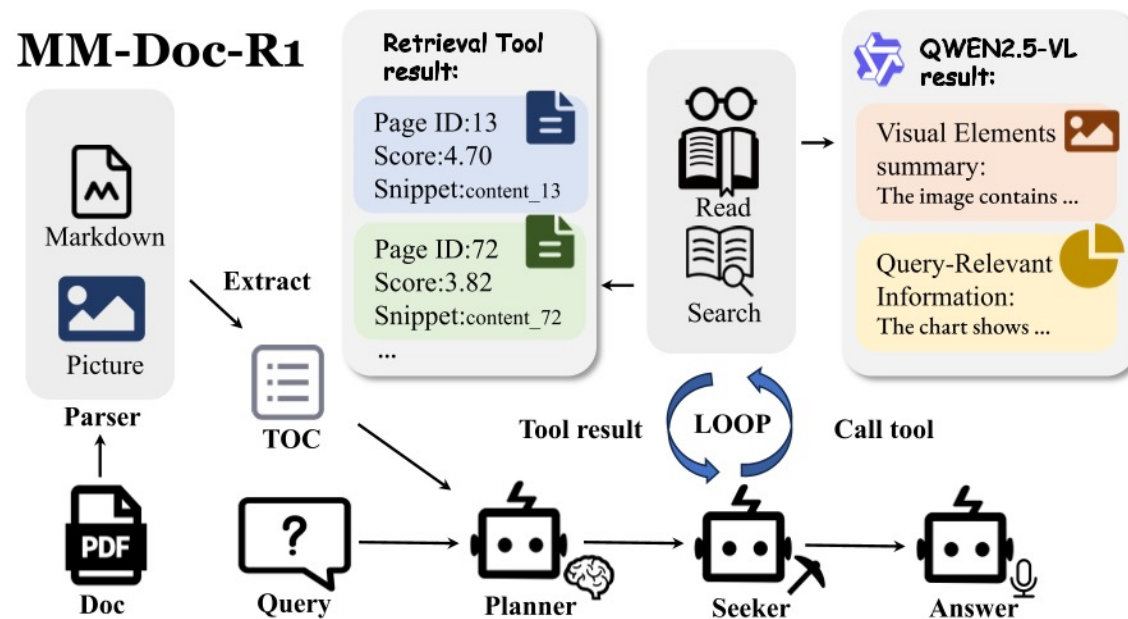
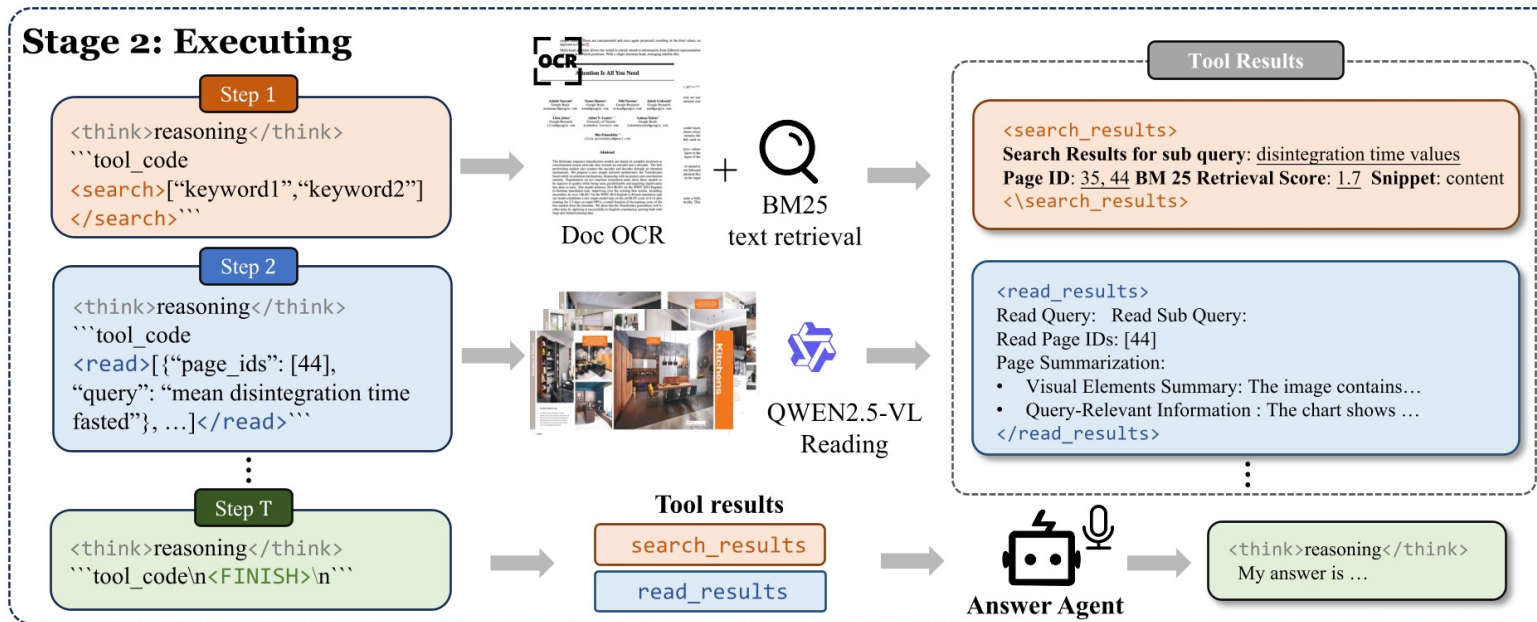
