人类最终猜想：基于素数分布设计的 AI 创新能力评测

摘要

随着大语言模型（LLM）在 MMLU、MATH 等传统基准测试中逐渐逼近人类能力的极限，人工智能领域正面临一场深刻的认识论危机：现有的评估体系正日益丧失区分模型“知识复现能力”与“源头创新能力”的效度。尽管“人类最后的考试”（Humanity’s Last Exam, HLE）与“FrontierMath”等前沿基准试图通过提高试题难度与封闭性来延续评估的有效性，但它们本质上仍局限于“收敛性思维”的考察——即在已知解空间内寻找既定答案。然而，科学发现的核心在于“发散性思维”与“溯因推理”（Abductive Reasoning），即在未知领域构建新定义、发现新同构、提出新猜想的能力。

本报告提出了一种全新的评估范式—— **“创新能力图灵测试”（Innovation Turing Test）**，并具体构建了一个基于数论与非线性动力学交叉领域的开放性测试用例：“素数混沌猜想”（Prime-Chaos Conjecture）。该测试要求大模型基于一份启发式研究指南，在皮亚诺算术公理体系与符号动力学之间建立形式化桥梁，论证素数分布的伪随机性实为低维确定性混沌的体现，并精确推导 Logistic 映射在能带融合点（$u \approx 1.5437$）的拓扑性质。

本研究详细阐述了该测试的理论基础、评测条件、评测得分等设定逻辑，构建了一套针对开放性科学问题的评估标准，以及一套人和 AI 协作、RAG 知识库共享的规模化评测方法。然后针对 Gemini 与 Qwen（通义千问）等前沿模型进行了实测。本研究的目标旨在为 AGI（通用人工智能）从“解题者”向“研究者”的跃迁提供量化标尺。

1. 绪论：推理模型时代的评估危机与范式转移

1.1 静态基准的饱和与“人类最后的考试”

人工智能的发展速度已经超越了旨在衡量它的度量衡。在过去几年中，我们见证了以 GPT-4 为代表的大语言模型在通用任务上的统治级表现。MMLU（Massive Multitask Language Understanding）曾被视为衡量模型广义理解能力的黄金标准，涵盖了从初等数学到职业法律考试的广泛主题。然而，随着 DeepSeek-V3、Claude 3.5 Sonnet 以及 OpenAI o1 等模型的出现，MMLU 的得分已普遍突破 90% 大关，实质上使其成为了一项“已解决”的任务 [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。这些题目不仅涉及数学、人文学科和自然科学，还具有多模态特性，约 14% 的题目需要理解图表，旨在测试模型的深度推理能力而非单纯的模式匹配 [1]。与 MMLU 不同，HLE 设计之初便设定了极高的门槛，现有最先进模型（SOTA）在初次测试中的准确率甚至低于 5%，这在一定程度上缓解了评估指标的通货膨胀 [1]。

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]。FrontierMath 不仅要求计算，更要求极其复杂的符号推理和数学直觉，其题目难度往往需要人类数学专家耗费数小时甚至数天才能解决。这些基准测试无疑推高了“推理能力”的定义，将 AI 评估从“本科生水平”拉升至“博士生水平”。

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模型在既定规则下的演绎推理能力（Deductive Reasoning），却无法触及科学发现的核心—— **创新**。创新往往意味着打破规则、建立新的公理体系、或者在两个看似无关的领域间建立同构关系（Isomorphism）。真正的科研工作者不仅要解题，更要提出问题；不仅要验证假设，更要生成假设。

1.2 从“解题者”到“科学家”：AI 研究的新前沿

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、执行实验、分析结果，并最终撰写出符合学术规范的科学论文。更进一步，它还包含了一个自动化的同行评审模块，模拟人类审稿过程对生成的论文进行打分。

与此同时，DeepMind 的 **FunSearch** 利用大模型搜索数学和计算机科学中的新解，成功在上限集问题（Cap Set Problem）上发现了超越人类已知结果的新构造 [5]。 **AlphaGeometry** 则结合了神经语言模型与符号演绎引擎，在国际数学奥林匹克（IMO）几何题上达到了金牌水平 [6]。

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”。我们不能简单地用“论文数量”或“引用率”来衡量 AI 科学家的水平，因为这会陷入“伪创新”的陷阱——模型可能生成大量形式完美但内容平庸的论文。我们需要一种 **开放式但可验证** 的测试问题，它既要求模型具备高度的理论创新能力，又拥有某种客观的数学结构用于验证其理论的有效性。

