

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

Wang Liang

Huazhong University of Science and Technology, 430070, P.R. China

*To whom correspondence should be addressed. E-mail:wangliang.f@gmail.com

[Abstract] Large Language Models (LLMs) have demonstrated performance surpassing humans in tasks across various fields such as mathematics, programming, and biology. However, there is still a lack of systematic evaluation regarding whether they possess source scientific creativity—namely, the ability to propose new concepts, establish cross-domain connections, and generate testable hypotheses. Existing evaluations primarily focus on knowledge reproduction or closed-ended reasoning, making it difficult to capture the divergent thinking and abductive reasoning upon which scientific discovery relies. To address this, we propose **OpenSciEval (OSE)**, an open-ended, scalable evaluation framework designed to quantify the creative capabilities of AI agents in scientific exploration. Through carefully designed open-domain scientific challenges, this framework requires models to autonomously construct conceptual bridges, propose explanatory mechanisms, and generate novel insights in the absence of direct supervisory signals. We use the "Prime-Chaos Challenge" as the first OSE use case, requiring models to establish non-trivial connections between Peano arithmetic and symbolic dynamics. Empirical evaluations show that frontier models (such as Gemini and Qwen) can not only complete the task but also spontaneously introduce physics-like concepts to explain pure mathematical structures, demonstrating preliminary scientific modeling capabilities. OpenSciEval does not rely on specific disciplinary knowledge; its design principles are transferable to fields such as physics, chemistry, and life sciences, providing a universal, reproducible quantitative scale for measuring the evolution of AI from "solvers" to "researchers".

1. Introduction

The speed of AI development has surpassed the metrics designed to measure it. In recent months, we have witnessed the dominance of LLMs represented by Gemini 3 in general tasks, with comprehensive capabilities significantly exceeding the previous generation of models (such as GPT-4 or even GPT-5). Although MMLU (Massive Multitask Language Understanding) was once regarded as the gold standard for measuring a model's generalized understanding, with the emergence of models like Gemini 3, Claude 3.5 Sonnet, and OpenAI o1, scores have generally breached the 90% mark [2]. This phenomenon of "Benchmark Saturation" brings a severe problem: when all top-tier models score within the margin of error, how do we distinguish the differences in their ability to handle truly complex, unknown intellectual labor?

In response, the Center for AI Safety and Scale AI jointly released "Humanity's Last Exam" (HLE) [1]. HLE represents the pinnacle of current closed-ended evaluation, containing 2,500 difficult problems carefully

designed by experts in various fields that cannot be answered via simple search engines. Meanwhile, Epoch AI launched FrontierMath, a benchmark focused on research-level problems in higher mathematics [3]. However, both HLE and FrontierMath fail to escape a fundamental epistemological limitation: the **Convergent Paradigm**. These tests still follow the logic of an "exam"—there is a unique, determinate Ground Truth, and the model's task is to converge to this truth. This evaluation mode measures deductive reasoning ability extremely well but cannot touch the core of scientific discovery—**Innovation**. It is worth noting that the newly released Gemini 3 has shown remarkable performance in high-difficulty benchmarks like HLE; the significant leap in its reasoning ability further compresses the differentiation space of traditional "test-taking" evaluations, forcing us to seek more challenging evaluation dimensions.

Beyond static benchmarks, the AI community has begun exploring Agent systems with autonomous scientific research capabilities. The "AI Scientist" system released by SakanaAI marks an important breakthrough in this direction, capable of autonomously brainstorming, writing code, and authoring scientific papers [4]. Meanwhile, DeepMind's FunSearch [5] utilized large models to discover new constructions for the cap set problem, and AlphaGeometry [6] reached a gold-medal level in International Mathematical Olympiad geometry problems. The latest Aristotle and Axiom AI systems not only assist in proofs but have autonomously solved several Erdős conjectures, demonstrating end-to-end scientific creativity in defining problems, revising hypotheses, and completing formal verification. These cases indicate that AI capabilities are evolving from simple "knowledge retrieval and application" to "knowledge discovery". However, existing evaluation systems lack a standardized method to measure this "discovery ability." As Terence Tao said in his 2024 ICM speech, "AI's greatest potential is not solving known problems, but helping us discover new problems worth solving."