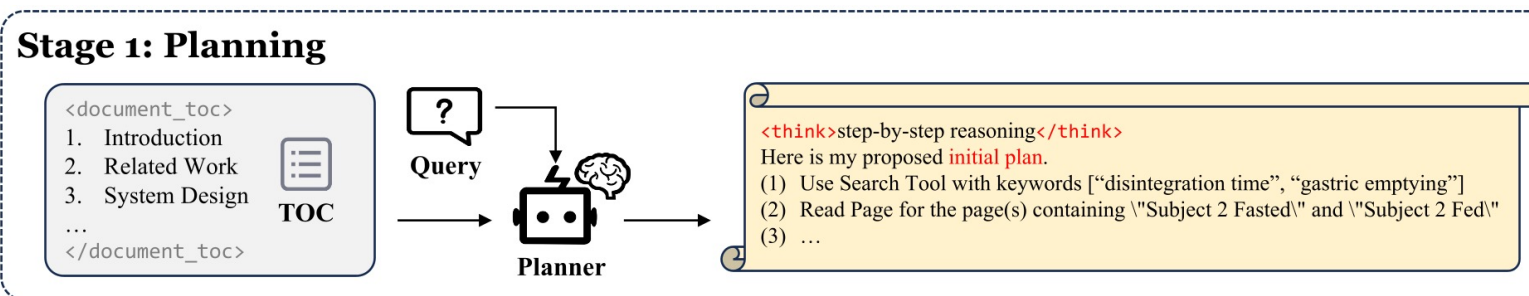


