AIGS：全自主AI科学发现探索

李鹏

清华大学

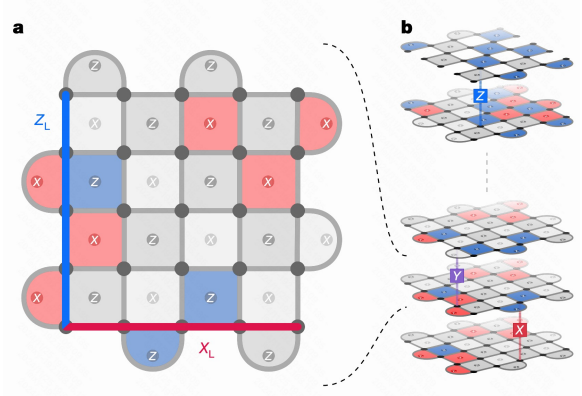
AI深刻改变科学研究范式

清華大學 Tsinghua University

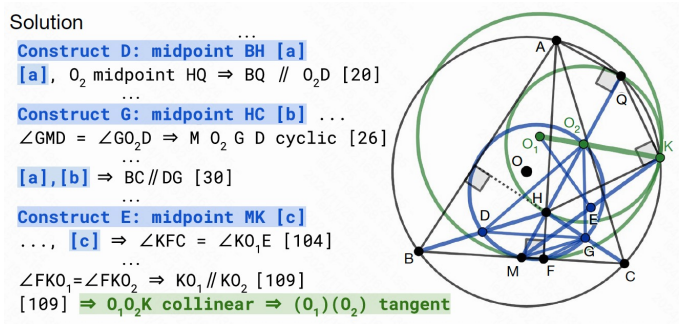
- 深度学习的发展赋能AI4Science，极大地推动了科学的发展。

John J. Hopfield and Geoffrey Hinton. Ill. Niklas Elmehed © Nobel Prize Outreach

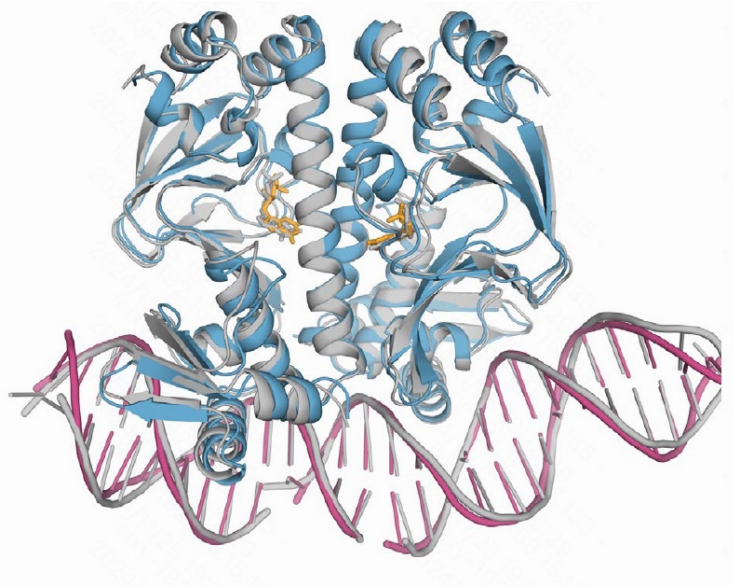
David Baker, Demis Hassabis and John Jumper. Ill. Niklas Elmehed © Nobel Prize Outreach



AlphaQuant



AlphaGeometry

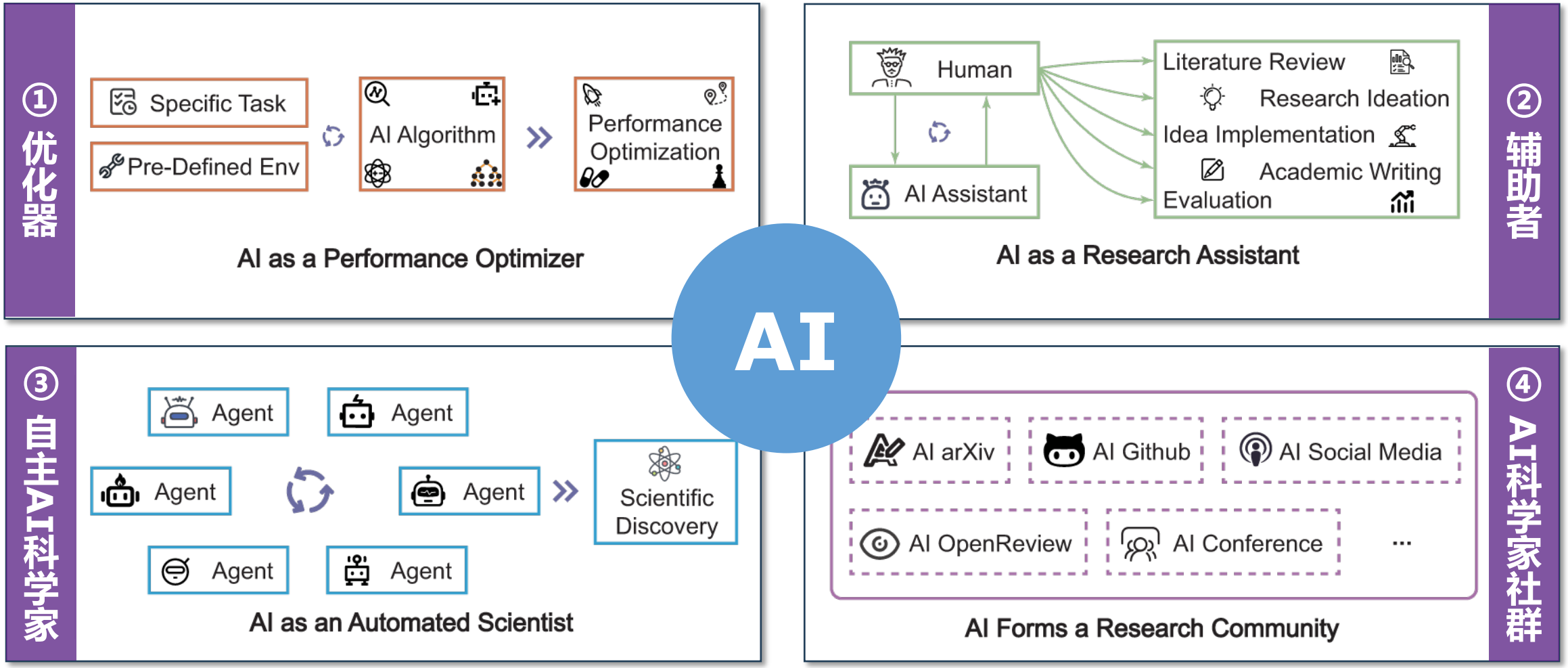


AlphaFold

https://www.nobelprize.org/all-nobel-prizes-2024/
https://www.nature.com/articles/s41586-024-07487-w
https://www.nature.com/articles/s41586-024-08148-8
https://www.nature.com/articles/s41586-023-06747-5