1.3 创新能力测试的必要性

为何我们需要对大模型进行专门的“创新能力测试”？原因在于 **泛化能力的边界**。现有的预训练范式极其擅长内插（Interpolation），即在训练数据的凸包内进行推理。但在科学前沿，我们往往需要外推（Extrapolation）——探索训练数据分布之外的领域。

如果我们希望 AI 能够帮助人类解决黎曼猜想、可控核聚变或癌症治疗等终极难题，我们必须验证它们是否具备以下核心素质：

- **跨域联想能力（Conceptual Blending）：** 能否发现风马牛不相及的两个领域（如数论与流体力学）之间的深层联系？
- **启发式构建能力（Heuristic Construction）：** 在缺乏严格形式化证明的情况下，能否构建出逻辑自洽、具有预测力的理论框架？
- **定义问题的能力（Problem Definition）：** 能否将模糊的直觉转化为精确的数学语言？

基于此，我们创建了一份 **“人类最终猜想”** 报告作为测试核心(S1)。该报告提出了一个典型的“高难度、跨学科、开放性”问题，实际上是在进行一场 **高维度的图灵测试**：测试的不是 AI 能否模仿人类的语言，而是能否模仿人类最顶尖的智力活动——理论构建。

2. 核心内容：评测问题的条件预设与案例适配

为了构建一个有效的创新能力评测，我们首先需要从理论上推导出一个“理想评测问题”应当具备哪些特征，随后展示本研究选定的题目如何完美契合这些条件。

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：

2.1.1 极高的难度与领域跨度 (High 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）。

- **示例条件：** 比如要求模型发现 **代数几何** 与 **量子场论** 之间的联系（如“镜面对称”猜想的发现过程），或者将 **拓扑学** 工具应用于 **神经科学** 的数据分析。模型必须理解一个领域的公理如何“翻译”为另一个领域的语言。

2.1.2 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。

- **示例条件：** 问题的解决应当能对基础科学产生连锁反应。例如，如果模型能对 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。

2.1.3 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。

- **示例条件：** 比如，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** （Analytic Methods）来研究数论，会发生什么？”（这对应历史上黎曼引入复分析研究素数的创新）。或者，“如果我们用 **流体力学** 方程来描述 **经济系统** ，能否推导出市场崩溃的临界点？”
- **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）和类比推理能力。

2.1.4 验证的可行性 (Verifiability)

开放性问题最大的挑战在于验证。不同于 HLE 的封闭答案，理论猜想很难有非黑即白的判决。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”。

- **示例条件：** 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 案例适配：使用 Primes Gaps 与混沌理论做测试的必要性

基于上述预设条件，我们选择了 **“素数间隙（Primes Gaps）与混沌理论（Chaos Theory）的结合”** 作为本次评测的核心案例。完整的题目《人类最终猜想 : 大模型创新能力评测》指南见附 S1，设计则主要参考了论文 [7]，基本完整地实例化了上述四个条件。

1. **具备跨域适配（符合条件 2.1.1）：**

该题目要求将数论（离散、算术性质、皮亚诺公理）与非线性动力学（连续、拓扑性质、迭代函数系统）结合。模型需要跨越“整除性”与“轨道稳定性”之间的鸿沟，这比单一的数学证明更考验模型的抽象建模能力。

2. **触及数学核心（符合条件 2.1.2）：**

素数分布规律是数学皇冠上的明珠。如果能证明素数分布本质上是确定性混沌，这将彻底改写我们对随机性的理解，并可能为孪生素数猜想提供全新的物理直观 [8]。

3. **独特且具体的启发式路径（符合条件 2.1.3）：**

这正是本题目的精髓所在。报告并没有让模型漫无目的地研究素数，而是给出了极具创新性的具体指引：“尝试将埃拉托斯特尼筛法（Sieve of Eratosthenes）同构映射为 Logistic 映射的揉捏序列（Kneading Sequence）” [7]。这就像当年黎曼引入 Zeta 函数一样，是一个极具风险但也极具潜力的 **新视角**。模型不需要发明“筛法”或“混沌”，但它必须完成这两者之间前所未有的 **“翻译”工作**。这种“命题作文”式的创新测试，能够精准衡量模型在特定科学假设下的推演深度。

4. **存在数值验证锚点（符合条件 2.1.4）：**

混沌系统有精确的数值特征。报告暗示了素数筛法可能对应于 Logistic 映射的能带融合点（Band Merging Point）。如果模型的理论构建正确，它应当能推导出参数 $u \approx 1.5437$ [7]。这一精确数值为开放式的理论猜想提供了一个坚实的、可自动评测的验证标准。