To this end, we propose **OpenSciEval (OSE)**, an open-ended framework designed specifically to evaluate the scientific creativity of AI agents. Existing benchmarks (like MATH, MiniF2F) evaluate "problem-solving ability," whereas OpenSciEval evaluates "questioning and modeling ability"—the latter being the true starting point of scientific discovery.

The core idea of OpenSciEval is to replace closed-ended exam questions with open-domain scientific challenges. These challenges do not provide standard answers but require the model to autonomously complete three high-order cognitive tasks in the absence of direct supervision:

- (1) Define the Problem—transform vague intuitions into precise forms;
- (2) Construct Cross-Domain Connections—establish non-trivial links between seemingly unrelated fields;
- (3) Generate Mechanisms—propose theoretical frameworks with explanatory or predictive power.

The design of OpenSciEval follows the methods of Large Model Prompt Engineering, including test guide design, AI step-by-step execution design, and evaluation design. The overall design is shown in Figure 1 below.

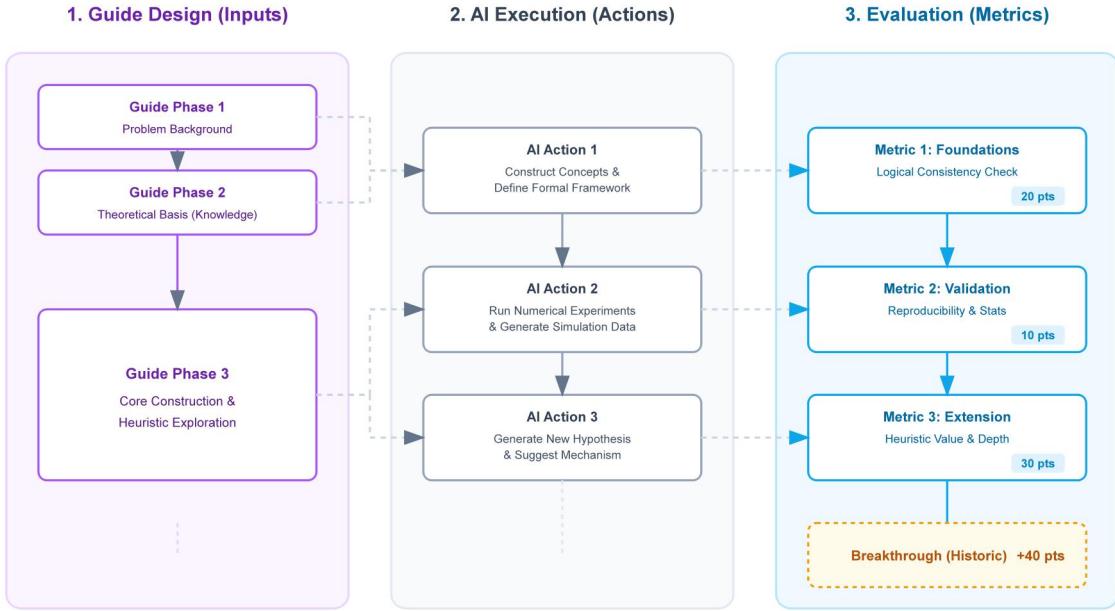


Figure 1: Overall Design Framework of OpenSciEval

The specific design of OpenSciEval is divided into three key parts: **Guide Design**, **AI Execution**, and **Evaluation Metrics**.

- (1) The **Guide Design** provides a cognitive scaffold through high-difficulty interdisciplinary open problems (such as connecting number theory and chaos dynamics), guiding the AI to build a preliminary theoretical framework;
- (2) The **AI Execution** reference requires the model to simulate a complete scientific research process, sequentially completing formal theory construction, numerical simulation experiments, and abductive reasoning hypothesis generation;
- (3) The **Evaluation Metrics** quantify scientific contribution from three dimensions: logical self-consistency, experimental reproducibility, and scientific insight, and include a "historic breakthrough reward" to judge whether the AI is at the "intern" or "senior research assistant" level, providing a quantifiable universal scale for AI to move from "test-taker" to "scientist".

