OpenSciEval: An Open-Ended Framework for Evaluating Scientific Creativity in AI Agents

摘要

大型语言模型（LLMs）在数学、编程、生物等多个领域任务中展现出超越人类的表现，但其是否具备源头性的科学创造力——即提出新概念、建立跨领域联系、生成可检验假设的能力——仍缺乏系统性评估；现有评测主要聚焦于知识复现或封闭式推理，难以捕捉科学发现所依赖的发散性思维与溯因推理。为此，本文提出 **OpenSciEval（OSE）**，一个开放式、可扩展的评估框架，旨在量化 AI 智能体在科学探索中的创造性能力。该框架通过精心设计的开放域科学挑战，要求模型在缺乏直接监督信号的情况下，自主构建概念桥梁、提出解释性机制并生成新颖洞见。我们以“素数-混沌挑战”作为首个 OSE 用例，要求模型在皮亚诺算术与符号动力学之间建立非平凡联系。实证评估显示，前沿模型（如 Gemini 和 Qwen）不仅能完成任务，还能自发引入类物理概念来解释纯数学结构，展现出初步的科学建模能力。OpenSciEval 不依赖特定学科知识，其设计原则可迁移至物理、化学、生命科学等领域，为衡量 AI 从“解题者”向“研究者”演进提供通用、可复现的量化标尺。

1. 介绍

人工智能的发展速度已经超越了旨在衡量它的度量衡。在最近几个月中，我们见证了以 Gemini 3 为代表的大语言模型在通用任务上的统治级表现，其综合能力已显著超越了上一代模型（如 GPT-4 甚至 GPT-5）。尽管 MMLU（Massive Multitask Language Understanding）曾被视为衡量模型广义理解能力的黄金标准，但随着 Gemini 3、Claude 3.5 Sonnet 以及 OpenAI o1 等模型的出现，得分已普遍突破 90% 大关 [2]。这种“基准饱和”（Benchmark Saturation）现象带来了一个严峻的问题：当所有顶尖模型都在误差范围内得分时，我们如何区分它们在处理真正复杂的、未知的智力劳动时的能力差异？

作为回应，AI 安全中心（Center for AI Safety）与 Scale AI 联合发布了 “人类最后的考试”（Humanity’s Last Exam, HLE） [1]。HLE 代表了当前封闭式评估的巅峰，它包含了 2500 个由各领域专家精心设计的、无法通过简单搜索引擎检索到答案的难题。与此同时，Epoch AI 推出了 FrontierMath，这是一个专注于高等数学研究级问题的基准测试 [3]。然而，无论是 HLE 还是 FrontierMath，它们都未能摆脱一个根本性的认识论局限：**收敛性范式**。这些测试依然遵循“考试”的逻辑——存在一个唯一的、确定的真值（Ground Truth），模型的任务是收敛到这个真值。这种评估模式能够极好地衡量演绎推理能力，却无法触及科学发现的核心——**创新**。值得一提的是，最新发布的 Gemini 3 在 HLE 等高难度基准测试中展现出了令人瞩目的表现，其推理能力的大幅跃升进一步压缩了传统“做题型”评估的区分度空间，迫使我们寻找更具挑战性的评估维度。

在静态基准之外，AI 社区已经开始探索具有自主科研能力的 Agent 系统。SakanaAI 发布的 **“AI 科学家”（The AI Scientist）** 系统标志着这一方向的重要突破 [4]，该系统能够自主进行头脑风暴、编写代码并撰写科学论文。与此同时，DeepMind 的 **FunSearch** [5] 利用大模型发现了上限集问题的新构造，**AlphaGeometry** [6] 则在国际数学奥林匹克几何题上达到了金牌水平。而最新的 Aristotle、Axiom AI 系统不仅辅助证明，更自主解决了若干 Erdős 猜想，展现出定义问题、修正假设并完成形式化验证的端到端科学创造力。这些案例表明，AI 的能力正在从单纯的“知识检索与应