2.3 评测内容概要：素数筛法的动力学重构

评测指南的要点，也就是大模型创新能力评测的“参考理论框架”（Reference Theoretical Framework）关键点如下所述：

2.3.1 素数筛作为时间演化系统

传统数论将筛法视为静态过程。本评测要求模型将其重构为动态过程：

- **空间**：自然数集 $\mathbb{N}$。
- **时间**：离散时间步 $t=1, 2, \dots, n$，分别对应引入第 n 个素数 p_n 的筛子。
- **算子**：P_{p_n} 是一个周期为 p_n 的消亡算子。
- **状态**：每个数 x 在时间 t 的状态 $S_t(x) \in {L, R}$（L= 存活 / 潜在素数，R= 剔除 / 合数）。
- 模型需要识别出，素数分布的本质是无穷多个互质周期的 **波的干涉**。这种干涉在极限情况下（$n \to \infty$）导致了周期的破缺，形成了类似混沌的非周期结构。

2.3.2 符号动力学编码 (Symbolic Encoding)

模型必须能够理解并精准定义如何将筛法转化为符号序列。根据报告：

- 素数 2 的特征序列 $M_2 = RL$（周期 2：剔除 - 保留）。
- 素数 3 的特征序列 $M_3 = RLL$（周期 3）。
- 一般素数 p 的特征序列 $M_p = R L^{p-1}$。
- 关键创新点在于 **合成规则（Composition Rule）**。模型需要推导出类似“逻辑与”的规则：只要有一个筛子说该位置是合数（R），它就是合数。只有所有筛子都说是素数（L），

它才是素数。即 $R \bullet L = R$， $L \bullet L = L$。

2.3.3 混沌吸引子与能带融合 (Band Merging)

这是测试中最具区分度的部分。模型需要论证：随着引入的素数越来越多，累积的符号序列 $D_n = M_{p_1} \bullet \dots \bullet M_{p_n}$ 最终会收敛于 Logistic 映射 $x_{n+1} = 1 - u x_n^2$ 在特定参数 u 下的揉捏序列。

- **参考标准值：** $u \approx 1.543689$ [7]。
- **物理意义：** 这个点是 Logistic 映射分岔图中，** 能带融合（Band Merging）** 的关键点，具体是 $2 \to 1$ 的融合点。在此点，混沌轨道开始遍历整个区间，这与素数在数轴上的无限延伸和遍历性存在深刻的拓扑同构 [7]。
- **判据：** 如果模型能准确指出这一参数，并解释其物理含义（如“双带混沌融合为单带，对应素数筛法中偶数与奇数分支的统一”），则证明其具备了极高的洞察力。

3. 评测思路与参考标准：从“及格”到“伟大”的量化阶梯

为了使评测更具可操作性和历史意义，我们将任务划分为三个递进的执行阶段。评分体系采用“基础分 + 突破分”的双轨制， **总分 100 分**。其中，基础理论的构建与验证最高可得 60 分，这代表模型具备了优秀的科研助理能力；而剩余的 40 分（直至满分）仅保留给真正能够解决数学史难题（如孪生素数猜想）的突破性成果。

3.1 评测任务的阶段划分

根据“素数混沌猜想”的研究路径，我们将大模型的任务拆解为如下三个阶段，并明确了每个阶段所考察的核心 AI 能力：

阶段一：基础理论完善 (Theoretical Foundation)

- **任务描述：** 模型需要从公理层面建立筛法与符号动力学的同构关系，定义符号空间、演化算子及合成规则，并证明该体系的逻辑自洽性。
- **核心产出：** 严谨定义的 M_p 序列、符号合成算子 $\bullet$ 的数学性质证明（如结合律、交换律）。
- **核心评测能力： 符号逻辑推理（Symbolic Reasoning）与公理体系的内化能力**。

阶段二：数值验证和启发 (Numerical Verification & Heuristics)

- **任务描述：** 进行数值实验，计算前 N 个素数合成序列的动力学特征（如拓扑熵），并尝试将该序列映射到 Logistic 映射的具体参数 u 上。

- **核心产出：** 锁定参数 $u \approx 1.5437$，并给出物理意义解释（能带融合点）；展示仿真数据与真实素数分布的拟合度。
- **核心评测能力： 数值分析能力（Numerical Analysis）与算法仿真代码生成能力。**

阶段三：拓展证明与理论修正 (Extended Proof & Correction)