Figure 1: MM-Doc-R1 系统概览 — 迭代检索 + VLM 阅读

MM-Doc-R1: Multi-turn RL for Long Document Visual QA



MM-Doc-R1: SPO vs GRPO 优势估计对比

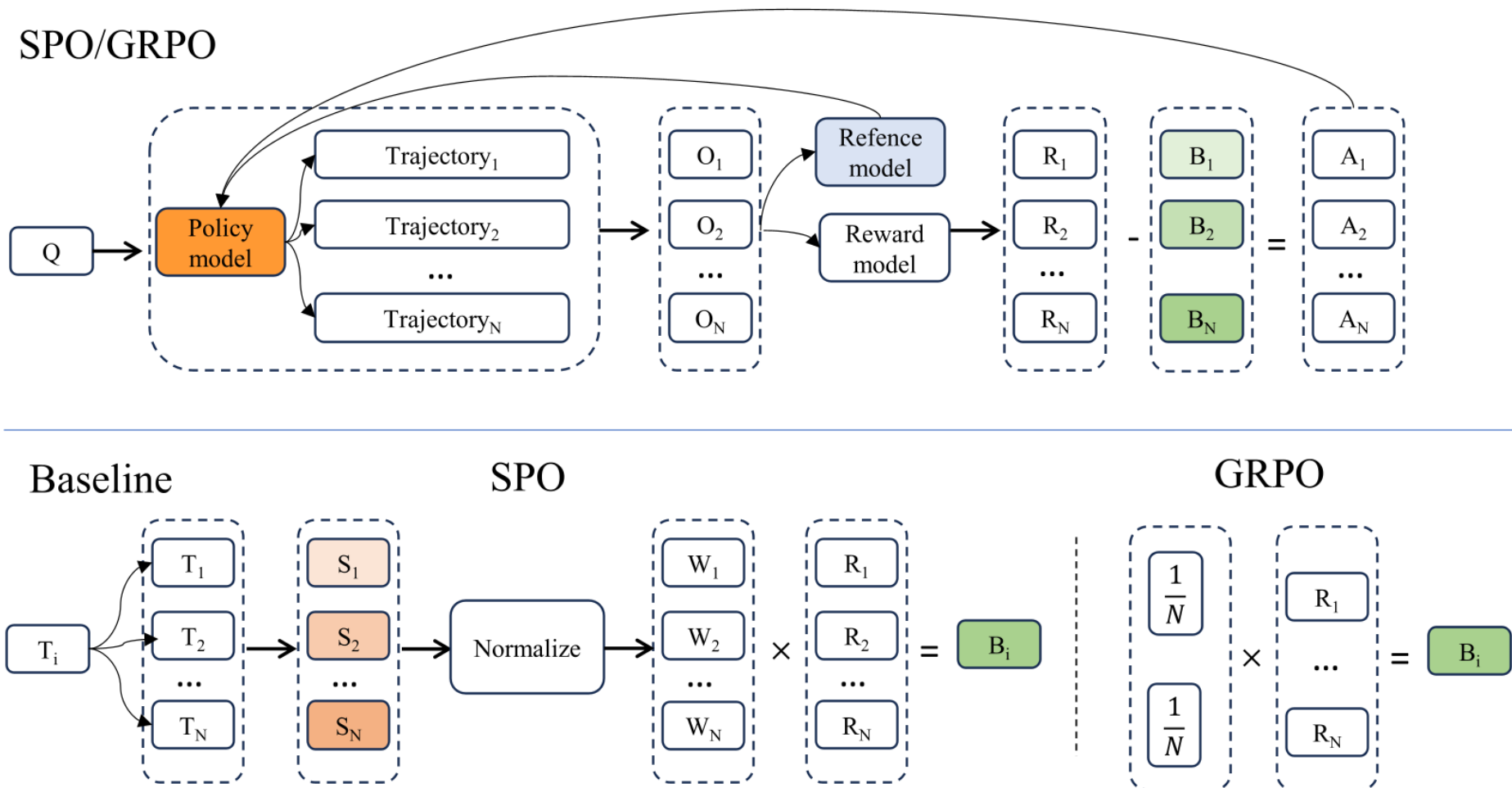


Fig. 3: SPO and GRPO advantage estimation comparison

SPO: 语义越相似的轨迹 \rightarrow 中间状态重叠越大 \rightarrow 共享基线估计越准确

利用 BGE-M3 嵌入计算轨迹间余弦相似度作为动态权重

MM-Doc-R1: SPO vs GRPO 优势估计对比

Method	Evidence Modality					Evidence Count			Overall	
	Text	Layout	Chart	Table	Figure	Single	Multi	Unans.	ACC	F1
<i>Human Performance</i>										
Human	—	—	—	—	—	—	—	—	65.8	66.0
<i>Upper Bounds (Ground-Truth Evidence)</i>										
Qwen2.5-VL-7B	33.8	38.7	31.8	32.3	34.1	46.6	20.5	92.8	46.8	41.7
Qwen3-8B	44.3	37.7	25.7	59.3	22.8	42.7	35.7	89.2	49.6	46.9
<i>RAG Baselines</i>										
BM25	30.9	23.4	22.3	28.5	9.2	30.7	14.1	88.3	36.4	31.0
BGE-M3	32.0	20.8	21.7	40.3	14.7	35.4	18.5	84.3	39.3	34.8
Colqwen	27.8	25.0	16.5	22.4	23.7	33.9	13.7	82.5	36.5	31.2
Mdoc agent	33.1	29.3	25.8	32.6	30.0	43.7	18.4	43.4	35.0	33.3
M3doc RAG	39.2	26.7	29.8	39.0	32.0	50.3	21.2	40.7	38.4	36.7
<i>Ours: MM-Doc-R1</i>										
Qwen3-4B	28.9	23.8	23.1	35.3	22.4	37.1	18.6	72.2	37.7	32.2
+GRPO	36.3	35.2	29.5	40.2	27.1	44.5	22.5	58.7	39.9	36.3
+SPO	41.1	37.2	35.6	47.2	30.5	50.5	27.5	68.0	46.0	41.2
Qwen3-8B	39.6	37.6	37.8	45.6	27.3	47.3	27.5	73.9	45.7	41.5
+GRPO	40.9	36.8	35.7	49.9	28.3	48.5	30.2	60.5	44.7	41.9
+SPO	46.2	38.1	40.8	52.8	35.9	56.0	31.2	68.2	49.7	46.1