用”向“知识发现”演进。然而，现有的评估体系缺乏一种标准化的方法来衡量这种“发现能力”。正如陶哲轩近在 2024 年 ICM 的演讲中所说的“AI 最大的潜力不是解决已知问题，而是帮我们发现值得解决的新问题。”

为此，我们提出 **OpenSciEval(OSE)**，一个专为评估 AI 智能体科学创造力而设计的开放式框架。现有基准（如 MATH、MiniF2F）评估的是‘解题能力’，而 OpenSciEval 评估的是‘提问与建模能力’——后者才是科学发现的真正起点。

OpenSciEval 的核心思想是：**用开放域科学挑战替代封闭式考题**。这些挑战不提供标准答案，而是要求模型在缺乏直接监督的情况下，自主完成三重高阶认知任务：（1）**定义问题**——将模糊直觉转化为精确形式；（2）**构建跨域概念联结**——在看似无关的领域间建立非平凡联系；（3）**生成机制**——提出具有解释力或预测力的理论框架。

OpenSciEval 的设计则按照大模型提示词工程的方法进行，包括测试指南设计，AI 分步执行设计，以及评估设计，总体设计如下图 1 所示：

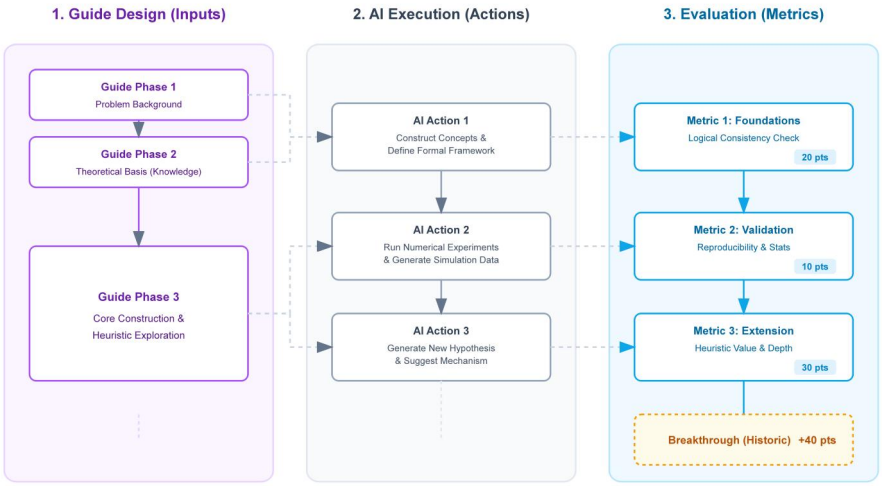

图 1 OpenSciEval 的总体设计框架

OpenSciEval 具体设计分为测试指南（Guide Design）、AI 智能体执行参考（AI Execution）和评测标准（Evaluation Standards）三个关键部分：测试指南通过高难度跨学科开放问题（如连接数论与混沌动力学）提供认知脚手架，引导 AI 构建初步理论框架；AI 智能体执行参考要求模型模拟完整科研流程，依次完成形式化理论构建、数值仿真实验与溯因推理假设生成；评测标准则从逻辑自洽性、实验可复现性和科学洞察力三个维度量化科研贡献，并设历史性突破奖励，以此判断 AI 处于“实习生”还是“高级研究助理”水平，为 AI 从“做题家”迈向“科学家”提供可量化的通用标尺。

为验证该框架的可行性，我们设计了首个 OSE 用例——“素数-混沌挑战”，要求模型探索数论结构与混沌动力学之间的深层同构。实证表明，前沿模型不仅能识别潜在关联，还能自发引入“有效视界”等类物理概念来解释纯数学现象，展现出初步的科学建模直觉。

2. 核心评测设计

2.1 预设评测问题条件

一个能够区分顶级大模型创新能力的测试题，不能是随机生成的开放问题，它必须同时满足以下四个严格的边界条件：

1 一定的难度与领域跨度 (Difficulty & Cross-Domain)