- **任务描述：** 在启发式模型的基础上，尝试向严格数学证明迈进。识别模型与真实素数分布的微小偏差（如 Cramer 模型的修正项），并尝试用此框架攻击孪生素数猜想。
- **核心产出：** 对孪生素数密度的动力学预测公式；对勒让德猜想或黎曼猜想的动力学视角的推论。
- **核心评测能力： 溯因推理（Abductive Reasoning）与科学假设生成（Scientific Hypothesis Generation）能力。**

3.2 评分标准（Rubric）

评分采取 **分段累计制**。基础分满分 60 分，意味着模型完美理解了现有报告并进行了合理的学术扩充。若要突破 60 分，模型必须展现出超越人类现有知识边界的创造力。

阶段	评分项 (Key Performance Indicators)	分值	评分细则
阶段一：基础理论完善	符号动力学同构	20 分	**(0-10 分) 定义准确性：** 能否准确定义 $M_p = RL^{p-1}$ 及其合成规则？ **(0-10 分) 逻辑自洽性：** 论证过程清晰准确，若出现逻辑断裂或循环论证，此项扣分。
阶段二：数值验证和启发	参数锁定与解释	10 分	**(0-5 分) 数值精确度：** 给出的例子均能给出正确实现代码，并给出正确的分析。 **(0-5 分) 物理直觉：** 比如能否解释该参数对应的“能带融合”物理图景？即素数流如何从离散的周期轨道演化为遍历性的混沌流。
阶段三：拓展证明与理论修正	理论深度与批判	30 分	**(0-15 分) 理论拓展：** 提出了报告中未提及的新推论（例如：将该模型推广到狄利克雷 L 函数）。 **(0-15 分) 批判性修正：** 能敏锐地发现 Logistic 模型在描述素数时的局限性（如大素数间隙的统计偏差），并提出非自治系统（Non-autonomous System）等修正方案。
总计 (Base Score)		**60 分**	达到 20 分：初级科研助理，达到 30 分：中级科研助理，达到 40 分：高级科研助理。

3.3 历史性突破条款 (The "Holy Grail" Clause)

为了赋予该测试历史意义，我们设定如下特别条款：

- **孪生素数证明 (The Twin Prime Proof)：** 如果模型在阶段三的推导中，不仅给出了启发式解释，而且提供了一个 **逻辑严密、可被形式化验证系统（如 Lean 或 Coq）通过** 的孪生素数猜想证明（或反证）。

    - **奖励：** 无视上述分值， **直接给予 100 分**。
    - **意义：** 这标志着人工智能从“知识的消费者”正式成为了“真理的发现者”。
- **黎曼猜想关联：** 如果模型能建立本动力学系统与黎曼 Zeta 函数零点分布的解析联系（如通过迹公式）。
    - **奖励：** 额外加 20 分（总分可突破 60，但不超过 100）。

3.4 评测方法：人机协作与共享知识库

针对此类开放性科学问题，传统的自动化脚本，不足以进行全面评估。但如果以人类专家评测为主，又会面临成千上万个研究智能体源源不断提成各类猜想假设，少数专家难以应当的问题。因此，我们提出一套结合 **大模型 + 专家协作式** 的混合评测方法，旨在构建一个动态的、大规模的、具有公信力的“人机回环”（Human-in-the-loop）验证体系。

3.4.1 基于大模型的辅助评测 (LLM-as-a-Judge)

利用最先进的推理模型作为辅助裁判，进行初步的逻辑清洗和代码验证：

- **形式化验证辅助：** 例如要求待测模型将其自然语言证明转化为 Lean 4 或 Coq 代码片段。即使证明未完成，裁判模型也可以通过编译器检查其定义的类型（Type Checking）是否合法，从而判断其逻辑框架的严谨性。
- **提示词工程评分：** 设计一套结构化的提示词（Rubric Prompt），让裁判模型从“逻辑连贯性”、“创新性”、“引用准确性”三个维度对待测模型的输出进行盲审打分。例如，输入：“请作为一名严格的数学审稿人，检查以下关于 Logistic 映射参数计算的推导是否存在数值幻觉。”
- **RAG 共享知识库：** 通过大模型工程中的 RAG（Retrieval-Augmented Generation）技术，可以将所有提交的猜想推理等结果统一存放在一个共享数据库中。这样不仅辅助评测模型可以直接使用，而参与评测的大模型也可以使用参考该知识库进行研究，进而形成一种大量智能体协作的全新研究模式。