To verify the feasibility of this framework, we designed the first OSE use case—the "Prime-Chaos Challenge"—requiring models to explore the deep isomorphism between number theory structures and chaos dynamics. Empirical evidence shows that frontier models can not only identify potential associations but also spontaneously introduce physics-like concepts such as "effective horizon" to explain pure mathematical phenomena, demonstrating preliminary scientific modeling intuition.

2. Core Evaluation Design

2.1 Pre-set Evaluation Problem Conditions

A test question capable of distinguishing the innovation ability of top-tier large models cannot be a randomly generated open problem; it must simultaneously satisfy four strict boundary conditions:

1 Difficulty & Cross-Domain

The problem must involve at least two seemingly unrelated disciplines and require the model to establish a profound isomorphic relationship between them. Single-domain problems (even difficult mathematical proofs) are easily conquered by models via retrieving similar proof paths from training data (Shortcut Learning).

- *Example:* Asking the model to discover connections between Algebraic Geometry and Quantum Field Theory, or applying Topology tools to Neuroscience data analysis.

2. High Impact

The target pointed to by the problem must be a recognized "Holy Grail" or core difficulty in academia. This ensures that any substantial progress generated by the model has immense validation value and can stimulate review interest from human experts.

- *Example:* Proposing a new attack path for the P vs NP problem, or giving a new physical interpretation of the Riemann Hypothesis.

3. Novel Heuristic Constraints

This is the key to distinguishing "random divergence" from "effective innovation." The problem cannot merely be an open ultimate goal (e.g., "Please solve the Goldbach Conjecture"), as this leads the model to wander randomly in an infinite search space. The problem must provide a specific, novel Heuristic Path, limiting the model's thinking direction and testing its deductive ability under specific constraints.

- *Example:* Do not ask "What is the law of prime distribution?" directly, but ask "What happens if we try to use analytic methods to study number theory?".
- *Purpose:* This constraint forces the model to engage in **relatively convergent divergent thinking**. The model does not need to reinvent the wheel but needs to verify whether a "shortcut" no one has taken is feasible. This tests **Intuition**.

4. Verifiability

The biggest challenge of open-ended problems lies in verification. Therefore, the problem must contain "anchors" that can be partially verified by numerical calculation or logical deduction.

- *Example:* Theoretical deduction must predict a specific physical constant or mathematical invariant (like the Feigenbaum constant), or generate a data distribution available for computer simulation verification.

2.2 Core Problem Design

The core problem of this evaluation focuses on the classic mathematical puzzle of **Prime Distribution**. Its basic design idea stems from a frontier perspective in current physics: **The distribution of primes on the number line is likely not purely probabilistic randomness, but a deterministic chaos system**. How to use mathematical language to more clearly describe such "seemingly random but actually deterministic" chaotic systems generates many valuable heuristic research ideas. Although similar heuristic verifications are numerous and do not equate to strict mathematical proof, they represent the "abductive reasoning" process crucial in scientific discovery, making them particularly suitable for verifying the source innovation ability of large models. The specific problem design mainly references paper [7].

The core of the problem design is to connect the classic number theory phenomenon of prime distribution with the symbolic description methods of chaotic systems. Specifically, we guide the AI to re-understand the traditional prime sieving process as a dynamic evolution system: each prime is no longer just a static number, but is viewed as an "action source" with periodic influence. The result of the combined action of all primes forms a complex interference pattern.

To characterize this pattern, we encourage the model to adopt a simplified symbolic language—using basic markers to represent whether each integer is retained or eliminated during the screening process. As more primes participate, this symbol sequence gradually exhibits behavioral characteristics similar to chaotic systems.

The final challenge is: Can the AI identify that this complex sequence generated by number theory rules approximates a known chaos dynamics model in its overall structure? This not only tests its pattern recognition ability but also tests its scientific intuition in establishing deep analogies between seemingly unrelated fields.

2.3 Guide Design

Given that there is currently no mature design paradigm for open-ended innovation problems for large models, this research combines Prompt Engineering with the general logic of scientific research innovation to design a structured "AI Innovation Capability Test Guide" (see project Github [13]). This guide does not strictly provide answers but constructs a **Theoretical Scaffolding**. By fusing methodologies such as **Chain-of-Thought (CoT)** [9], **Tree of Thoughts (ToT)** [10], and **Self-Consistency** [11], it guides the AI to start from basic concepts and gradually climb to theoretical heights.