AI参与科研的范式

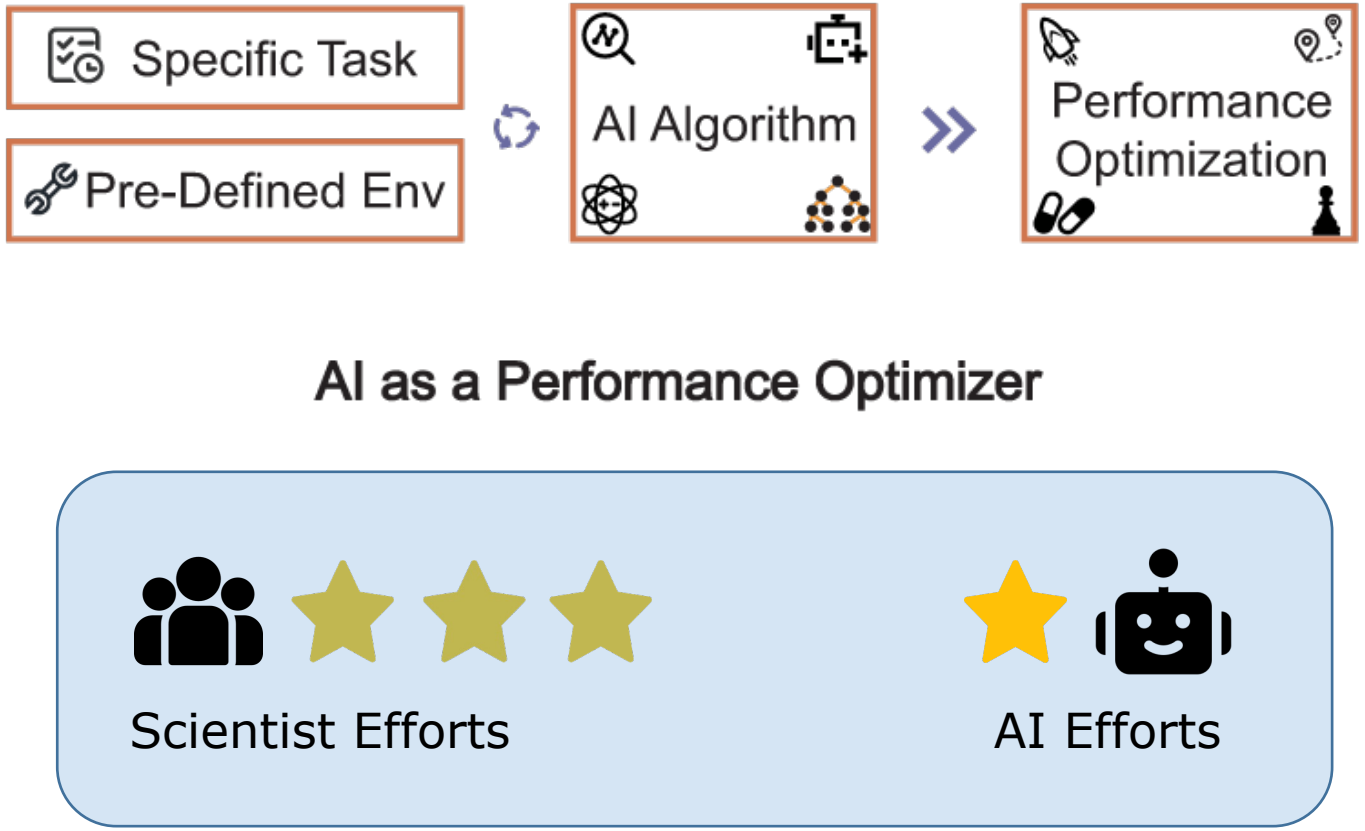
- 根据AI在科研中的角色可以划分出 4 种AI参与科学研究的范式。

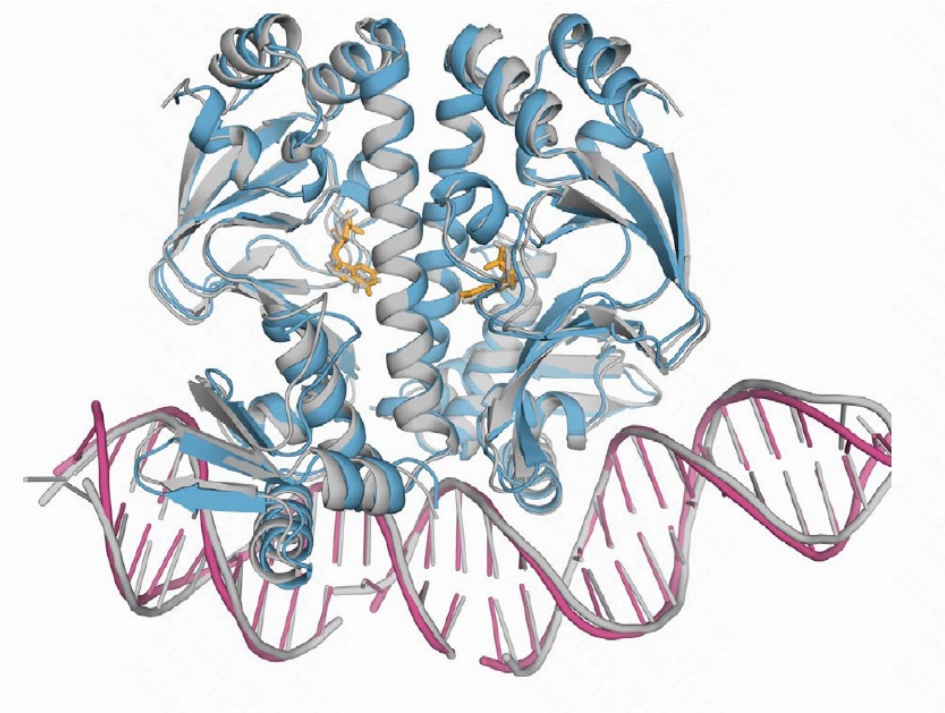


Liu et al., 2024. AIGS: Generating Science from AI-Powered Automated Falsification. arXiv:2411.11910.