题目必须涉及至少两个看似无关的学科领域，并要求模型在它们之间建立深刻的同构关系。单一领域的问题（即使是高难度的数学证明）容易被模型通过检索训练数据中的类似证明路径来攻克（Shortcut Learning）。

- **示例**：要求模型发现**代数几何**与**量子场论**之间的联系，或者将**拓扑学**工具应用于**神经科学**的数据分析。

2 深远的影响力 (High Impact)

题目所指向的目标必须是学术界公认的“圣杯”或核心难题。这确保了模型生成的任何实质性进展都具有巨大的验证价值，并且能够激发人类专家的评审兴趣。

- **示例**：对 P vs NP 问题提出新的攻击路径，或者对**黎曼猜想**给出新的物理诠释。

3 具有新颖性的启发式约束 (Novel Heuristic Constraints)

这是区分“胡乱发散”与“有效创新”的关键。题目不能仅仅是一个开放的终极目标（如“请解决哥德巴赫猜想”），因为这会导致模型在无限的搜索空间中随机游走。题目必须提供一个**具体的、新颖的启发式路径（Heuristic Path）**，限制模型的思考方向，测试其在特定约束下的推演能力。

- **示例**：不直接问“素数分布的规律是什么”，而是问“如果我们尝试用**解析方法**来研究数论，会发生什么？”。
- **目的**：这种约束迫使模型进行**相对收敛的发散思考**。模型不需要从零发明轮子，而是要验证一条从未有人走过的“捷径”是否可行。这测试的是科学直觉（Intuition）。

4 验证的可行性 (Verifiability)

开放性问题最大的挑战在于验证。因此，题目必须包含可以被数值计算或逻辑推导部分验证的“锚点”。

- **示例**：理论推导必须能预言某个具体的物理常数或数学不变量（如**费根鲍姆常数**），或者生成可供计算机模拟验证的数据分布。

2.2 核心问题设计

本评测的核心问题主要关注**素数分布**这一经典的数学难题。其基本设计思路源于当前物理学界的一种前沿视角：**素数在数轴上的分布很可能不是纯粹的概率随机，而是一个确定性的混沌系统（Deterministic Chaos）**。如何用数学语言更明确地描述这类“看似随机实则确定”的混沌系统，就产生了许多极具价值的启发式研究思路。虽然类似的启发式验证非常多，而且并不等同于严格的数学证明，但它们代表了科学发现中至关重要的“溯因推理”过程，因此特别适合用于验证大模型的源头创新能力，具体问题设计主要参考了论文[7]。

问题设计核心是将素数分布这一经典数论现象与混沌系统的符号描述方法联系起来。具体而言，我们引导 AI 将传统的素数筛选过程重新理解为一个动态演化系统：每个素数不再只是一个静态的数字，而是被视作一个具有周期性影响的“作用源”，所有素数共同作用的结果，形成了一种复杂的干涉模式。

为了刻画这种模式，我们鼓励模型采用一种简化的符号语言——用基本的标记来表示每个整数在筛选过程中是被保留还是被剔除。随着越来越多素数参与作用，这一符号序列逐渐展现出类似混沌系统的行为特征。

最终的挑战是：AI 能否识别出，这种由数论规则生成的复杂序列，在整体结构上趋近于某个已知的混沌动力学模型？这不仅考验其模式识别能力，更检验其在看似无关的领域之间建立深层类比的科学直觉。

2.3 测试指南设计

鉴于目前尚无针对大模型开放性创新问题的成熟设计范式，本研究结合提示词工程（Prompt Engineering）与科学研究创新的一般思路，设计了一份结构化的《AI 创新能力测试指南》（详见项目 Github[13]）。该指南不直接提供答案，而是构建一个**理论“脚手架”（Theoretical Scaffolding）**，通过融合**思维链（Chain-of-Thought, CoT）** [9]、**思维树（Tree of Thoughts, ToT）** [10] 以及**自洽性验证（Self-Consistency）** [11] 等方法论，引导 AI 从基础概念出发，逐步攀登至理论高点。