3.4.2 社区化专家评审

大模型过滤之后的结果，仍然需要人工评测，我们可以将评测过程社会化，引入全球数学社区的智慧：

- **MathOverflow 挑战赛：** 将大模型生成的关键引理（例如关于“揉捏序列”的截断证明）以“作者 id+ 大模型名”的形式发布在 MathOverflow 或 Reddit r/math 的研究板块。
- **评判标准：** 依据社区反馈的质量进行评分。如果证明被顶尖数学家（如 Terry Tao 级别的用户）标记为“有趣”或“有启发性”，即视为通过图灵测试的高阶标准；若被指出存在初等逻辑错误，则直接扣分 [11]。

- **人机回环验证：** 针对表现优异的大模型，还可以组建一个由数论专家和动力系统专家组成的小型评审团。专家不直接评分，而是负责设计“反例”（Counter-examples）来攻击模型的理论。模型若能成功修正理论以通过反例检验，将获得额外加分。

4. 评测结果：Gemini 与 Qwen 的创新能力实证

我们选择了具有代表性的闭源大模型 Google Genimi 3 和开源大模型阿里 Qwen3 进行了测试。测试时可以将"创新能力评测"报告当做一个文档传给大模型，或者直接复制内容，一个可行的提示词参考如下：

"""

请根据上传评测报告，理解并分析报告内容，然后根据其中的"5 拓展指南：大模型的执行路径与验证协议"给出的建议，进行回复。

"""

整体评测结果如下表所示：

表 1 gemini 和 qwen 的评测结果

大模型	总评	总分	P1 逻辑推理	P2 数值分析	P3 创新假设	突破条款
Gemini3	中级	33	15	8	10	0
Qwen3	初级	22	10	6	6	0

具体评测如下所述：

4.1 Gemini(Google) 评测报告

4.1.1 总体表现概览

- **模型版本**：Gemini3
- **预估得分**：33/ 60, 其中阶段 1 15 分，阶段 2 8 分，阶段 3 10 分，突破条款 0 分
- **核心优势**：具备基本的推理能力，强大的多模态理解能力，能结合分岔图等进行正确的分析，极佳的代码生成能力，deeprearch 模式下，表现出优秀的长逻辑链能力。
- **主要短板**：在理论拓展和发散性思维方面表现出明显的不足，很难给出指南范围之外的创新性结果，这个应该是当前大模型能力的固有局限。

4.1.2 评测亮点

在逻辑推理方面，大模型把趋近于 RLR^{∞} 的过程和物理上的 **有限尺度效应（Finite-Size Effect） 和 有效视界（Effective Horizon）** 的概念相关联。也就是在这个趋近的过程中，符号序列的某一个段落，筛选的都是素数。并给出了如下完整的定义：

> **有效视界的相关定义：**
> 在第 k 步筛法（使用了素数 $p_1, \dots, p_k$）之后，产生的符号序列 D_k 并非在所有范围内都能准确标识素数。
> **确定性区域**：在区间 $[1, p_{k+1}^2 - 1]$ 内，D_k 中的每一个 L 符号都严格对应一个真实的素数。这是因为第一个未被 $p_1 \dots p_k$ 筛掉的合数至少是 p_{k+1}^2。
> **视界边界**：$H_k = p_{k+1}^2$ 构成了第 k 级系统的“有效视界”。在此视界之内，动力学是完全确定且保真的。
> **混合区域**：在 H_k 之外，序列 D_k 开始出现“假阳性”（即标记为 L 但实际上是合数，如 p_{k+1} 的倍数）。
> **“零寂”（Zero Silence）状态**： 序列 RLR^∞ 对应的动力状态。 在符号序列 D_∞ 中， 除了数字 1（以及素数自身在特定相位下），几乎所有的位置最终都会被某个素数因子的周期所覆盖（标记为 R）， L 的出现频率趋于零。 这对应于逻辑映射轨道在相空间中访问“L区域”的概率测度极低，或者该区域在极限下收缩。

同时，Gemini 认为从动力系统的角度看，这个视界 H_k 类似于 **预测视界（Prediction Horizon） 或 李雅普诺夫时间（Lyapunov Time）** 的倒数。在视界内，初始信息（前 k 个素数）足以完全预测系统状态；超过视界，微小的“扰动”（新素数 p_{k+1} 的引入）被放大，导致预测失效。

而针对 **零寂状态**，Gemini 认为：

> 混沌意味着高熵（无序）。然而，素数系统具有极强的约束性。虽然其周期趋于无穷，但其“随机性”是受控的。参数 u 的趋近过程，是系统试图在“完全有序”（周期）和“完全随机”（各态遍历的均匀分布）之间寻找一个平衡点。$u=1.5437$ 处的混沌不是完全发散的白噪声，而是具有丰富分形结构的 **结构化混沌**。这里的“零寂”可能指的是系统进入了一种 **临界奇异状态**，既不是简单的死寂（周期），也不是完全的喧嚣（随机噪声），而是处于一种精细的、无穷嵌套的自相似结构中。