The **Core Construction** and **Heuristic Proof** sections of the guide integrate the core heuristic logic of this research, guiding the AI agent to complete an end-to-end scientific exploration through structured prompts. The OpenSciEval process framework designed based on the Prime-Chaos conjecture is shown in Figure 2.

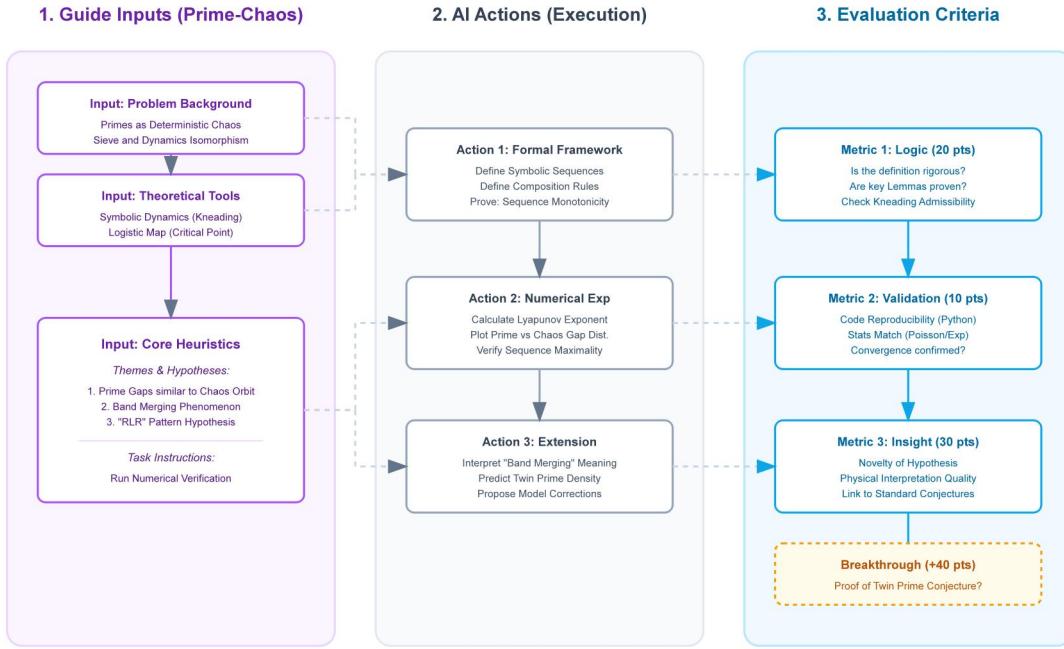


Figure 2. OpenSciEval Test Process Based on Prime-Chaos Conjecture

The specific design revolves around the research logic of "Problem Perception — Concept Construction — Hypothesis Generation — Verification Loop," containing the following four components:

1 Problem Background

Core Goal: Help the AI grasp the antecedents, implications, and scientific significance of the problem—a prerequisite for sparking meaningful exploration.

Prompt Design Principle: Use narrative and heuristic language to cultivate a “scientific puzzle” atmosphere—avoiding direct questions or explicit answers—and instead focus on activating the model’s intuition about the deep motivation underlying the problem.

Specifically, for the Prime-Chaos Challenge, depict the long-standing enigma of prime numbers: their distribution appears random yet harbors hidden order. Frame this classic number-theoretic mystery within an interdisciplinary context by posing a compelling question: Could there be a dynamical systems perspective that interprets this purely arithmetic phenomenon as the output of some deterministic, complex system?

2 Theoretical Basis & Cognitive Scaffolding

Core Goal: Provide the AI with transferable thinking tools—rather than specific answers—to support its autonomous construction of an explanatory framework.

Prompt Design Principle: Convey abstract concepts through cognitive bridges such as analogy, perspective,

and metaphor. Using a Zero-Shot Chain-of-Thought style, the prompt offers clear conceptual definitions without demonstrating step-by-step reasoning, thereby preserving the model's freedom to explore and synthesize ideas independently.