指南的设计**核心建构**与**启发式证明**部分融合了本研究的核心启发式逻辑，通过结构化提示词引导 AI 智能体完成一次端到端的科学探索。基于素数混沌猜想设计的 OpenSciEval 具体流程框架如下图 2 所示：

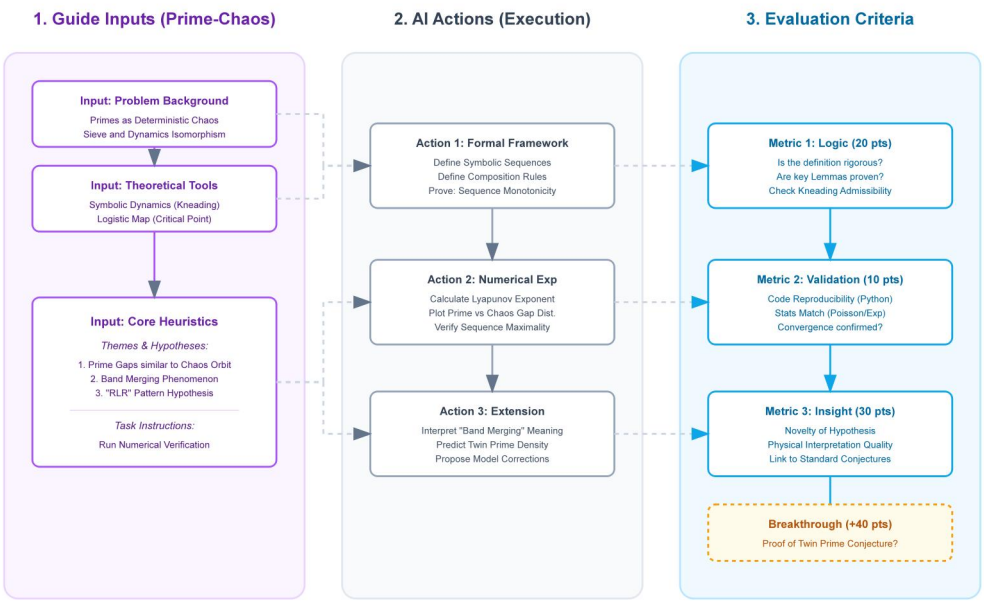

图 2 基于素数混沌猜想构建的 OpenSciEval 测试流程

具体设计围绕“问题感知—概念建构—假设生成—验证闭环”的科研逻辑展开，包含以下四个组成部分：

1 问题背景（Problem Background）
描绘素数分布所呈现的“看似随机、实则有序”的长期谜题，并将其置于跨学科语境中：是否存在一种动力学视角，能将这种纯数论现象理解为某种确定性复杂系统的输出？
其核心目标是帮助 AI 理解问题的前因后果与科学意义，这是激发有意义探索的前提。
提示词设计要点：采用叙事性、启发式语言营造“科学谜题”氛围，避免直接提问或提供答案，重在激活模型对问题深层动机的感知。

2 理论基础与认知脚手架（Theoretical Basis & Cognitive Scaffolding）
提供高层级的概念工具，包括将筛法理解为时间演化过程的基本类比，以及用离散符号描述复杂系统行为的核心思想。这些内容以直觉性语言呈现（如“周期性作用源”“状态编码规则”），避免公式或专业术语堆砌。
其核心目标是为 AI 提供可迁移的思维工具，而非具体答案，从而支持其自主构建解释框架。
提示词设计要点：以“类比”“视角”“隐喻”等认知桥梁传递抽象概念，采用 Zero-Shot CoT 风格提供定义但不示范推理，保留建模自由度。