范式一：优化器

- AI在科研中的角色：专注于优化特定任务的表现。



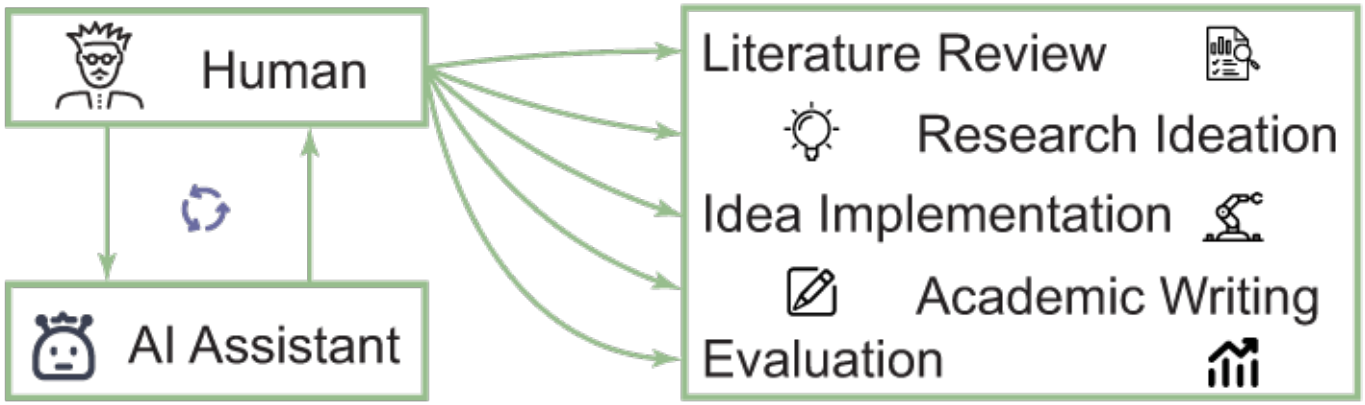


AlphaFold

https://www.nature.com/articles/s41586-024-08148-8

范式二：辅助者

- AI在科研中的角色：专注于在特定科研环节上帮助人类研究者。



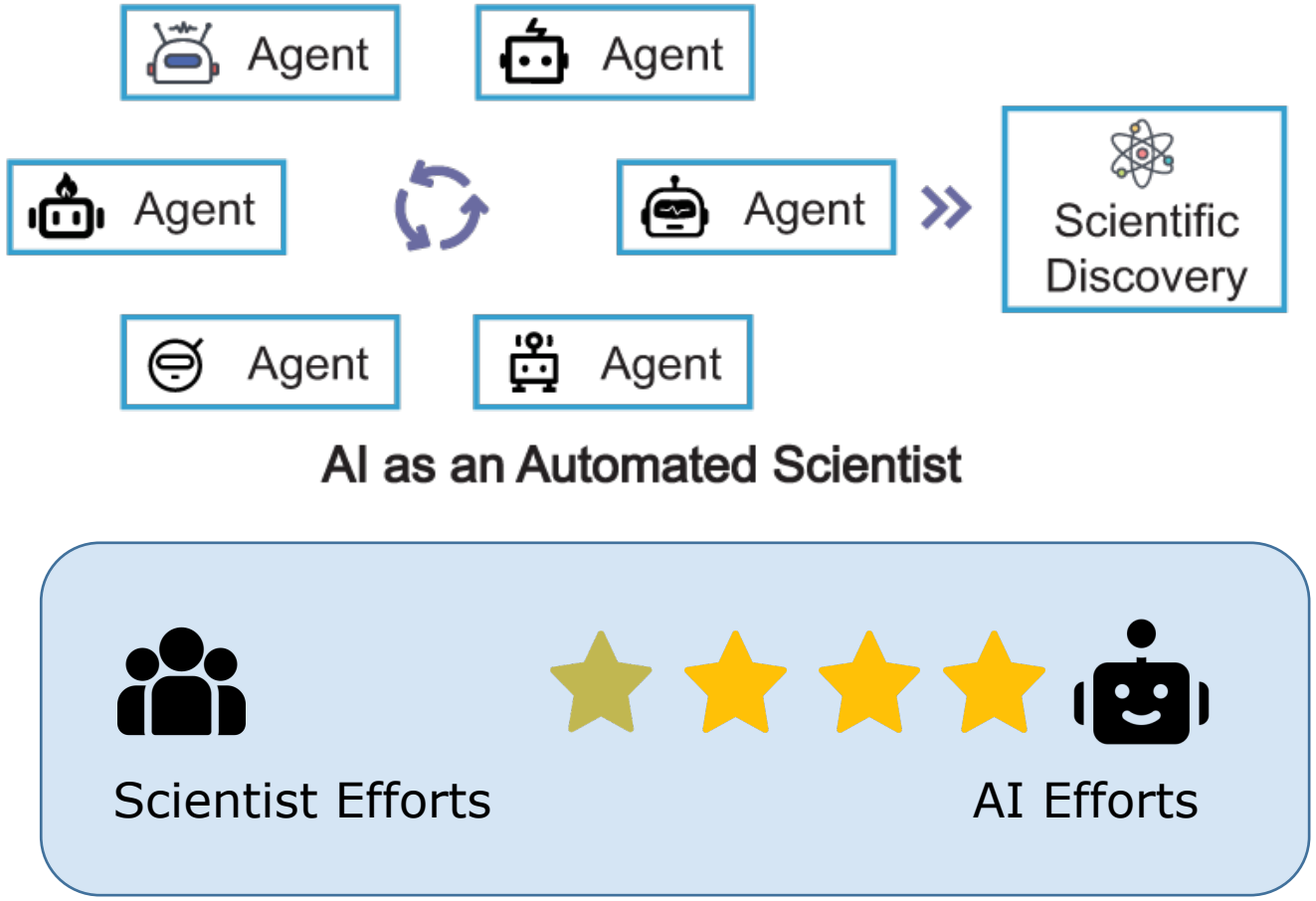
AI as a Research Assistant

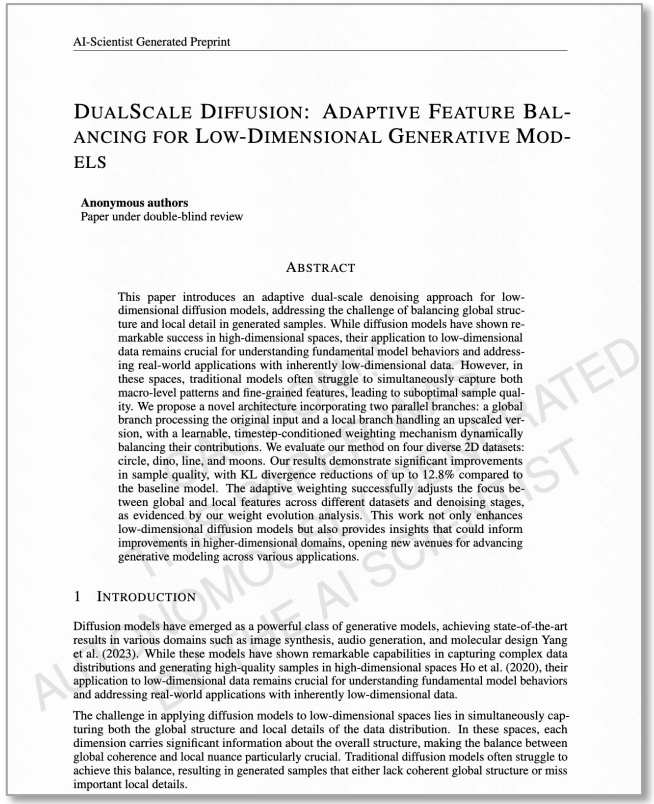
Github Copilot

https://github.com/features/copilot

范式三：自主AI科学家

- AI在科研中的角色：专注于端到端进行自主科学研究。





AI-Scientist Generated Preprint

DUALSCALE DIFFUSION: ADAPTIVE FEATURE BALANCING FOR LOW-DIMENSIONAL GENERATIVE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper introduces an adaptive dual-scale denoising approach for low-dimensional diffusion models, addressing the challenge of balancing global structure and local detail in generated samples. While diffusion models have shown remarkable success in high-dimensional spaces, their application to low-dimensional data remains crucial for understanding fundamental model behaviors and addressing real-world applications with inherently low-dimensional data. However, in these spaces, traditional models often struggle to simultaneously capture both macro-level patterns and fine-grained features, leading to suboptimal sample quality. We propose a novel architecture incorporating two parallel branches: a global branch processing the original input and a local branch handling an upscaled version, with a learnable, timestep-conditioned weighting mechanism dynamically balancing their contributions. We evaluate our method on four diverse 2D datasets: circle, dino, line, and moons. Our results demonstrate significant improvements in sample quality, with KL divergence reductions of up to 12.8% compared to the baseline model. The adaptive weighting successfully adjusts the focus between global and local features across different datasets and denoising stages, as evidenced by our weight evolution analysis. This work not only enhances low-dimensional diffusion models but also provides insights that could inform improvements in higher-dimensional domains, opening new avenues for advancing generative modeling across various applications.

1 INTRODUCTION

Diffusion models have emerged as a powerful class of generative models, achieving state-of-the-art results in various domains such as image synthesis, audio generation, and molecular design Yang et al. (2023). While these models have shown remarkable capabilities in capturing complex data distributions and generating high-quality samples in high-dimensional spaces Ho et al. (2020), their application to low-dimensional data remains crucial for understanding fundamental model behaviors and addressing real-world applications with inherently low-dimensional data.

The challenge in applying diffusion models to low-dimensional spaces lies in simultaneously capturing both the global structure and local details of the data distribution. In these spaces, each dimension carries significant information about the overall structure, making the balance between global coherence and local nuance particularly crucial. Traditional diffusion models often struggle to achieve this balance, resulting in generated samples that either lack coherent global structure or miss important local details.

AI Generated Paper

Lu et al., 2024. The AI Scientist: Towards Fully Automated Open-ended Scientific Discovery. arXiv:2408.06292.