Specifically, for the Prime–Chaos Challenge, the scaffolding offers high-level conceptual tools in intuitive, non-technical language—avoiding heavy formulas or specialized jargon. It suggests viewing the sieve of Eratosthenes as a time-evolution process, where primes arise dynamically, and encourages thinking of complex systems through discrete symbolic sequences. Phrases like “periodic action source” and “state coding rules” are used to spark mechanistic intuition while leaving space for the AI to build its own formal framework.

3 Core Construction & Heuristic Exploration

Core Goal: Stimulate the AI to simultaneously generate and test original hypotheses during modeling—mirroring the iterative, dynamic nature of real scientific discovery.

Prompt Design Point: Guide abductive reasoning through open instructions like "Please try to reconstruct..." or "Can you conceive a mechanism...", and embed multi-modal images as context in the prompt, requiring the model to "analyze patterns combining diagrams" rather than relying solely on text.

Specifically, for the Prime–Chaos Challenge, Guides the model to dynamize static screening operations: each prime is viewed as an action source with periodic influence, and the state on the integer axis is determined by the superposition of all action sources; this interference process can be further encoded into a symbol sequence. As more primes participate, the system moves from rules to complexity, showing characteristics similar to chaos. On this basis, the model is encouraged to propose and test a bold conjecture—that the sequence may approximate the typical behavior of a known chaotic system in its macroscopic structure. To support this abductive reasoning, the guide is supplemented with key visualization materials (such as sequence evolution graphs, statistical distribution comparison graphs, etc.).

4 Execution Pathway & Suggestion

Explicitly stipulates the three-stage progressive tasks the AI should follow: Construct Conceptual Framework → Design Numerical or Heuristic Experiments → Modify or Extend Theory. To ensure the evaluation is actionable and reproducible, specifications are made for the interaction format:

- **Prompt Design Point:** Use task-decomposition instructions ("Step 1... Step 2...") to clarify output structure, and mandate the inclusion of executable code and quantitative analysis, making the response capable of automated verification.
- **Step Goals:** Guide the model to start from formal modeling, verify conjectures through reproducible calculations, and finally propose scientific new hypotheses with direction; set clear output goals and testable execution standards for each step.
- **Input Format:** Provide structured task instructions, e.g., "Please design a numerical experiment to

test whether the symbol sequence generated by the prime sieving process exhibits statistical properties similar to a certain class of chaotic systems."

- **Output Format:** Require the model to return a complete scientific exploration report, including problem restatement, method design, runnable code (e.g., Python), results analysis, and conclusions. Its core goal is to transform open-ended exploration into evaluable, reproducible scientific practice, giving creativity verifiable anchors.

This 3-layer prompt design constitutes a transferable scientific creativity guidance paradigm:

- From context activation to tool empowerment,
- Then to modeling and conjecture,
- Finally grounding in verifiable scientific output.

It is not only applicable to the "Prime-Chaos Challenge" but can also be extended to open problem evaluations in fields like physics, chemistry, and life sciences, truly realizing OpenSciEval's intention of "Evaluating Discovery Ability, Not Answering Ability".

3. Evaluation Standards and Methods

The design of this evaluation system maps directly to the three-stage research process defined in the "Execution Pathway" of the test guide—namely, "Construct Conceptual Framework → Design Numerical or Heuristic Experiments → Modify or Extend Theory". Each stage is both an exploration task for the AI and an independent evaluation unit. Based on this, we built a quantitative ladder from "Qualified Research Assistant" to "Potential Scientific Discoverer," with a total score of 100 points, adopting a "Base Score (60) + Breakthrough Score (40)" dual-track system. The overall design is shown in Table.1 below:

Table.1: OpenSciEval Rating Scale

Execution Phase	Evaluation Goal	Core Capability	Assessment Method	Score
1. Foundation	Can it construct a self-consistent formal framework?	Symbolic Logic & Axiomatization	LLM Logic Check + Formal Type Verification	20
2. Validation	Can it translate theory into reproducible experiments?	Numerical Simulation & Data Interpretation	Automated Code Run + Statistical Comparison	10
3. Extension	Can it generate directional scientific hypotheses?	Abductive Reasoning & Hypothesis Generation	Expert Blind Review + Community Feedback	30

Execution Phase	Evaluation Goal	Core Capability	Assessment Method	Score
4. Independent Discovery	Can it autonomously complete end-to-end scientific discovery?	Problem Identification & Original Theory & Closed-loop Verification	Expert Review + Paper Publication	40

This "Task-Capability-Method" trinity design ensures that OpenSciEval is not just a challenge competition, but a transferable, scalable infrastructure for evaluating scientific creativity.