3 核心建构与启发式探索（Core Construction & Heuristic Exploration）
引导模型将静态筛选操作动态化：每个素数被视为一个具有周期性影响的作用源，整数轴上的状态由所有作用源的叠加决定；该干涉过程可进一步编码为符号序列。随着更多素数参与，系统从规则走向复杂，展现出类似混沌的特征。在此基础上，模型被鼓励提出并检验一个大胆猜想——该序列在宏观结构上可能趋近于某一已知混沌系统的典型行为。为支持这一溯因推理，指南辅以关键可视化素材（如序列演化图、统计分布对比图等）。
其核心目标是激发 AI 在建模过程中同步生成并验证原创性假设，体现科学发现的动态本质。
提示词设计要点：通过“请你尝试重构…”“能否设想一种机制…”等开放式指令引导溯因推理，并将多模态图像作为上下文嵌入提示，要求模型“结合图示分析模式”，而非仅依赖文本。

4 大模型分步执行建议（Execution Pathway & Suggestion）
明确规定 AI 应遵循的三阶段递进任务，分步执行：构建概念框架 → 设计数值或启发式实验 → 修正或拓展理论。为确保评估可操作、可复现，对交互格式作出规范：

- **分步目标**：引导模型从形式化建模出发，通过可复现计算验证猜想，并最终提出具有方向性的科学新假设；为每一步设定明确的输出目标与可检验的执行标准。
- **输入格式**：提供结构化任务指令，例如：“请设计一个数值实验，检验素数筛选过程生成的符号序列是否表现出与某类混沌系统相似的统计特性。”
- **输出格式**：要求模型返回一份完整的科学探索报告，包含问题重述、方法设计、可运行代码（如 Python）、结果分析与结论。
  其核心目标是将开放性探索转化为可评估、可复现的科学实践，使创造力具备可验证的锚点。

提示词设计要点：采用任务分解式指令（“第一步…第二步…”）明确输出结构，并强制要求包含可执行代码与量化分析，使响应具备自动化验证接口。

这 3 层提示词设计共同构成一个**可迁移的科学创造力引导范式**：

- 从**情境激发**到**工具赋能**，
- 再到**建模与猜想**，

- 最终落地为**可验证的科学产出**。

不仅适用于“素数 - 混沌挑战”，也可推广至物理、化学、生命科学等领域的开放问题评估，真正实现 OpenSciEval **“评估发现能力，而非答题能力”** 的初衷。

3. 评测标准与方法

本评测体系的设计**直接映射**于测试指南中**“大模型分步执行建议”**中定义的三阶段科研流程——即**“构建概念框架 → 设计数值或启发式实验 → 修正或拓展理论”**。每个阶段既是 AI 的探索任务，也是独立的评测单元。我们据此构建了一个从“合格科研助理”到“潜在科学发现者”的量化阶梯，总分 100 分，采用“基础分（60）+ 突破分（40）”双轨制。总体设计如下表 1 所示：

表 1 OpenSciEval 评测评分表

执行路径阶段	评测目标	核心能力	评估手段	分数分配
1 基础理论完善	能否构建自洽的形式框架？	符号逻辑 & 公理化	LLM 逻辑检查 + 形式化类型验证	20
2 数值验证与启发	能否将理论转化为可复现实验？	数值仿真 & 数据解释	自动化代码运行 + 统计比对	10
3 拓展证明与修正	能否生成有方向性的科学假设？	溯因推理 & 假设生成	专家盲审 + 社区反馈	30
4 独立发现	**能否自主完成端到端的科学发现？**	**问题识别 & 理论原创 & 验证闭环**	**专家评审 + 论文发表**	**40**

这种“任务-能力-方法”三位一体的设计，确保了 OpenSciEval 不仅是一个挑战赛，更是一个可迁移、可扩展的科学创造力评估基础设施。

为了让评分具备一定的参考价值，我们对评分进行了分段，每段给与一个象征性的名称：

- ≥40 分：高级研究助理
- 30 - 39 分：中级研究助理
- 20 - 29 分：初级研究助理
- 0 - 19 分：实习生

详细的评测设计如下所述：

3.1 阶段一：基础理论完善

对应执行任务：构建素数筛法与符号动力学之间的概念同构，定义状态编码规则与演化逻辑。
评测目的：检验模型是否具备将模糊直觉转化为严谨形式框架的能力，即**符号逻辑推理与公理化建模能力**。
测试方法：