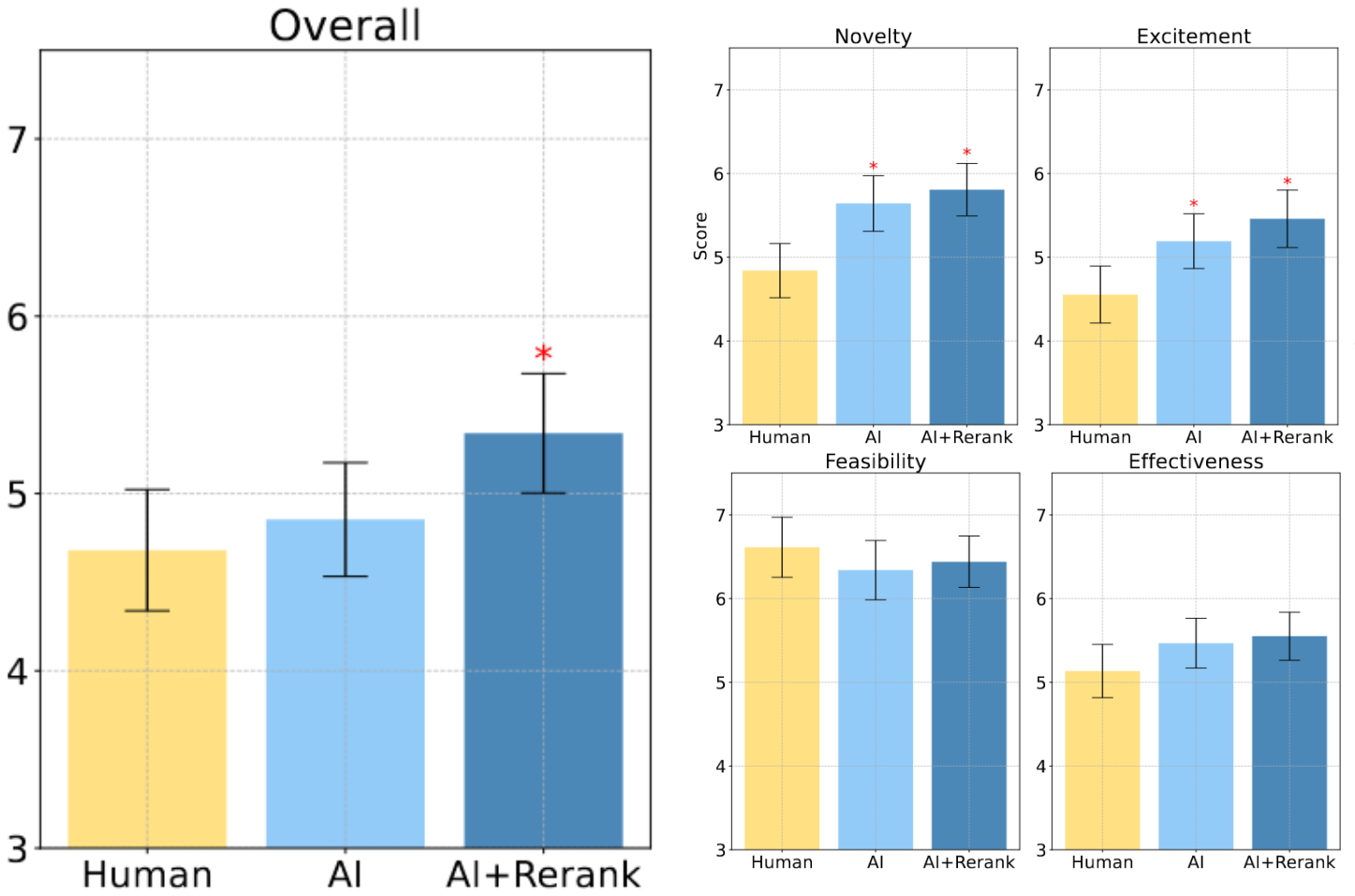
范式三的初步探索

清華大學 Tsinghua University

- **已有相当数量工作探索利用AI自主提idea可行性，展现出可观潜力。**

斯坦福大学研究显示大模型智能体在特定条件下可以产生创新度优于人类的idea

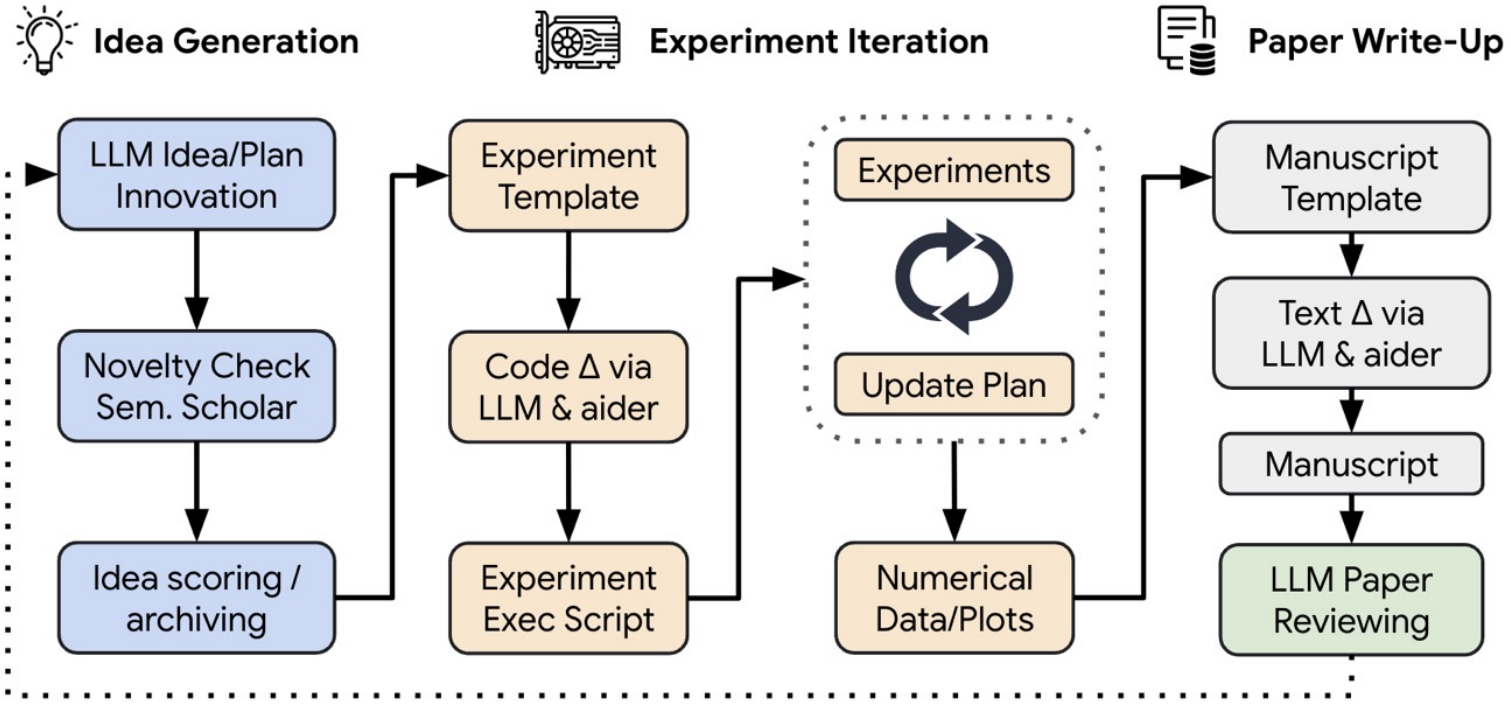
- ❑ 考查了7个NLP主题
- ❑ 约束在提示学习子领域内
- ❑ 系统无需对idea进行验证
- ❑ 人类被试提供的未必是其个人想到的最优idea

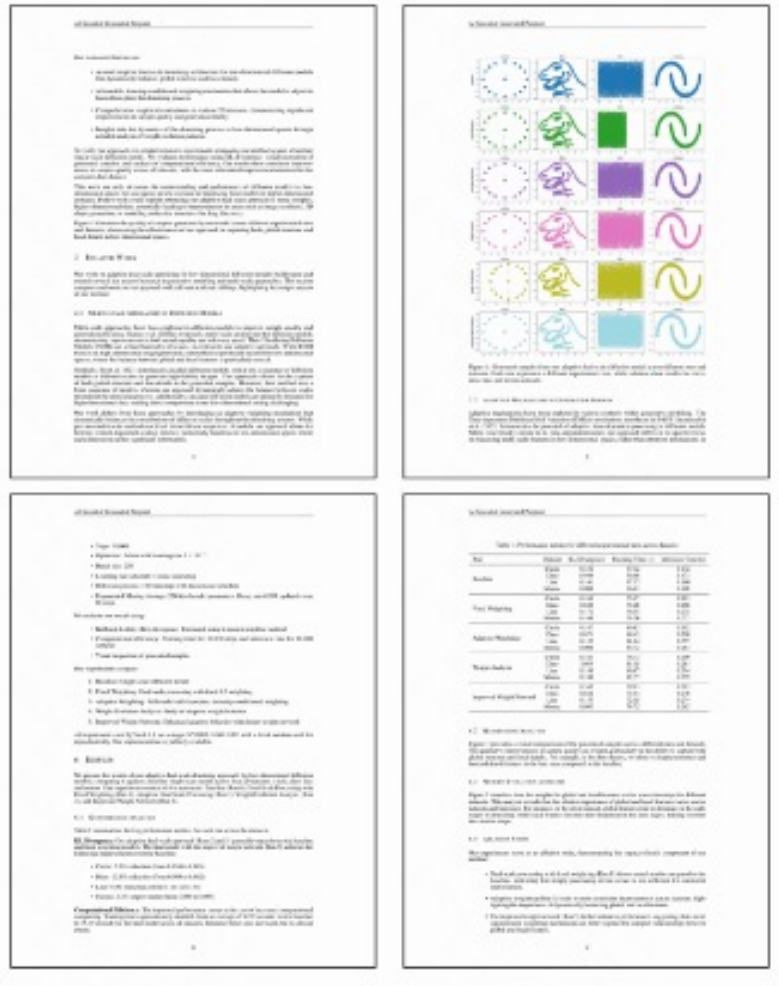


Si et al., 2024. Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers. arXiv:2409.04109.

范式三的初步探索

- 全流程验证性系统快速涌现，已展现出智能体作为AI科学家的可行性。





AI Scientist系统展现了智能体系统自主完成”提idea→做实验→写论文”全流程的可行性，被Nature新闻报道

Lu et al., 2024. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. arXiv:2408.06292.

可信任性挑战

- 虽然AI Scientist 产生的论文图文并茂，但难于被信任。

“它是由AI产生的”是影响您判断的根本性因素吗?

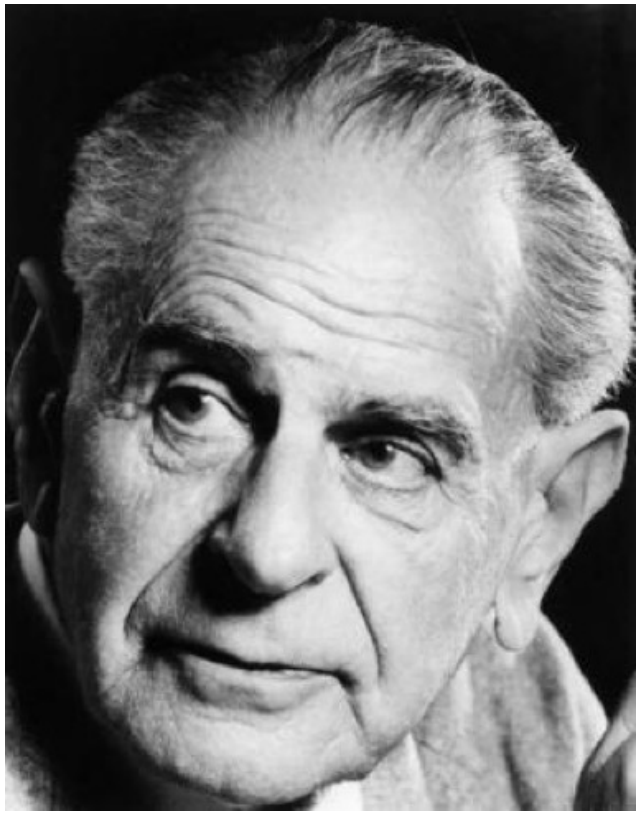
我们为什么会相信AlphaFold产生的结果?

Lu et al., 2024. The AI Scientist: Towards Fully Automated Open-ended Scientific Discovery. arXiv:2408.06292.

科学与非科学的分水岭

- **可证伪性是区分科学与非科学的关键因素：**

卡尔·波普尔在其著作《科学发现的逻辑》中提出了“可证伪性”的概念，用以区分科学与非科学的理论。他认为，**一个理论只有在能够被经验或实验观察所反驳的情况下，才具有科学性**。例如，命题“所有的天鹅都是白色的”是可证伪的，因为只需观察到一只黑天鹅即可推翻该命题。因此，波普尔强调，**科学理论应具备可证伪性，即存在被经验事实否定的可能性**。



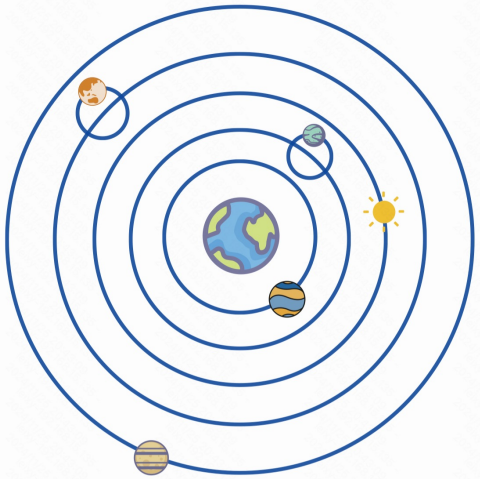
卡尔·波普尔

Popper, 1935. The Logic of Scientific Discovery.