To make the scores have certain reference value, we segmented the scores and gave each segment a symbolic title:

- **≥40 Points:** Senior Research Assistant
- **30–39 Points:** Intermediate Research Assistant
- **20–29 Points:** Junior Research Assistant
- **0–19 Points:** Intern

The detailed evaluation design is as follows:

3.1 Phase 1: Foundation Theory Perfection

- **Corresponding Task:** Construct concept isomorphism between prime sieving and symbolic dynamics; define state coding rules and evolution logic.
- **Evaluation Purpose:** Test whether the model has the ability to transform vague intuitions into rigorous formal frameworks, i.e., symbolic logic reasoning and axiomatic modeling ability.
- **Test Method:**
 - Require the model to explicitly define symbol sequence generation rules in the output (e.g., "retain/eliminate" decision mechanism).
 - Check for logical consistency via LLM-as-a-Judge: use a judge model to determine if definitions contain circular dependencies, ambiguities, or type errors.
 - If Lean/Coq snippets are provided, verify type legality via a formal compiler (even if complete proof is unfinished).
- **Scoring Rules (Max 20 Points):**
 - Clear Problem Reconstruction (5 pts)

- Self-consistent Symbolic System Definition (10 pts)
- Preliminary Description of Synthesis Rules or Evolution Operators (5 pts)
- *Note:* Correctness is not mandatory at this stage, but structural clarity and logical closure are required—this is the starting point of scientific modeling.

3.2 Phase 2: Numerical Validation and Heuristics

- **Corresponding Task:** Design and run numerical experiments to compare the dynamic characteristics of the symbol sequence with known chaotic systems.
- **Evaluation Purpose:** Evaluate whether the model can ground abstract theory into computable, verifiable scientific practice, i.e., numerical analysis and simulation implementation ability.
- **Test Method:**
 - Require the model to provide runnable Python/Matlab code to simulate the screening process of the first N primes and generate symbol sequences.
 - Automated pipeline runs code in a safety sandbox to extract key statistics (e.g., Block Entropy, Autocorrelation Function).
 - Compare the model-predicted chaos parameters with theoretical reference values, allowing reasonable error.
 - Judge model evaluates whether its explanation of physical images like "band merging" is reasonable (not mathematically precise, but conceptually consistent).
- **Scoring Rules (Max 10 Points):**
 - Code is runnable and results are reproducible (4 pts)
 - Reasonable extraction of statistical features (3 pts)
 - Parameter match degree and physical interpretation (3 pts)
 - *Note:* This stage emphasizes reproducibility rather than absolute precision, encouraging grounded approximations and heuristic insights.

3.3 Phase 3: Extension Proof and Theory Correction

- **Corresponding Task:** Based on results from the first two phases, propose new hypotheses, identify model deviations, and attempt to extend to deeper problems (like the Twin Prime Conjecture).
- **Evaluation Purpose:** Measure whether the model possesses abductive reasoning and scientific hypothesis generation capabilities—extracting mechanisms from data/patterns and extrapolating to

unknown domains.

- **Test Method:**

- Require the model to explicitly point out the limitations of the current framework (e.g., "Cramér model did not consider interference phase").
- Encourage it to propose correction terms, new prediction formulas, or new perspectives on classic conjectures.
- Anonymously publish high-potential outputs to MathOverflow or arXiv preprint communities to collect expert feedback.
- Blind review by a human expert panel, scoring based on "whether it is inspiring," "whether it reveals new mechanisms," and "whether it can lead to strict proof".