Table 1: MMLongBench-Doc 基准评测 — 各方法在不同证据模态和数量维度上的性能对比

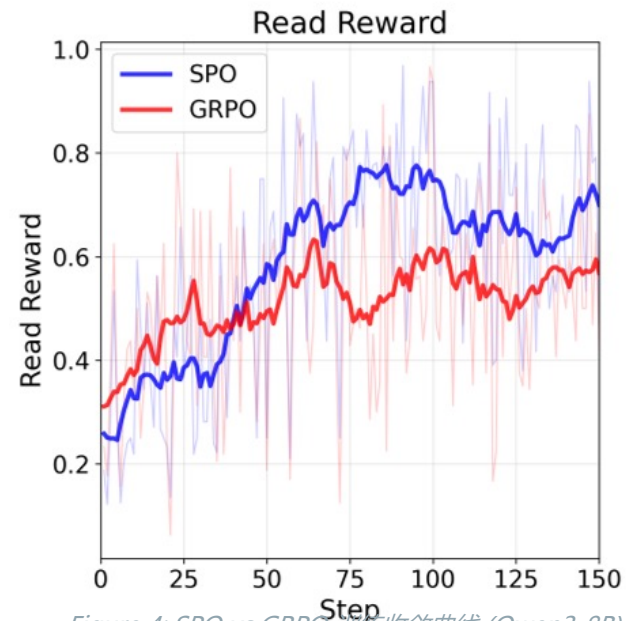


Figure 4: SPO vs GRPO 训练收敛曲线 (Qwen3-8B)

🏆 核心发现

- ▶ 未训练即超越最佳 RAG 基线 +10.4%
- ▶ SPO + Qwen3-4B: 比 GRPO +6.1%
- ▶ SPO + Qwen3-8B: 比 GRPO +5.0%
- ▶ 多证据多跳问题增益最显著 +10.0%
- ▶ SPO 训练收敛更快、曲线更平滑

5

AI Scientist - WisPaper

Introduction of WisPaper Platform

从文献检索说起



深度搜索

🔍 查找近三年关于耐药机制的论文

🔍 AI Survey 🔍 🔍 🔍 🔍 🔍

搜索验证: 制备人脐静脉内皮细胞和人肺腺癌细胞融合细胞的新... 🔍

- ✓ 拆解问题: "drug resistance" AND mechanism 2024-2026 高级搜索
- ✓ 拆解问题: drug resistance mechanism 2024-2026
- ✓ 拆解问题: "drug resistance" mechanism 2024-2026
- ✓ 拆解问题: 查找近三年关于耐药机制的论文
- ✓ 验证条件: The paper discusses or investigates the mechanisms of drug resistance.
- ✓ 验证条件: The paper was published in 2024, 2025, or 2026.
- ✓ 搜索验证: 制备人脐静脉内皮细胞和人肺腺癌细胞融合细胞的新方法

Unveiling the mechanisms and challenges of cancer drug resistance Perfect

Sameer Ullah Khan, Kaneez Fatima, Shariqa Aisha, Fayaz Malik – Cell communication and signaling : CCS, 2024 – The University of Texas MD Anderson Cancer Center – Indian Institute of Integrative Medicine – Academy of Scientific and Innovative Research – doi.org

📄 Q1 📈 IF:8.9

该论文主要探讨了癌症治疗中耐药机制及其挑战, 并讨论了多种克服耐药的策略。论文内容与查找近三年关于耐药机制的论文高度相关, 符合研究方向。

☆加入知识库 📄 引用 📄 引用 被引用: 277

Emerging Therapeutic Strategies to Overcome Drug Resistance in Cancer Cells Perfect

Pankaj Garg, Jyoti Malhotra, Ravi Salgia, Sharad S Singhal – Cancers, 2024 – GLA University – City of Hope – City Of Hope National Medical Center – doi.org

📄 Q2 📈 IF:4.4

该论文主要探讨了癌症细胞中耐药机制的分子层面, 并提出了多种克服耐药性的新兴治疗策略。论文发表于2024年, 与近三年的要求相符, 因此与查询主题高度相关。

☆加入知识库 📄 引用 📄 引用 被引用: 202

Drug resistance in ovarian cancer: from mechanism to clinical trial Perfect

Ling Wang, Xin Wang, Haixia Wang, Dongling Zou – Molecular cancer, 2024 – Chongqing University – Chongqing Cancer Hospital – Chongqing Medical and Pharmaceutical College – doi.org

📄 Q1 📈 IF:33.9

该论文主要探讨卵巢癌中耐药机制的分类及其在临床试验中的应用, 通过文献综述详细介绍了耐药的多种分子机制。论文发表于2024年, 内容与耐药机制密切相关, 符合相关要求。

☆加入知识库 📄 引用 📄 引用 被引用: 178

Drug-resistant tuberculosis: a persistent global health concern Perfect

Maha Farhat, Helen Cox, Jennifer Furin, Madhukar Pai – Nature reviews. Microbiology, 2024 – Harvard University – Massachusetts General Hospital – University of Cape Town – Wellcome Centre for Infectious Diseases Research in Africa – German Center for Infection Research – Heidelberg University – Universitv Hospital Heidelberg – P. D. Hinduia Hospital and Medical Research Centre – Southern Africa Nazarene