科学研究的核心是证伪

- 在科学研究的过程中，什么是证伪 (Falsification)?
- 以日心说的科学革命为例：

地心说

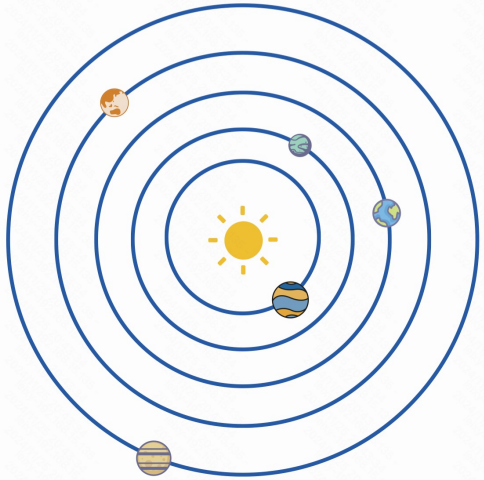


Geocentric Theory:

- Earth is stationary;
- Celestial bodies revolve around the Earth;
- Circular motion.
- ...

V.S.

日心说



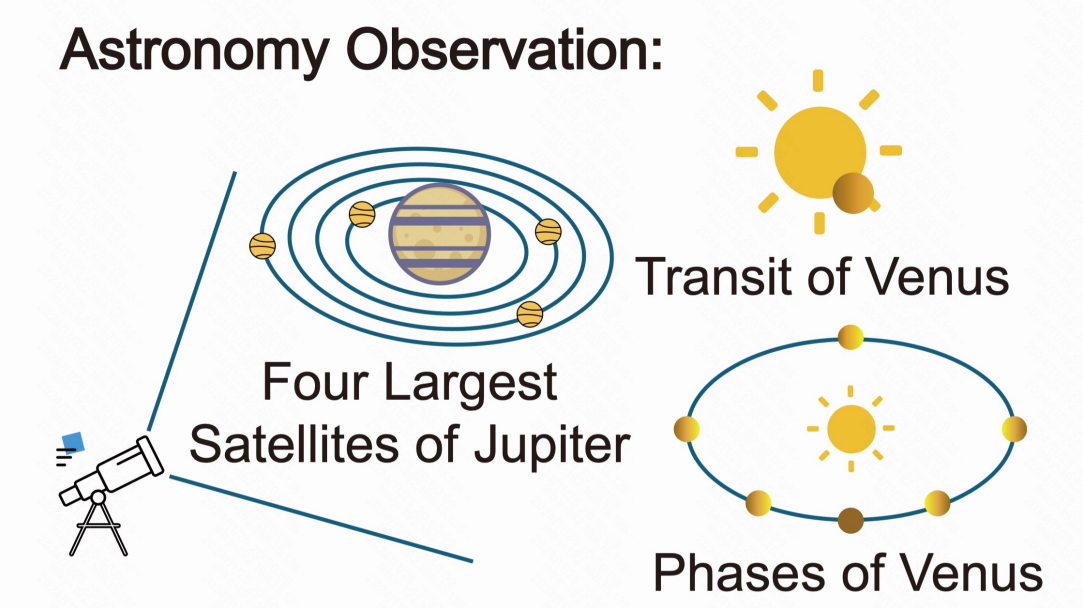
Heliocentric Theory:

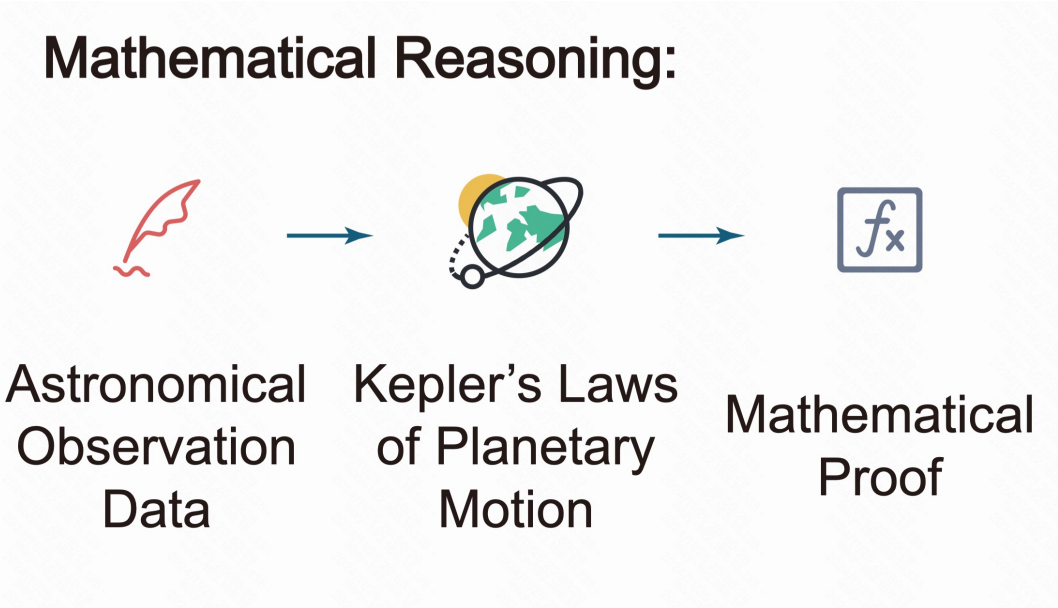
- Earth is rotating;
- Celestial bodies revolve around the Sun;
- Circular motion.
- ...

Liu et al., 2024. AIGS: Generating Science from AI-Powered Automated Falsification. arXiv:2411.11910.

科学研究的核心是证伪

- 在科学研究的过程中，什么是证伪 (Falsification)?
- 需要设计实验证明假设正确或错误：





设计观察性实验验证假说

数学形式化推理验证假说

Liu et al., 2024. AIGS: Generating Science from AI-Powered Automated Falsification. arXiv:2411.11910.

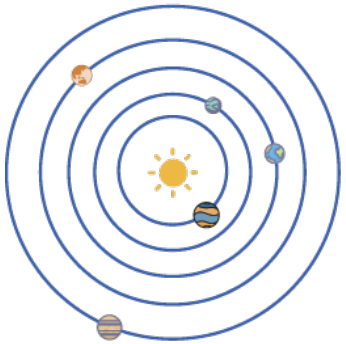
科学研究的核心是证伪

- 只有经过严格的证伪过程，才能产生可靠的科学发现。

Hypothesis

Geocentric Theory:

- Earth is stationary;
- Celestial bodies revolve around the Earth;
- Circular motion.
- ...

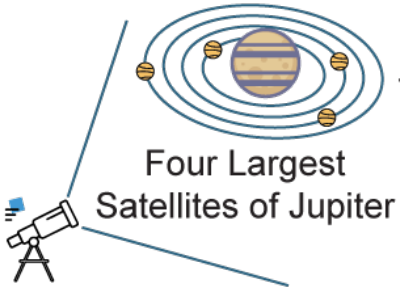


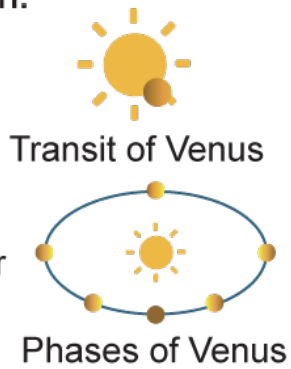
Heliocentric Theory:

- Earth is rotating;
- Celestial bodies revolve around the Sun;
- Circular motion.
- ...

Falsification

Astronomy Observation:





Mathematical Reasoning:

Astronomical Observation Data → Kepler's Laws of Planetary Motion → Mathematical Proof

Scientific Discovery

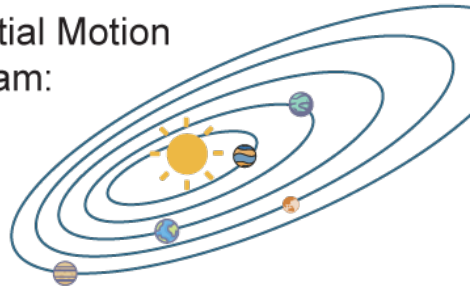
Final Discovery:

- Earth is rotating;
- Celestial bodies revolve around the Sun;
- Elliptical orbit motion.
- ...