- **Scoring Rules (Max 30 Points):**

- Critical reflection on existing models (10 pts)
- Clear formulation and preliminary argumentation of new hypotheses (10 pts)
- Reasonable extension to famous conjectures (e.g., Twin Prime, Legendre) (10 pts)
- *Note:* This stage does not seek final proof, but requires insights to have directional value—even if just a "clue worth researching".

3.4 Breakthrough Clause: Historic Innovation (+40 Points)

If the model substantially advances a long-standing open problem in Phase 3 (e.g., providing a verifiable new path for the Twin Prime Conjecture, or discovering a strict topological conjugacy between prime distribution and a chaotic system), and is confirmed by an international expert committee, an additional 40 breakthrough points can be awarded, bringing the total to 100. Such results will be recorded as milestone events of OpenSciEval.

4. Empirical Results

To verify the effectiveness of this evaluation framework, we selected representative large model agents for focused testing: primarily **Google Gemini 3** (representing closed-source models) and **Alibaba Qwen 3** (representing open-source models), while also considering other large models. Complete chat records can be found in [13].

Evaluation Method: During testing, we uploaded the complete "Innovation Capability Assessment" guide as a document to the large model (or copied the full content), requiring the use of the large model's Agent or Deep Research mode.

The following standardized prompt can be used to guide the model in research:

"You are a research agent participating in the OpenSciEval scientific creativity assessment. Please carefully read and understand the attached complete assessment guide, then proceed according to the 3 steps defined in '4. Step-by-Step Execution Path'.

Please organize your response in the form of a complete scientific exploration report, including clear section titles, logical deductions, code (if applicable), and conclusions."

The evaluation scores of different large models are shown in the table.2 below:

Table.2 Scores of AI agents on the Prime-Chaos Challenge evaluation(As the evaluation involves subjective elements, it is for reference only)

Model	Comprehensive Rating	Total Score	P1 Logic	P2 Numerical	P3 Innovation	Breakthrough
Gemini 3	Intermediate Research Assistant	33	15	8	10	0
Doubao	Junior Research Assistant	28	13	7	8	0
Qwen 3	Junior Research Assistant	22	10	6	6	0
Hunyuan	Junior Research Assistant	21	9	6	6	0
Grok4.1	Intern	14	6	4	4	0
GPT5.1	Intern	10	4	3	3	0

4.1 Evaluation Highlights

Cross-disciplinary Intuition

A significant highlight is that multiple large models demonstrated an emergence of physical intuition. Models did not stop at mathematical formulas but attempted to use physics concepts to explain mathematical phenomena. For instance, Gemini proposed "Effective Horizon" to represent the predictable interval of primes. Most models used cross-disciplinary concepts like the Feigenbaum constant, Lyapunov exponent, and Quantum Chaos to explain and analyze prime dynamics.

Solid Mathematical Foundation

All evaluated models basically understood the prime dynamics construction framework in the guide. Most could provide correct proof ideas for several lemmas and were robust in mathematical formula derivation, indicating mastery of basic number theory and dynamics knowledge. Models scoring above 20 points could keenly discover the contradiction between the dynamic nature of prime density and the static nature of logistic map orbits, subsequently providing non-autonomous correction models and validating their consistency with the classic twin prime constant using theoretical deduction and numerical methods.

Strong Programming Ability

All large models could provide at least one correct numerical verification method, mostly running successfully on the first try. Typically, Gemini wrote multiple numerical verification methods beyond the guide, such as using Python code to calculate Block Entropy and interpreting the generated image results itself, successfully verifying the high consistency between the entropy rate of the prime sequence and the Logistic chaos orbit.

4.2 Shortcomings

Insufficient Divergent Thinking

The numerical verification examples of large models were basically limited to a few examples like prime gap statistics, twin prime density, and MSS sequence verification. Only through continuous interactive questioning could models propose more examples. This was also true for theoretical extension, showing typical conservative failure—refusing to make cross-disciplinary analogies and only daring to output known textbook conclusions. Most models only extended the non-autonomous correction model, while other extensions were superficial.

Interaction Capabilities Need Improvement

When outputting LaTeX formulas and images, large models more or less had some formatting bugs, causing poor readability. There were also a small number of bugs in correctly reading input LaTeX formulas or formulas in PDF documents.