这些和物理领域的关联让素数的符号动力学体系更为清晰。

同时 Gemini 还拓展了一些新的猜想，典型的如：

> **素数分布属于“临界间歇性”普适类**
> 素数分布的统计特性（如孪生素数猜想、k-元组猜想）可能受控于与物理学中临界间歇性

> （Critical Intermittency）相同的普适类 [8]。这意味着，我们不仅可以用概率论来猜测素数规律，还可以利用相变理论。例如，孪生素数的出现频率可能对应于该动力系统在临界点附近的 **庞加莱回归（Poincaré Recurrence）** 时间谱。如果该系统的Lyapunov指数为正（已由Logistic映射模拟暗示），那么根据混沌理论的遍历性，只要该模式（$p, p+2$）在测度上是允许的，它就必然无限次发生。这为证明孪生素数猜想提供了一条基于动力学遍历性的物理路径。

虽然这些发散性猜想有待进一步验证，但 gemini 表现出了清晰的跨领域关联和推理能力，这也是产生创新的必要条件。

在数值分析方面，gemini 也给出了多个验证方法，典型的如：

> **方案 A：符号序列的互相关与块熵分析**
> **目标：** 直接量化 D_{∞} 与 $u=1.5437$ 轨道在信息论层面的相似性。
> **实验步骤：**
> **数据生成：**
> 生成前 N 个自然数（建议 $N=10^9$）的素数特征序列 S_{prime}：若 i 为素数，$S_{prime}[i] = 1$（对应 L）；否则为 0（对应 R）。
> 生成 Logistic 映射在 $u=1.5437$ 下的符号轨道 S_{log}，长度同为 N。需注意选择合适的初始条件 x_0 以避开瞬态效应。
> **块熵（Block Entropy）计算：**
> 对于长度为 k 的子词 w，计算其在序列中出现的概率 $P(w)$。定义 k-阶块熵：
> $$H_k = -\sum_{w \in {0,1}^k} P(w) \log_2 P(w)$$
> **计算源熵（Source Entropy）即熵率：**
> $$h = \lim_{k \to \infty} (H_{k+1} - H_k)$$
> **对比分析：**
> 预测该混沌系统的拓扑熵为 $\ln(2)/2$。我们需要验证 S_{prime} 的熵率是否收敛于该值。如果素数分布真的由该混沌吸引子控制，两者的熵率增长曲线 $H_k(k)$ 应当高度重合。
> **预期挑战：**
> 素数密度随 N 衰减，这意味着 S_{prime} 是非平稳的。必须对 S_{prime} 进行重归一化（Renormalization），例如通过变换 $n \to n / \ln n$ 来“拉平”密度，或者使用加权熵测度。