癌症治疗耐药机制与应对策略综述 ✕

🔍 基于搜索结果生成的AI文献综述

癌症治疗耐药机制与应对策略综述

耐药机制分类

基因与表观遗传调控

肿瘤微环境影响

人工智能与治疗前景

其他疾病耐药机制

[1] 癌症耐药分子机制与挑战

[2] 肿瘤耐药: 从机制到临床

[3] 多层级癌症耐药机制

[4] 多层级癌症耐药机制

[1] 人工智能在肿瘤耐药中的机

继续生成

①拆解关键词 + 验证条件 -> ②AI自动过滤 -> ③每篇论文总结

一键生成论文摘要、
抓核心贡献点、数据图标解读

AI Copilot 深度文献阅读

AI Survey



自动生成思维导图
识别研究“热点”与“空白点”
并提供证据支持

领域趋势洞察

识别研究缺口

基于语义分析自动定位研究
空白并提供证据支持，精准
锁定创新切入点。

Idea发现

生成可落地研究构思

通过苏格拉底式对话
逐步明确方向并形成研究计划

WisPaper

肿瘤 nature communication 相关论文 使用机器学习算法 根据临床数据

Chatbox File Browser

Search complete Found 10 papers

好的，经过几轮检索，我找到了以下与你研究方向直接相关的论文。让我先整理一下，然后我们聊聊。

目前在 Nature Communications 上找到一篇高度匹配的：

Nature Communications 核心论文

"Metabolomic machine learning predictor for diagnosis and prognosis of gastric cancer"

- 期刊: Nature Communications, 2024
- 引用: 159 次
- DOI: 10.1038/s41467-024-46043-y
- 核心思路: 用 702 例多中心血浆样本做靶向代谢组学，机器学习模型 (10 种代谢物) 用于胃癌早期诊断 (外部验证灵敏度 0.905, 远超传统蛋白标志物的 <0.40)，同时还建了一个预后模型做风险分层。

其他相关的高影响力/高质量论文 (非 NC 但方向接近)

论文	期刊	年份	核心
Lung Cancer EHR Survival (Yuan et al.)	JAMA Network Open	2021	从 EHR 提取结构化+非结构化临床数据，半监督 ML 识别肺癌患者，预测 1-5 年生存
Breast Cancer			

Start DeepSeek Agent Start My Agent

New messages

Case history mode, input is locked

空白方向	状态
TabPFN 做肿瘤生存分析 (需回归/生存输出)	❌ 没人做过
TabPFN vs. Cox / RSF / DeepSurv 系统性基准测试	❌ 没人做过
TuneTables 微调 TabPFN 用于临床大样本	❌ 没人做过
TabPFN 做 pan-cancer 预后预测	❌ 没人做过
纯临床表格特征 (非影像组学) 的肿瘤预测	❌ 没人做过
TabPFN + 内部数据 外部验证	❌ 没人做过

Communications, aligning well with the focus on cancer learning in a clinical context.

Add to Library Cite Cited By: 0

Predicting response to enzalutamide and abiraterone in metastatic prostate cancer using whole-omics machine learning. 67.6% Perfect

Anouk C de Jong, Alexandra Danyi, Jeroen de Ridder, Martijn P Lokema - Nature communications, 2023 - Erasmus MC Cancer Institute - University Medical Center Utrecht - University of California, San Francisco - nih.gov

Q1 IF:15.7

The paper develops machine learning models using whole-genomics and transcriptomics data to predict treatment response in metastatic prostate cancer. It directly addresses cancer outcome prediction, employs machine learning on clinical data, and is published in Nature Communications, aligning with the specified criteria.

Add to Library Cite Cited By: 0

Advancements in Cancer Survival Prediction: A Systematic Review of Classical and Modern Approaches 58.6% Partial

Gitanjali Pradhani, Jonali Gogoi - Indian journal of community medicine : official publication of Indian Association of Preventive & So..., 2025 - Assam University - doi.org

Q4 IF:0.9

The paper is a systematic review examining advancements in cancer survival prediction using machine learning and deep learning models applied to clinical data. It highlights trends in prediction techniques and the increasing use of advanced methods. However, it was published in the Indian journal of community medicine, not in Nature Communications.

从流形角度分析大语言模型工作机制的研究进展

152.7K Credits ● 已终止

Chatbox 文件浏览

```
mcp_task-context_update_context mcp_task-context_update_context {'patch': {'id': "1079a68a-5b02-4edf-918e-e2d2012e959b", "url": "artifacts/1079a68a-5b02-4edf-918e-e2d2012e959b/proposal_experiment_plan.md"}}
```

Check line counts of both documents

Show command

两份文档都已生成并保存:

产出文件

- 研究对话与文献综述 (literature.md , 166 行)
 - 完整记录了我们对话的问题聚焦过程
 - 梳理了 20+ 篇相关论文, 按 7 个子主题分类
 - 记录了三个关键思想演进: 因果方向之辩、退火类比、post-training 几何不对称
 - 列出了 5 个开放问题
- 实验计划 (proposal_experiment_plan.md , 293 行)
 - 项目名称: GeoSFT — Geometry-Aware Supervised Fine-Tuning
 - 核心假说: SFT 的 entropy-seeking 膨胀中有大量“过度膨胀”, 几何约束可缓解 alignment tax
 - 三种策略: A (深层约束) / B (逐层梯度约束) / C (自适应约束)
 - 模型: Pythia-410M / 1B / 2.8B
 - 评估: 6 个 OOD benchmark + AlpacaEval + 逐层 erank 追踪
 - 预计工期: 10-15 天 (有 A100 的情况下)

有什么想调整的吗? 比如实验规模、策略细节、评估指标、或者优先级的安排?

案例历史模式中, 输入已锁定

搜索结果 搜索结果 搜索结果 搜索结果 搜索结果 搜索结果 研究计划 ...

3. 实验目标

P0 (核心)

- Pythia-1B 消融: Vanilla / v1 / SDP-only / SAP-only / SGR — 验证 SAP 独立贡献
- Qwen2.5-7B SGR: 验证是否消除 erank ↑ OOD ↓ 悖论

P1 (重要)

- Layer-Adaptive vs Uniform λ 对比
- Hyperparameter sweep: $\alpha \in \{0.1, 0.3, 0.5\}$, $\beta \in \{0.1, 0.3, 0.5\}$, $k \in \{8, 16, 32\}$

P2 (补充)

- Per-layer principal angles 训练轨迹可视化
- Erank vs Grassmann distance 解耦分析

P3 (可选)

- 与 SIFT 对比

4. 资源情况

可用代码/框架

- GeoSFT v1 代码: /root/autodl-tmp/geosft/code/ — 直接起点
- 推荐框架: PyTorch + HuggingFace Transformers + lm_eval

数据集

数据集	来源	规模	用途	获取
Alpaca	tatsu-lab/alpaca	52K	SFT 训练	公开
HellaSwag	lm_eval	~10K test	OOD	公开
ARC-Easv/Challenge	lm_eval	~2.5K/1.2K	OOD	公开

复制 下载

实验设计

交互式实验助手

通过自然语言交互和环境自动配置机制，降低用户运行实验的认知门槛。

云端容器化实验环境

为每个任务分配独立 GPU 环境，依赖与数据完全隔离。

The screenshot displays a web-based experimental environment. On the left, a chat interface shows a conversation with an AI assistant. The assistant has provided a detailed report on LLM mechanisms, including literature reviews and experimental plans. The user asks for GPU and image recommendations, and the assistant offers to start an A800 machine. On the right, a 'GPU 实例控制台' (GPU Instance Control Panel) shows a stopped RTX PRO 6000 instance. Below it, a terminal window displays the execution of various commands, including tailing logs and running training scripts. A notification at the bottom right indicates that the experimental report has been generated and is ready for download.

实验执行

从流形角度分析大语言模型工作机制的研究进展

152.7K Credits 已终止

Chatbox 文件浏览

mcp_gpu-runtime_stop_gpu mcp_gpu-runtime_stop_gpu {}

实验报告已写入完成, GPU 已释放。以下是 WizardLM 196K 全量实验的核心结论:

核心结果

方法	Mean erank	OOD 平均	WinoGrande
Pretrained	6.00	68.05	68.27
Vanilla SFT 196K	6.02	68.55	67.40
SGR SFT 196K	6.01	68.53	67.95 (+0.55)

关键发现

- Qwen2.5-7B 在 196K 规模下不存在 erank 坍塌 — Vanilla 和 SGR 的 erank 都稳定在 ~6.0, 接近预训练水平。SGR 的主要价值 (防止谱坍塌) 在此场景下无用武之地。
- SGR 的收益随数据规模递减:
 - 10K: SGR OOD avg +0.23 明显优势
 - 50K: SGR OOD avg -0.26 略低
 - 196K: SGR OOD avg -0.02 ≈ 持平
- WinoGrande 是 SGR 唯一稳定的优势项 — 在全部三个规模下 SGR 均提升 WinoGrande (+0.79 / +0.08 / +0.55), 说明 SAP 确实保护了常识推理子空间。
- 未来方向: 建议引入 scale-adaptive α/β (数据量越大正则化越弱)、选择性 SAP (只对特定层/子空间施加约束)、或结合 LoRA 测试 SGR。