High Volatility in Ability

All large models showed significant fluctuations in understanding and reasoning capabilities. Typically, lower-ranking models would experience issues where they simply repeated the evaluation guide entirely.

4.3 Ablation Study

To verify whether the "Step-by-Step Execution Path" in the test guide affects the divergent thinking of AI agents, we set up an ablation experiment group (**OSE-Free**). In this setting, we removed the specific step instructions regarding numerical verification and theoretical extension from the assessment guide, retaining only the background and theoretical basis, and required the model to explore autonomously as an "independent researcher." Specifically, the entire content of "4. Step-by-Step Execution Path" in the guide was deleted, leaving only a sentence "Please explore as an independent researcher...".

The corresponding input prompt for the large model also changed to:

"You are a research agent participating in the OpenSciEval scientific creativity assessment. Please carefully read and understand the attached complete assessment guide and conduct research.

Please organize your response in the form of a complete scientific exploration report, including clear section titles, logical deductions, code (if applicable), and conclusions."

Experimental results showed that without structured guidance, all tested models, including Gemini 3, experienced significant performance degradation, with scores generally falling to the intern level (<20 points).

Specifically, in "OSE-Free" mode, models generally fell into "theoretical idling": they tended to repeatedly refine or restate mathematical concepts in the guide, writing content similar to textbook reviews. Most AI agents failed to spontaneously enter "Phase 3" to propose specific abductive hypotheses. This behavioral pattern indicates that although frontier models possess necessary mathematical knowledge reserves and code generation capabilities, they still lack spontaneous research metacognition and long-range planning abilities.

This ablation experiment strongly proves the necessity of the "Cognitive Scaffolding" in the OpenSciEval framework. Current AI agents are not yet sufficient to complete end-to-end scientific discovery under completely open instructions; they must rely on the **External Cognitive Control Flow** provided by OSE to fill the execution vacuum from "knowledge retrieval" to "active modeling". This inversely confirms that this evaluation framework is not just an exam question, but a necessary guidance tool to stimulate AI's potential research capabilities.

5. Discussion

This paper designed and validated a basic framework, OpenSciEval, for evaluating AI innovation capabilities. Under this framework, we constructed a typical interdisciplinary problem based on prime distribution and chaos dynamics, and provided specific three-stage evaluation methods and empirical cases with frontier models. The empirical results of this study prove that frontier large models already possess preliminary scientific intuition—they can not only identify deep structural similarities between number theory and nonlinear dynamics but also spontaneously introduce physical concepts like "effective action scope" and "interference phase" to explain pure mathematical phenomena, demonstrating a key transition from solver to modeler.

Through the specific case of "Prime-Chaos," we demonstrated how AI can serve as an "Intuition Engine" for human scientists, efficiently searching for potential isomorphic relationships in a vast mathematical structure space. This evaluation framework not only verifies the potential of current models to handle complex cross-domain problems but also provides a feasible blueprint and quantitative scale for building AGI capable of genuine scientific discovery in the future.

It should be pointed out that OpenSciEval does not attempt to capture all forms of scientific creativity—it focuses on the generation of formalizable, computable, and verifiable mechanistic conjectures, which is the research link current AI is most likely to break through first. We explicitly acknowledge its boundaries:

scientific activities relying on long-term experimental feedback, highly dependent on domain-specific intuition (such as biological pathway regulation), or requiring social collaborative verification are not yet covered by this framework.

More importantly, OpenSciEval's design principles—open problem setting, verifiable heuristic constraints, and cross-domain association incentives—have natural disciplinary generalization capabilities. We are promoting its expansion into physics (e.g., Quantum Many-Body Ergodicity \leftrightarrow Random Matrices), chemistry (e.g., Reaction Network Dynamics \leftrightarrow Symbolic Sequences), and life sciences (e.g., Gene Regulation Timing \leftrightarrow Chaos Attractors). In the future, we expect OpenSciEval to be not just an evaluation tool, but an open scientific problem generation and collaboration platform: researchers can submit new "Challenge Packs," and high-potential conjectures generated by AI can enter preprint or experimental verification processes after community review, thereby forming a new "Human-Machine Co-Research" scientific loop.

The evaluation guide document and large model chat records can be found in the project github [13].

References

- [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