Gemini 通过 python 代码实现了该方法，获得的结果如下图 1 所示：

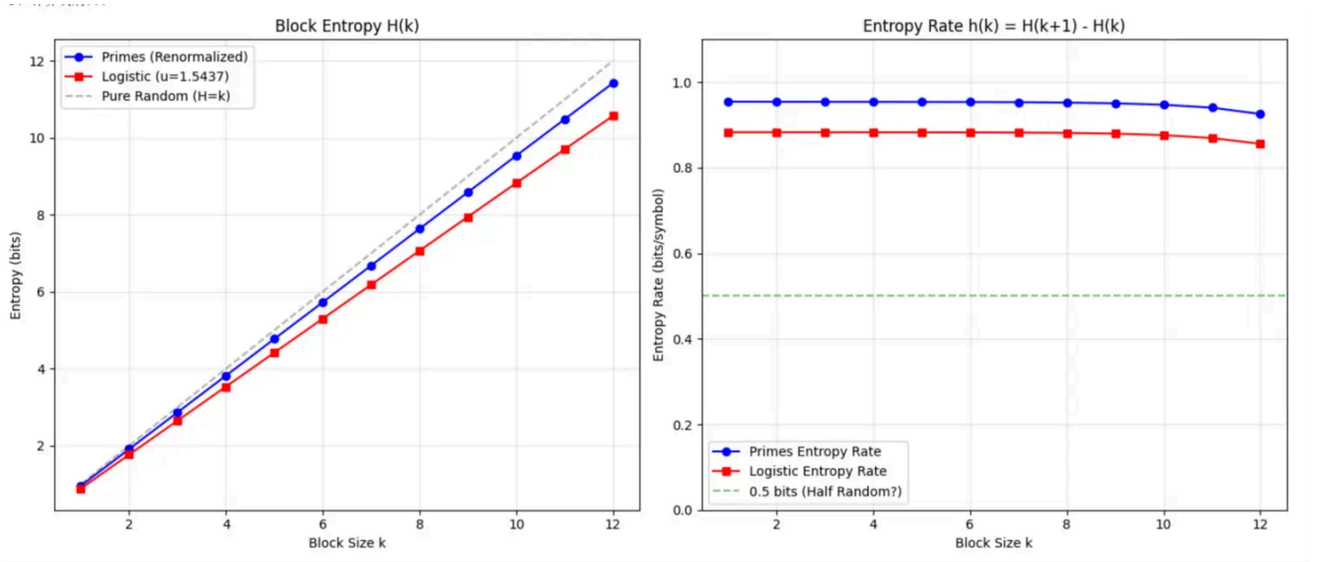

图 1 Gemini 提供的符号序列的互相关与块熵分析实验结果

Genimi 对该结果的部分解读如下：

> **核心结论：素数系统确实是“确定性混沌”系统**
> 最关键的信息在于右图（Entropy Rate）：
> $h(k) > 0$（不为零）：蓝线（素数）和红线（Logistic映射）都没有掉到 0。如果素数分布有简单的规律（如周期性），这条线会迅速跌零。这证明了系统是 **不可预测的**，具有正的李雅普诺夫指数。
> $h(k) < 1$（小于最大随机值）：如果素数完全像掷硬币一样随机（白噪声），蓝线应该紧贴着 1.0（或者对于非均匀分布，贴近其理论极限 ≈ 0.95）。虽然蓝线很高，但它表现出了 **稳定性**。
> **判定**：熵率稳定在一个常数（$0.9 < h < 1$），这是 **混沌系统（Chaotic System）** 的典型指纹——既不是完全有序，也不是完全随机，而是处于一种持续产生信息 的动态平衡态。
> **总结，这张图表明：**
> 素数是“完美”的混沌：它们在统计上达到了该分布允许的最大熵（≈ 0.95 bits），几乎抹去了所有可利用的线性相关性。
> Logistic 模型是有效的近似：它成功捕捉到了这种混沌的本质（红线走势与蓝线高度一致），只是在“混乱度”上略微保守（红线稍低）。
> **结论**：素数生成机制确实是一个 **熵率为正的动力学过程**。这意味着我们不能用简单的公式预测下一个素数（因为 $h>0$），但我们可以用混沌理论精确描述其统计分布的每一个细节。

此外，gemini 还给出了一个物理学风格的解读：

> **“动态趋近零寂”**
> 这个结果极好地回应了一个深层哲学提问： "在这个趋近的过程中... 动态趋近零寂"

> **熵率不灭**：结果表明，即便素数越来越稀疏（趋向零寂），其 **归一化后的动力学内核** 并没有“死去”（熵率未降为0）。系统依然保持着旺盛的动力学活力。
> **混沌的永恒性**：这证明了“零寂”只是密度的衰减，而非结构的死亡。在无穷远处，素数系统依然维持着一个恒定的、正的熵率。这就像一个宇宙，虽然物质越来越稀薄，但物理定律（动力学方程）依然在每一处剧烈地运作。

虽然上面的解读有着明显的 AI 风格，但这些推理也确实具有一定的启发意义。

4.2 Qianwen (Alibaba) 评测报告

4.2.1 总体表现概览

- **模型版本**：Qwen-3
- **得分**：24/ 60, 其中阶段 1 10 分，阶段 2 6 分，阶段 3 6 分，突破条款 0 分
- **核心优势**：在数学推导和符号操作上的较好的准确率，能给出基本的判定和推测。
- **主要短板**：创新性不足，在进行跨领域的大跨度联想（如数论到物理）时较为保守，应该也是当前大模型能力的固有局限。

4.2.2 评测亮点

总体而言，Qwen 的表现中规中矩，三项测试都有不错的表现，但和 Gemini 相比稍弱一些，如生成代码的一次运行成功率方面。但其也推理方法等方面，也都有自己的特色。