→ Geocentric Theory ✗

Heliocentric Theory ✓

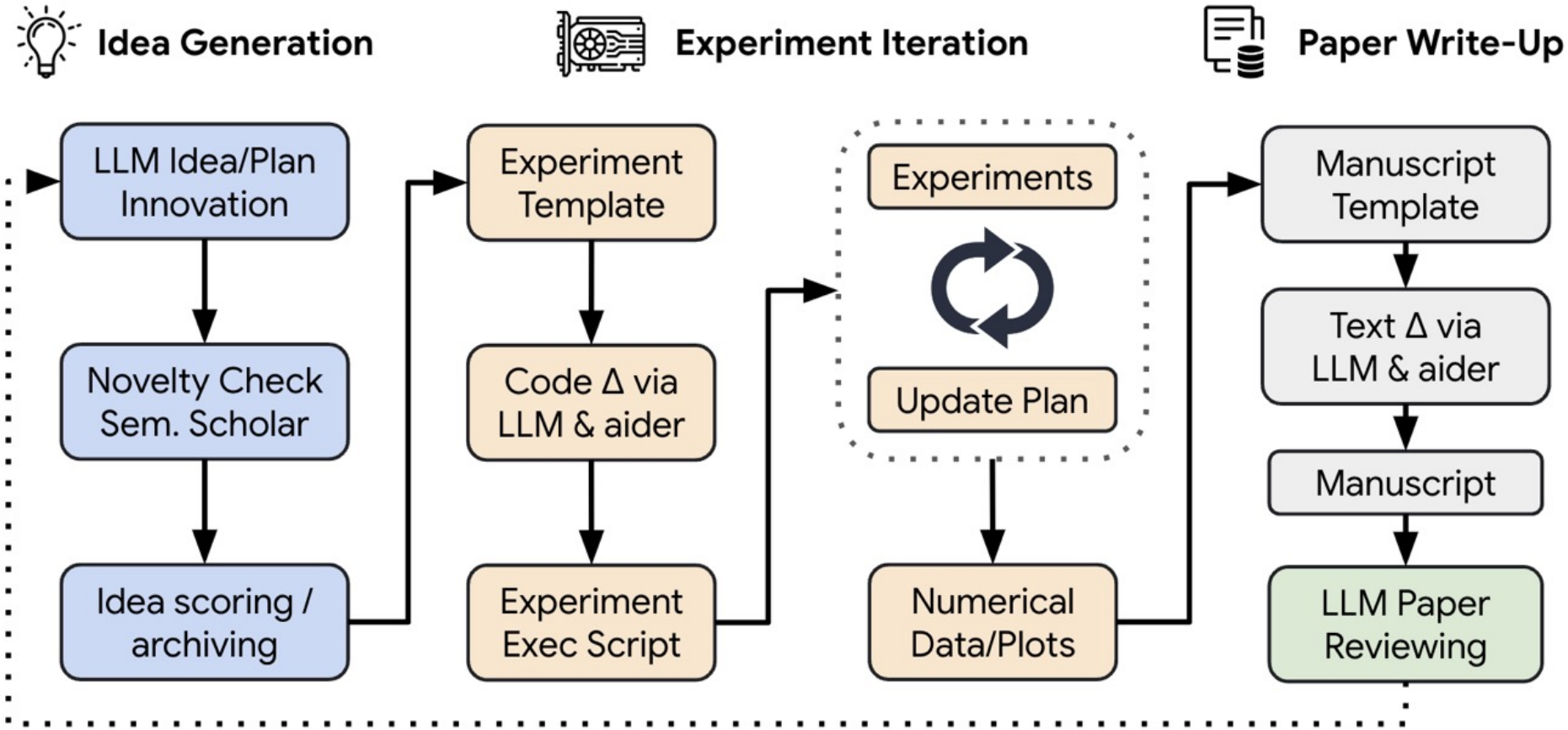
→ Celestial Motion Diagram:



Liu et al., 2024. AIGS: Generating Science from AI-Powered Automated Falsification. arXiv:2411.11910.

目前的自主科研系统缺少自主证伪

- 带着证伪的视角审视目前的科研智能体架构：普遍缺少自主证伪 (Falsification) 过程，**未能自主排除其他合理假设。**



Lu et al., 2024. The AI Scientist: Towards Fully Automated Open-ended Scientific Discovery. arXiv:2408.06292.

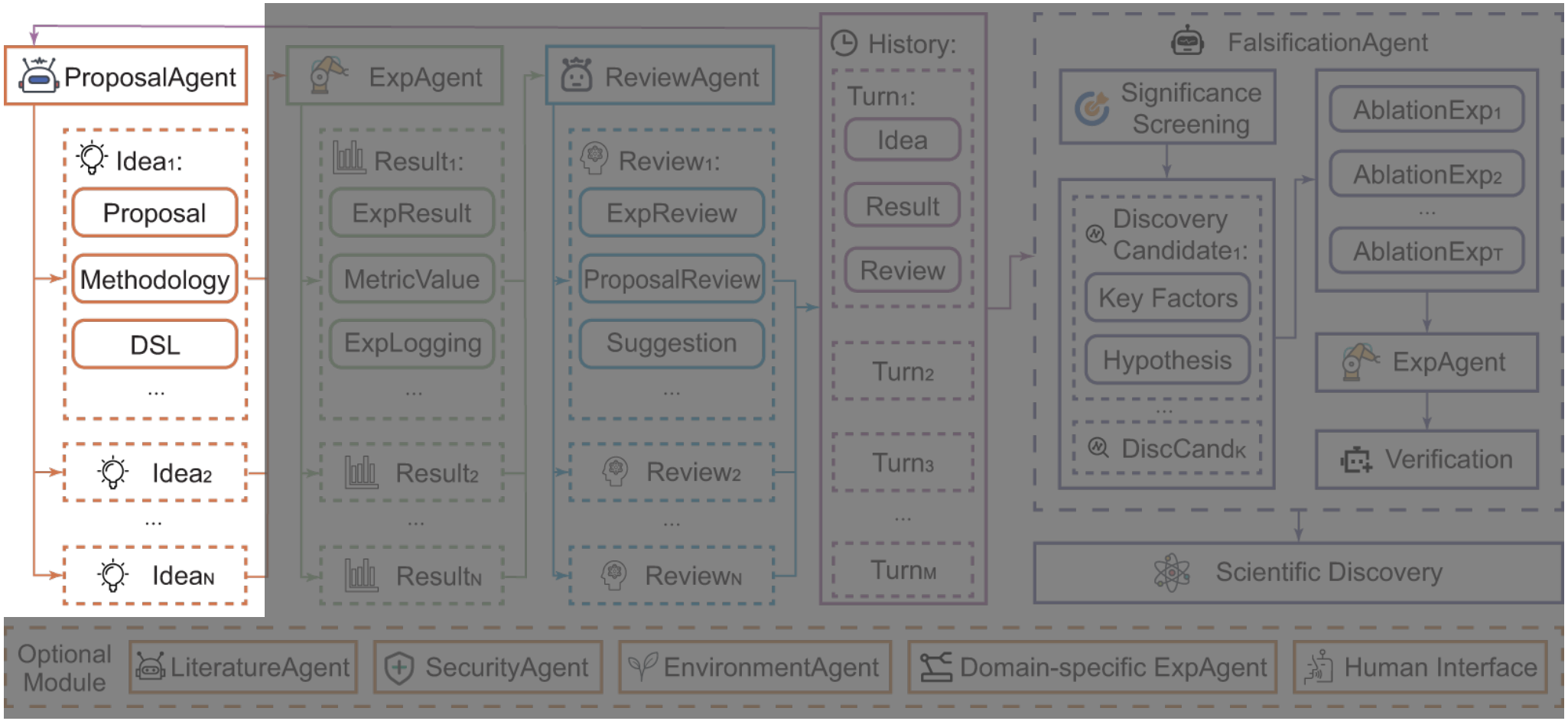
AIGS及核心原则

- AIGS（AI-Generated Science）：AI全自主作出的符合科学标准的科学发现。
- 实现AIGS的核心原则是保证系统具备证伪能力、创造性和可执行性：
  - 证伪 (Falsification) 是科学研究的核心
    - 设计并执行实验以验证或反驳假设
  - 创造性 (Creativity) 的想法是科学研究的起点
    - 科学发现需要不断提出新的假设
  - 可执行性 (Executability) 构成了证伪的基础
    - 科学假设需要可执行的实验来验证

Liu et al., 2024. AIGS: Generating Science from AI-Powered Automated Falsification. arXiv:2411.11910.