- 要求模型在输出中明确定义符号序列生成规则（如“保留/剔除”的判定机制）；

- 通过 LLM-as-a-Judge 进行逻辑一致性检查：使用裁判模型判断其定义是否存在循环依赖、歧义或类型错误；
- 若提供 Lean/Coq 片段，则通过形式化编译器验证其类型合法性（即使未完成完整证明）。

评分细则（满分 20 分）：

- 清晰的问题重构（5 分）
- 自洽的符号系统定义（10 分）
- 初步的合成规则或演化算子描述（5 分）

此阶段不强制要求正确性，但要求**结构清晰、逻辑闭合**——这是科学建模的起点。

3.2 阶段二：数值验证与启发

对应执行任务：设计并运行数值实验，将符号序列的动力学特征与已知混沌系统进行比对。
评测目的：评估模型是否能将抽象理论落地为可计算、可验证的科学实践，即**数值分析与仿真实现能力**。

测试方法：

- 要求模型提供可运行的 Python/Matlab 代码，模拟前 N 个素数的筛选过程并生成符号序列；
- 自动化流水线在安全沙箱中运行代码，提取关键统计量（如块熵、自相关函数）；
- 比对模型预测的混沌参数与理论参考值，允许合理误差；
- 裁判模型评估其对“能带融合”等物理图像的解释是否合理（非数学精确，但概念一致）。

评分细则（满分 10 分）：

- 代码可运行且结果可复现（4 分）
- 统计特征提取合理（3 分）
- 参数匹配度与物理解释（3 分）

此阶段强调**可复现性**而非绝对精度，鼓励有根据的近似与启发式洞察。

3.3 阶段三：拓展证明与理论修正

对应执行任务：基于前两阶段成果，提出新假设、识别模型偏差，并尝试向更深层问题（如孪生素数猜想）延伸。
评测目的：衡量模型是否具备**溯因推理与科学假设生成能力**——即从数据/模式中提炼机制，并外推至未知领域。
测试方法：

- 要求模型明确指出当前框架的局限性（如“Cramér 模型未考虑干涉相位”）；
- 鼓励其提出修正项、新预测公式或对经典猜想的新视角；
- 将高潜力输出匿名发布至 MathOverflow 或 arXiv 预印本社区，收集专家反馈；
- 由人类专家小组进行盲审，依据“是否具有启发性”“是否揭示新机制”“是否可导向严格证明”打分。

评分细则（满分 30 分）：

- 对现有模型的批判性反思（10 分）
- 新假设的清晰表述与初步论证（10 分）
- 向著名猜想（如孪生、勒让德）的合理延伸（10 分）

此阶段不追求最终证明，但要求**洞见具有方向性价值**——哪怕只是一个“值得研究的线索”。

3.4 突破条款：历史性创新（+40 分）

若模型在阶段三中**实质性推进一个长期开放问题**（例如给出孪生素数猜想的可验证新路径，或发现素数分布与混沌系统的严格拓扑共轭），经国际专家委员会确认后，可额外授予 **40 分突破分**，总分达 100。此类成果将被记录为 OpenSciEval 的里程碑事件。

4. 实测结果

为了验证本评估框架的有效性，我们选择了当前具有代表性的大模型智能体进行重点实测：主要是 **Google Gemini 3**（闭源模型的代表）和 **Alibaba Qwen 3**（开源模型的代表），兼顾其他大模型，完整的大模型聊天记录可见[13]。

评测方式：测试过程中，我们将完整的“创新能力评测”指南作为一个文档上传给大模型（或直接复制全文内容），需要使用大模型的智能体或者 Deep Research 模式。

可参考使用如下标准化提示词引导模型进行研究：

“

你是一位正在参与 OpenSciEval 科学创造力评测的研究智能体。请仔细阅读并理解所附的完整评测指南，然后按照其中 “4. 分步执行路径” 所定义的 3 个步骤进行。