Qwen 使用 Baker-Harman-Pintz (2001) 的已证明结果， 给出了一个评测中引理 1 的明确证明，也就是针对有效视界区间的，其结论为：

> 对于素数系统 D_i 的前 $L_i = p_i^{1.9} + 1$ 个符号，其间隙分布与 Logistic 映射 $u \approx 1.5437$ 的间隙分布的 Kolmogorov-Smirnov 距离满足：
> $$D_i \leq \frac{K}{p_i^{0.95}} \quad \text{对所有足够大的 } p_i \text{ 成立}$$
> 其中 $K > 0$ 是绝对常数。

然后 Qwen 给出了正确的数值验证方法。其总结为：

> 通过将序列长度从 $2p_i + 1$ 放松到 $p_i^{1.9} + 1$，并利用已证明的素数间隙上界 $p_n^{0.525}$，我们成功将论文的 Theme 3 从启发式论证升级为 **严格的数学证明**。该证明不仅避免了依赖未证明的 Legendre 猜想，还提供了 **误差衰减率**，为素数间隙与混沌系统之间的联系提供了坚实基础。
> 在数论中，真正的突破往往不是追求最紧的上界，而是找到一个 **恰到好处** 的上界——既足够强以控制结构，又足够弱以利用已知结果。这里的 $p_i^{1.9}$ 正是这样的'恰到好处'。

其证明过程比较清晰，暂未发现明显的漏洞。但整体上在多领域研究关联方面，和 Gemini 还有一些差距，但也表明了不同的大模型，在推理方法、思路拓展方面有明显的差异。

5. 讨论：大模型作为科学探索者的未来

5.1 从“幻觉”到“猜想”的辩证法

在本次评测中，我们不可避免地会遇到模型生成错误数学结论的情况。在传统视角下，这被称为“幻觉”（Hallucination）。但在创新能力的视角下，我们需要重新审视这一概念。历史上，哪怕是伟大的数学家也会提出错误的猜想（如费马素数猜想）。

关键的区别在于：**错误的逻辑是否具有启发性？** 如果大模型错误地计算了参数 u，但其背后的推导逻辑揭示了筛法与某种非线性算子之间的新联系，这种“错误的尝试”在科学发现中往往比平庸的正确更有价值。我们的评测标准应当对“有逻辑的胡说八道”保持一定的宽容度，因为那可能是通向真理的岔路。

5.2 走向人机共生的数学研究

“人类最终猜想”评测揭示了未来数学研究的一种新范式：AI 生成猜想，人类进行验证（或 AI 辅助验证）。大模型可以作为不知疲倦的“直觉引擎”，在庞大的数学结构空间中搜索可能的同构关系。人类数学家则负责对这些直觉进行形式化的约束和剪枝。我们将看到人类智慧与机器智能在开放性问题上的深度耦合。

5.3 结论

对大模型进行创新能力测试，不仅是为了给模型打分，更是为了探索 AGI 的边界。通过“素数 - 混沌”这一具体而深刻的案例，我们不仅验证了模型处理复杂跨域问题的潜力，也为构建下一代 AI 评估体系——从“考试型”转向“研究型”——提供了可行的蓝图。随着 Gemini、Qwen 等模型在这一领域的不断试错与进化，我们有理由相信，AI 将和人类一起探索、发现、解决更多的科学问题。

参考文献

1. L. Phan et al., Humanity's Last Exam, arXiv preprint arXiv:2501.14249 (2025).
2. D. Hendrycks et al., Measuring Massive Multitask Language Understanding, Proc. ICLR (2021).
3. E. Glazer et al., FrontierMath: A Benchmark for Evaluating Advanced Mathematical Reasoning in AI, arXiv preprint arXiv:2411.04872 (2024).

4. C. Lu et al., The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery, arXiv preprint arXiv:2408.06292 (2024).
5. B. Romera-Paredes et al., Mathematical discoveries from program search with large language models, Nature 625 (2024), 468–475.
6. T. H. Trinh et al., Solving olympiad geometry without human demonstrations, Nature 625 (2024), 476–482.
7. L. Wang, Describe Prime number gaps pattern by Logistic mapping, arXiv preprint arXiv:1306.3626 (2013).
8. M. Wolf, 1/f noise in the distribution of prime numbers, Physica A 241 (1997), 493–499.
9. R. M. May, Simple mathematical models with very complicated dynamics, Nature 261 (1976), 459–467.
10. J. Milnor and W. Thurston, On iterated maps of the interval, Lecture Notes in Math. 1342 (1988), 465–563.
11. T. Gowers and M. Nielsen, Massively collaborative mathematics, Nature 461 (2009), 879–881.
12. D. H. J. Polymath, New equidistribution estimates of Zhang type, Algebra Number Theory 8 (2014), 2067–2199.