Baby-AIGS

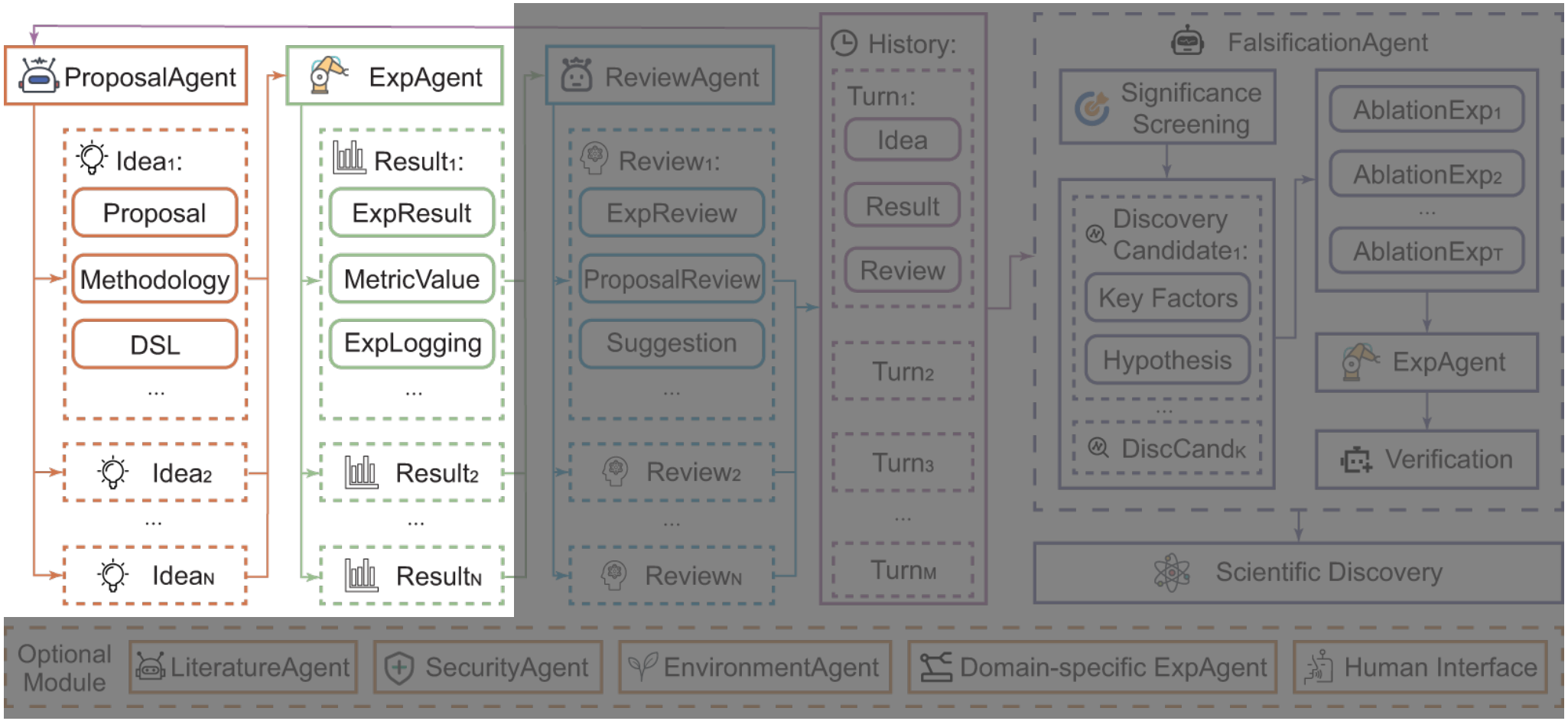
- **ProposalAgent：提出具有创造力的科研想法。**



Liu et al., 2024. AIGS: Generating Science from AI-Powered Automated Falsification. arXiv:2411.11910.

Baby-AIGS

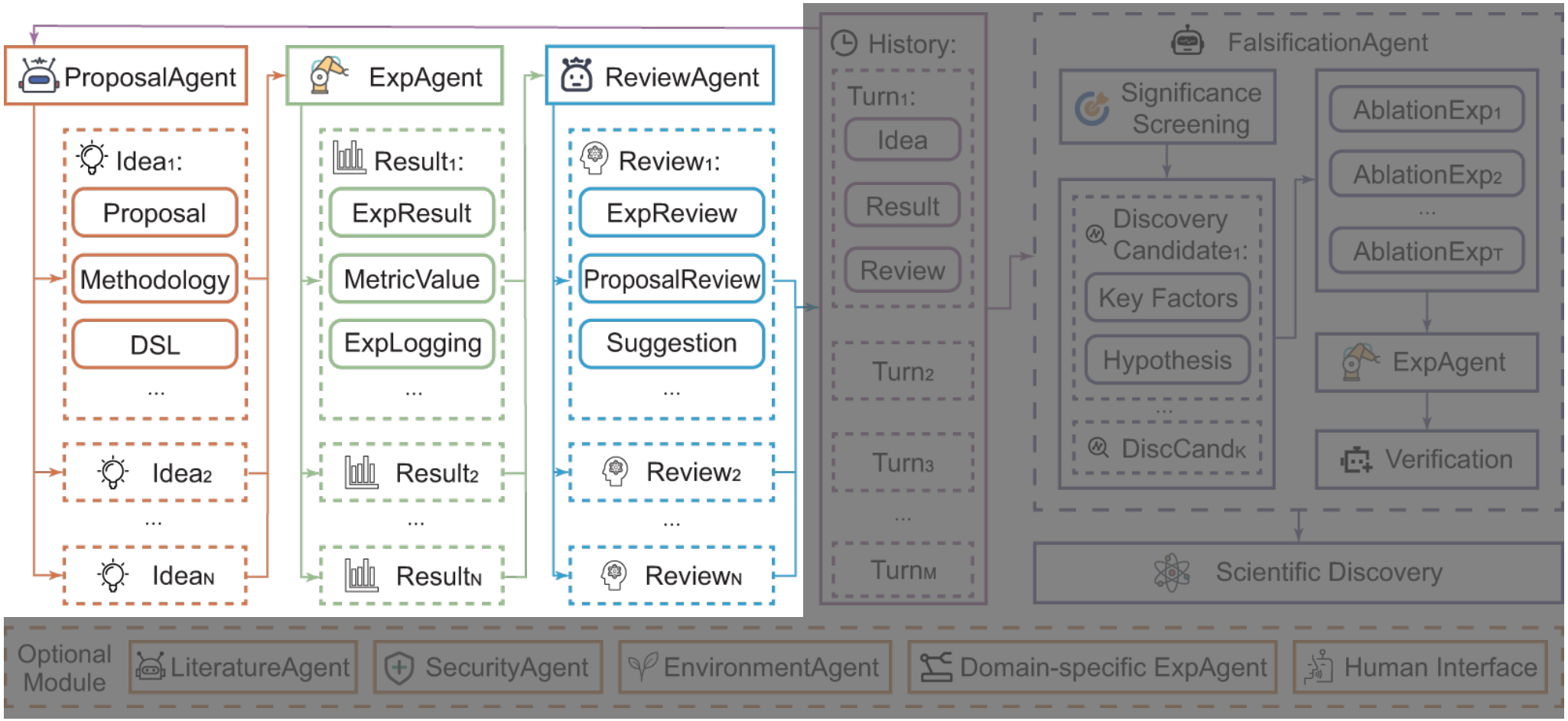
- **ExpAgent：实现想法并且进行科学实验。**



Liu et al., 2024. AIGS: Generating Science from AI-Powered Automated Falsification. arXiv:2411.11910.