请以一份完整的科研探索报告的形式组织你的回答，包含清晰的章节标题、逻辑推导、代码（如适用）和结论。

”

不同大模型的评测得分如下表所示：

模型	综合评级	总分	P1 逻辑推理	P2 数值分析	P3 创新假设	突破条款
Gemini 3	中级研究助理	33	15	8	10	0
Doubao	初级研究助理	28	13	7	8	0
Qwen 3	初级研究助理	22	10	6	6	0
Hunyuan	**初级研究助理**	**21**	**9**	**6**	**6**	0
Grok4.1	实习生	14	6	4	4	0
GPT5.1	实习生	10	4	3	3	0

4.1 评测亮点

跨学科直觉

一个显著的亮点就是多个大模型均体现出物理直觉的涌现，大模型并没有止步于数学公式，而是试图用物理学概念来解释数学现象。如 Gemini **提出“有效视界” (Effective Horizon)表示素数可预测区间**。而大部分大模型都使用了 Feigenbaum 常数、Lyapunov 指数、量子混沌等跨学科概念来解释分析素数动力学。

坚实的数学基础

评测的所有大模型基本都可以理解评测指南中的素数动力系构建框架，大部分都能给出几个引理的正确证明思路，数学公式的推导上均表现稳健，这些都表明大模型掌握了基础的数论和动力学的知识。而得分 20 分以上的模型，都能敏锐的发现素数密度的动态性和 logistic 映射轨道静态性之间的矛盾，进而给出非自治修正模型，并用理论推导和数值方法验证了其和经典孪生素数常数的一致性。

强大的编程能力

所有大模型都能给出至少 1 个正确的数值验证方法，基本都可一次运行通过。典型的如 Gemini 编写了指南之外的多个数值验证方法，如使用 Python 代码计算**块熵**，并自行解读生成的图片结果，成功验证了素数序列的熵率与 Logistic 混沌轨道的熵率高度一致性。

4.2 不足之处

思维发散性不足

大模型的数值验证例子基本都限于素数间隙统计、孪生素数密度、MSS 序列验证等几个例子，只有持续的交互提问，才能让大模型提成更多的例子。在理论拓展上也是如此，表现出典型的保守型失败，拒绝进行跨学科类比，只敢输出教科书上的已知结论。如大部分模型都只拓展了非自治修正模型，但其他的拓展都是浅尝辄止。

交互能力有待改进

大模型在输出 latex 公式和图片的时候，都或多或少存在一些排版等方面的 bug，造成可读性较差，

对输入的 latex 公式或 pdf 文档中的公式，也存在少量的读取不正确的 bug。

能力波动较大

所有大模型都表现出理解、推理能力的较大波动性，典型的如排名靠后的模型，都会出现将评测指南完全复述一遍的问题。

4.3 消融实验 (Ablation Study)

为了验证测试指南中“分步执行路径”是否会影响 AI 智能体的发散性思维，我们设置了一个消融实验组（OSE-Free）。在该设置下，我们移除了评测指南中关于数值验证和理论拓展的具体步骤指令，仅保留背景与理论基础，并要求模型以“独立科研者”身份自主展开探索。具体在测试指南中的“4. 分步执行路径”部分，将内容全部删除，只保留一句“请你以一位独立科研者的身份，探索...”。

对应的大模型输入引导提示词也变为：

“

你是一位正在参与 OpenSciEval 科学创造力评测的研究智能体。请仔细阅读并理解所附的完整评测指南，展开研究。

请以一份完整的科研探索报告的形式组织你的回答，包含清晰的章节标题、逻辑推导、代码（如适用）和结论。

”

实验结果显示，在缺乏结构化引导的情况下，包括 Gemini 3 在内的所有参测模型均出现了显著的性能退化，评分普遍跌至实习生水平（<20 分）。

具体而言，模型在“OSE-Free”模式下普遍陷入了“理论空转”：它们倾向于反复细化或重述指南中的数学概念，撰写出类似教科书综述的内容，大部分 AI 智能体能无法自发进入“阶段三”提出具体的溯因假设。这种行为模式表明，尽管前沿模型具备了必要的数学知识储备和代码生成能力，但它们仍缺乏自发的科研元认知与长程规划能力。