开启DeepSeek版Agent

开启我的Agent

案例历史模式中, 输入已锁定

搜索结果 搜索结果 研究计划 配置 GPU 配置 GPU 研究计划

GPU 实例控制台

由 GPU Hub 提供算力支持

GPU 型号 RTX PRO 6000

控制台

```
[17:24:47] nvidia-smi && python3 -c "import torch; print(f'PyTorch: {torch.__version__}, CUDA: {torch.cuda.is_available()}, Devices: {torch.cuda.device_count()})"
[17:24:52] which python && python -c "import torch; print(f'PyTorch: {torch.__version__}, CUDA: {torch.cuda.is_available()}, Devices: {torch.cuda.device_count()})"
[17:24:57] ls /root/miniconda3/bin/python* 2>/dev/null; ls /opt/conda/bin/python* 2>/dev/null; which conda 2>/dev/null; find / -name "python*" -maxdepth 4 -type f 2>/dev/null | head -10
[17:25:02] export PATH=/root/miniconda3/bin:$PATH && python -c "import torch; print(f'PyTorch: {torch.__version__}, CUDA: {torch.cuda.is_available()}, Devices: {torch.cuda.device_count()})"
[17:25:08] export PATH=/root/miniconda3/bin:$PATH && pip install transformers datasets accelerate lm_eval 2>&1 | tail -5
[17:26:16] export PATH=/root/miniconda3/bin:$PATH && python -c " from transformers import AutoModelForCausalLM, AutoTokenizer import os model_path = '/root/autodl-tmp/models/qwen2.5-7b' if os.path.exists(model_path) and os.listdir(model_path): print(f'Model already exists at {model_path}') print(os.listdir(model_path)) else: print('Downloading Qwen2.5-7B...') os.makedirs(model_path, exist_ok=True) tok = AutoTokenizer.from_pretrained('Qwen/Qwen2.5-7B', trust_remote_code=True) tok.save_pretrained(model_path) model = AutoModelForCausalLM.from_pretrained('Qwen/Qwen2.5-7B', trust_remote_code=True) model.save_pretrained(model_path) print('Done downloading') "
Command timed out after 600s
[17:27:37] export PATH=/root/miniconda3/bin:$PATH && ls /root/autodl-tmp/models/qwen2.5-7b/ 2>/dev/null && echo "----EXISTS----" || echo "----NOT FOUND----"
[17:27:44] export PATH=/root/miniconda3/bin:$PATH && python -c " from huggingface_hub import snapshot_download import os model_path = '/root/autodl-tmp/models/qwen2.5-7b' # Check if model weights exist safetensors = [f for f in os.listdir(model_path) if f.endswith('.safetensors')] print(f'Existing safetensors: {len(safetensors)}') if len(safetensors) == 0: print('Downloading Qwen2.5-7B model weights...') snapshot_download('Qwen/Qwen2.5-7B', local_dir=model_path, local_dir_use_symlinks=False) print('Download complete!') else: print('Model weights already present') " 2>&1
```

实验报告已生成

下载实验产物

数据分析

Self-RAG 方法可以用于解决rag 回答中的忠实性问题吗?

本次消耗 331.9K Credits ● 已终止

Chatbox 文件浏览

资源管理器

- universal_run-4e22a0d9-b625-4514-a53c-2f741a79f56b
 - artifacts
 - 3a1a3c03-b13d-439f-a90c-9264272fefec
 - a16a884d-a5c6-447a-b08d-399d87458e21
 - faithfulness_qa_data_report.md 2026-04-22 18:51:22
 - faithfulness_qa_data_report.pdf** 2026-04-22 18:56:06
 - faithfulness_qa_data_report.tex 2026-04-22 18:51:01
 - a8881159-e179-424f-92d0-42e40d3cc2a7

究计划 搜索结果 搜索结果 搜索结果 搜索结果 搜索结果 预览 ...

faithfulness_qa_data_report.pdf 267.1 KB

下载

faithfuln... 1 / 12 92% +

Faithfulness-QA: Dataset Generation Report

Counterfactual Entity Substitution for RAG Faithfulness Training

Qiang Zhang
Fudan University
<https://github.com/qzhangFDU/faithfulness-qa-dataset>

April 2026

Abstract

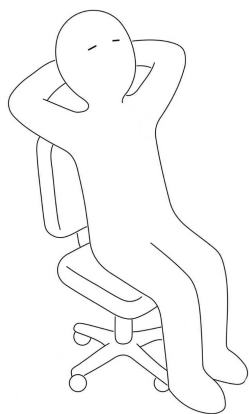
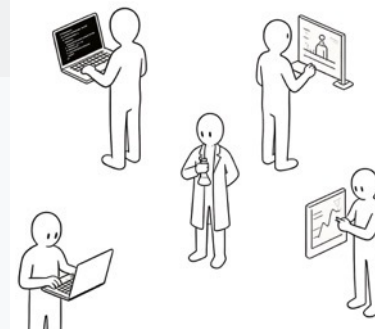
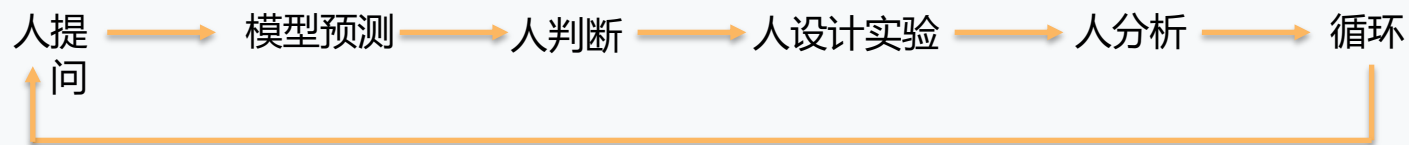
This report documents the data generation process and quality analysis for the **Faithfulness-QA** dataset—a benchmark designed to train and evaluate the faithfulness of Retrieval-Augmented Generation (RAG) models to retrieved context. The dataset is constructed via *counterfactual entity substitution*: for each QA sample, a key entity in the context is replaced with a same-type alternative, creating a controlled conflict between the context and a model's parametric knowledge. We successfully generated **99,094** high-quality samples from two source datasets (SQuAD: 49,094; TriviaQA: 50,000), covering 8 named entity types with a combined entity bank of 76,953 unique entities. All quality checks achieve 100% pass rates on 200-sample audits. The code and data are publicly available at <https://github.com/qzhangFDU/faithfulness-qa-dataset>.

Contents

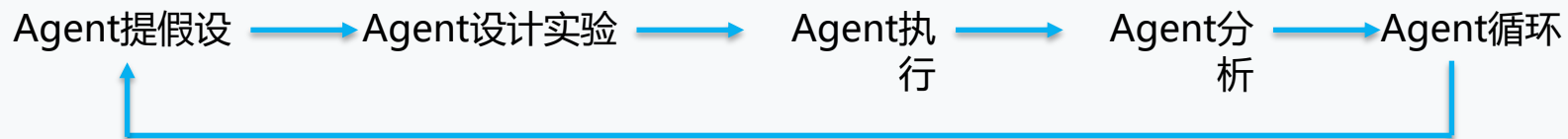
1 Introduction	3
1.1 Motivation	3
1.2 Core Idea	3
1.3 Report Scope	3
2 Data Generation Pipeline	3
2.1 Pipeline Overview	3
2.2 Stage 1: Entity Bank Construction	4

论文写作

Before 人 + 算力 + 模型



After Agent 自己做科研



人类：设定目标+监督边界

6

未来展望与总结

Future Outlook & Summary

从「AI 辅助」到「AI 自主」：渐进式自主化路线图

Level 1	AI 工具	2020-22	搜索引擎、统计软件
Level 2	AI 研究助手	2023-25	AI 辅助文献、编码、写作
Level 3	AI 科研合作者	2026-28	AI 自主完成子任务，人类把控方向 ← 当前
Level 4	AI 主导研究者	2028-30	AI 主导研究，人类监督和验证
Level 5	AI 独立科学家	2030+?	完全自主发现、设计、验证、发表

「Graded Autonomy」(分级自主) —— 系统在机器速度闭环运行，但始终锚定在人类的优先事项上

挑战与思考

技术挑战

引用幻觉	AI 编造不存在的参考文献
可重复性	缺乏标准化复现框架
深层理解	高级模式匹配 vs. 真正理解
工具使用	CCTU: 所有模型 < 20%
上下文学习	CL-BENCH: GPT-5.1 仅 23.7%

伦理与社会挑战

论文洪水	AI 不间断生产 → 审稿系统淹没
角色危机	年轻科学家丧失核心实验能力
责任归属	AI 发现出了错，谁负责？
认识论挑战	没有理解的发现算“知识”吗？

核心共识：「Graded Autonomy」—— 系统在机器速度闭环运行，但始终锚定在人类的优先事项、可验证的机制和学科适配的理解之上。

三个关键信息

1

AI Scientist 已从概念变为现实

- 能自主生成想法、做实验、写论文、通过评审
- 在特定领域已超越人类最佳水平 (DeepScientist)
- Nature 连发两篇 AI 科研系统论文 (Co-Scientist, ERA)

2

当前模型的基础能力仍有显著短板

- 工具使用: CCTU 上所有模型 PSR < 20%
- 上下文学习: CL-BENCH 上平均解决率仅 17.2%
- 科学工具使用: SciAgentGym GPT-5 仅能达到 30.9%

3

未来 5 年是关键窗口期

- 闭环实验室、跨学科通用化、科研平权化
- 需要同步建立伦理框架和评估体系
- WisPaper 等平台将降低 AI 科研门槛

Thank You & Q&A

参考资源: Google Co-Scientist (Nature 2026), ERA (Nature 2026),
OpenNovelty, CCTU, CL-BENCH, SciAgentGym, MM-Doc-R1