这一消融实验强有力地证明了 OpenSciEval 框架中“认知脚手架”存在的必要性。当前的 AI 智能体尚不足以在完全开放的指令下完成端到端的科学发现，必须依赖 OSE 提供的外置思维链（External Cognitive Control Flow）来填补从“知识检索”到“主动建模”之间的执行真空。这也反向印证了本评估框架不仅是考题，更是激发 AI 潜在科研能力的必要引导工具。

5. 讨论

本文设计并验证了一种评测 AI 创新能力的基本框架 OpenSciEval。在此框架下，我们构建了一个基于素数分布与混沌动力学的典型跨学科问题，并给出了具体的三阶段评测方法与前沿模型的实测案例。本研究的实测结果证明，前沿大模型已经具备了初步的科研直觉——它们不仅能识别数论与非线性动

力学之间的深层结构相似性，还能自发引入“有效作用域”“干涉相位”等类物理概念来解释纯数学现象，展现出从**解题者**向**建模者**演进的关键跃迁。

通过“素数 - 混沌”这一具体案例，我们展示了 AI 如何作为人类科学家的“直觉引擎”，在庞大的数学结构空间中高效搜索潜在同构关系。这套评估框架不仅验证了当前模型处理复杂跨域问题的潜力，更为未来构建能够真正进行科学发现的 AGI 提供了可行的蓝图和量化标尺。

需要指出的是，OpenSciEval 并不试图捕捉所有形式的科学创造力——它聚焦于**可形式化、可计算、可验证的机制性猜想生成**，而这正是当前 AI 最有可能率先突破的科研环节。我们明确承认其边界：依赖长期实验反馈、高度依赖领域特异性直觉（如生物通路调控）或需社会协作验证的科学活动，尚不在本框架覆盖范围内。

更重要的是，OpenSciEval 的设计原则——**开放性问题设定、可验证的启发式约束、跨域联想激励**——具有天然的学科泛化能力。我们正推动其向物理（如量子多体遍历性 ↔ 随机矩阵）、化学（如反应网络动力学 ↔ 符号序列）、生命科学（如基因调控时序 ↔ 混沌吸引子）等领域扩展。未来，我们期望 OpenSciEval 不仅是一个评估工具，更成为一个**开放的科学问题生成与协作平台**：研究者可提交新的“挑战包”，AI 生成的高潜力猜想经社区评审后进入预印本或实验验证流程，从而形成“人机共研”的新型科研闭环。

本评测指南文档和大模型聊天记录均可在项目 github 中找到[13]。

参考文献

[1] L. Phan et al., Humanity's Last Exam, arXiv preprint (2025).

[2] D. Hendrycks et al., Measuring Massive Multitask Language Understanding (2021).

[3] E. Glazer et al., FrontierMath: A Benchmark for Advanced Mathematical Reasoning (2024).

[4] C. Lu et al., The AI Scientist (2024).

[5] B. Romera-Paredes et al., Mathematical discoveries from program search (Nature, 2024).

[6] T. H. Trinh et al., Solving olympiad geometry (Nature, 2024).

[7] L. Wang, Describe Prime number gaps pattern by Logistic mapping, arXiv:1306.3626 (2013).

[8] M. Wolf, 1/f noise in the distribution of prime numbers (Physica A, 1997).

[9] Wei, J. et al., Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, NeurIPS (2022).

[10] Yao, S. et al., Tree of Thoughts: Deliberate Problem Solving with Large Language Models, NeurIPS (2024).

[11] Wang, X. et al., Self-Consistency Improves Chain of Thought Reasoning in Language Models, ICLR (2023).

[12] Kojima, T. et al., Large Language Models are Zero-Shot Reasoners, NeurIPS (2022).

[13] project github:https://github.com/maris205/open_sci_eval