Baby-AIGS

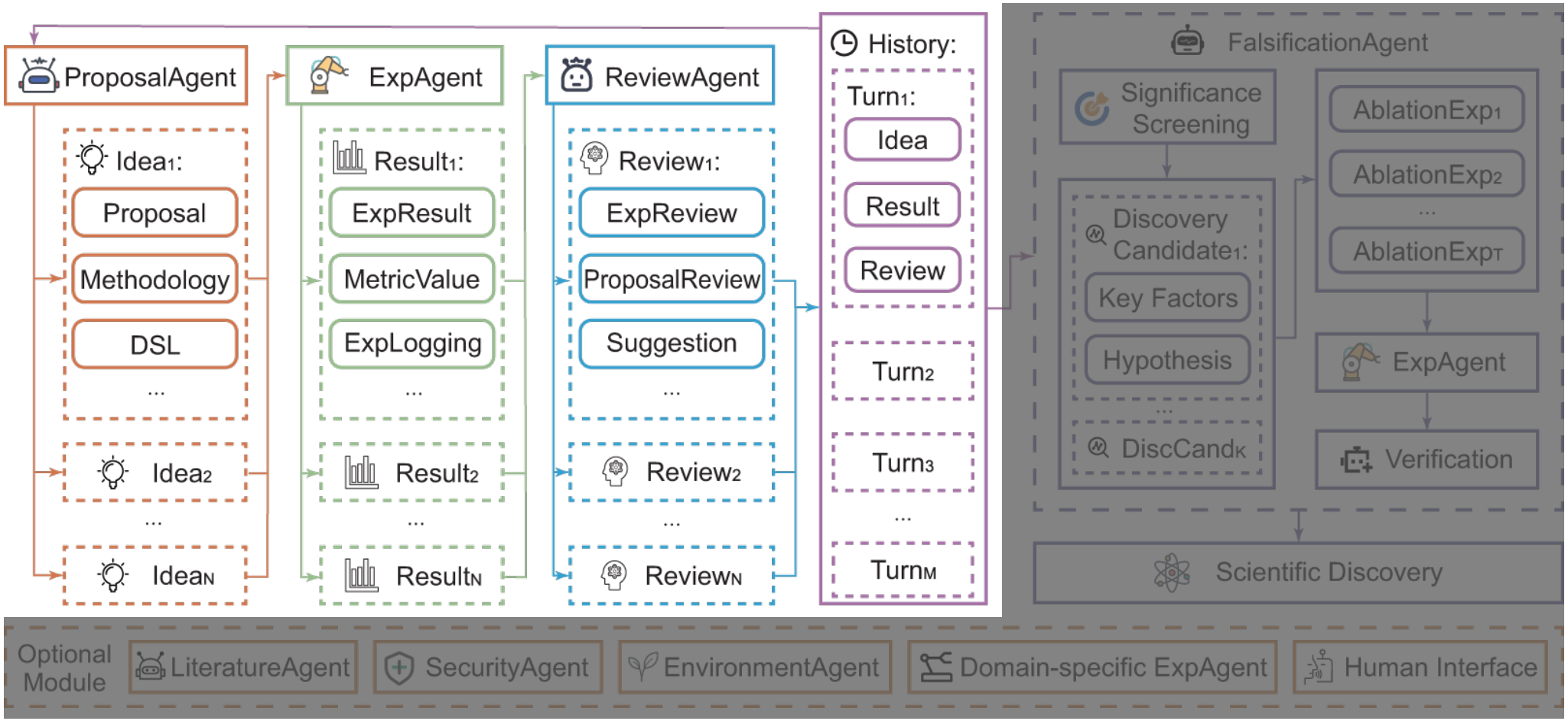
- **ReviewAgent：根据实验结果对想法提出建设性反馈。**



Liu et al., 2024. AIGS: Generating Science from AI-Powered Automated Falsification. arXiv:2411.11910.

Baby-AIGS

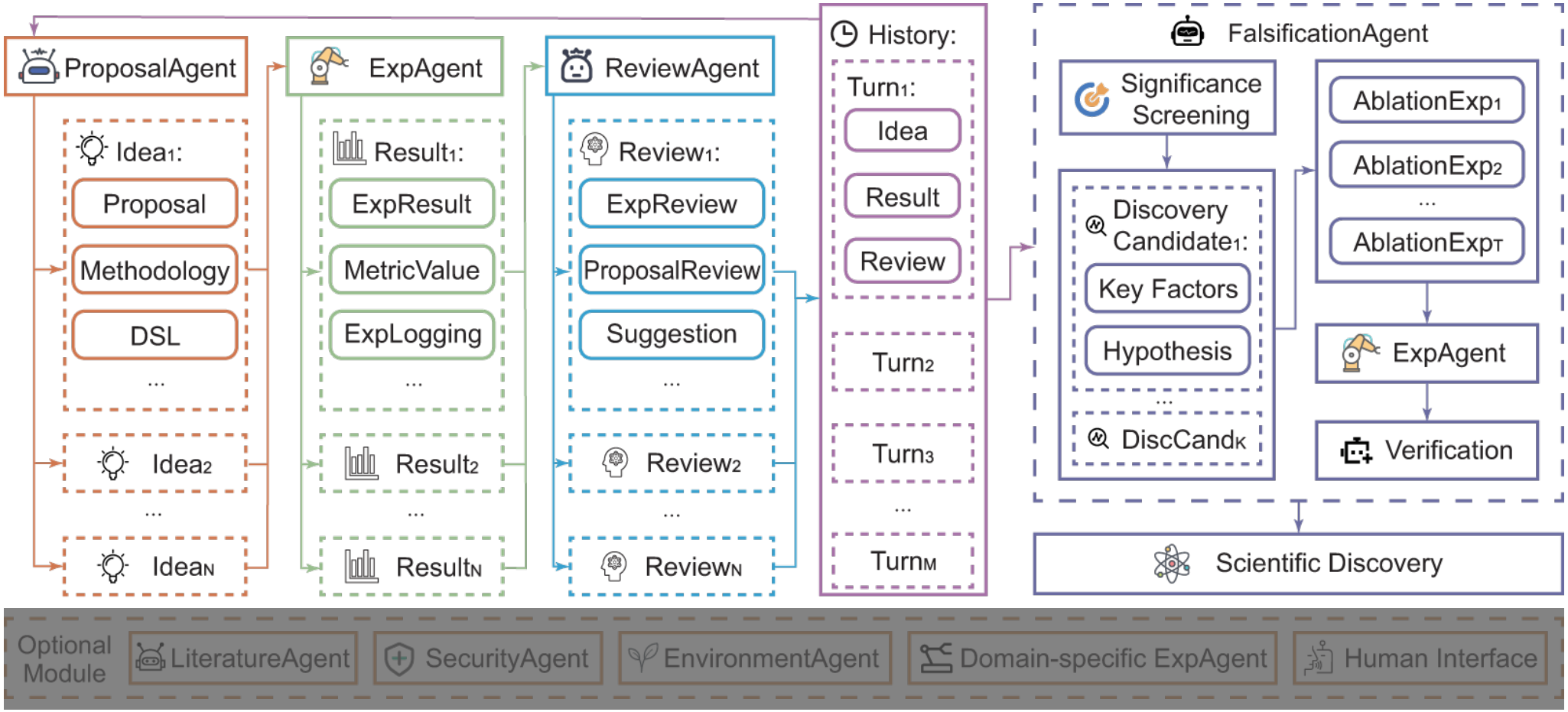
- **Iteration：循环迭代以提高实验表现。**



Liu et al., 2024. AIGS: Generating Science from AI-Powered Automated Falsification. arXiv:2411.11910.

Baby-AIGS

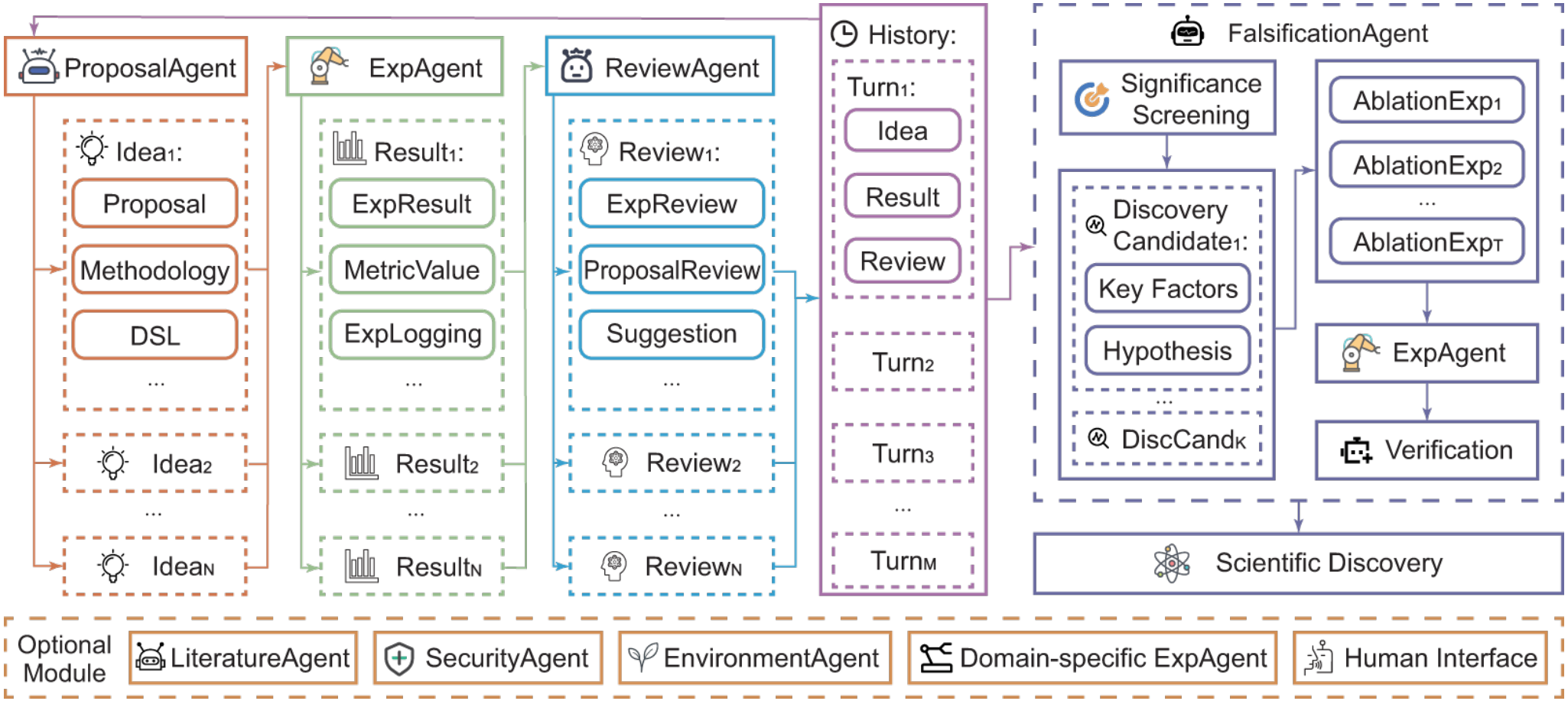
- **FalsificationAgent：消融实验来进一步验证科学发现。**



Liu et al., 2024. AIGS: Generating Science from AI-Powered Automated Falsification. arXiv:2411.11910.

Baby-AIGS

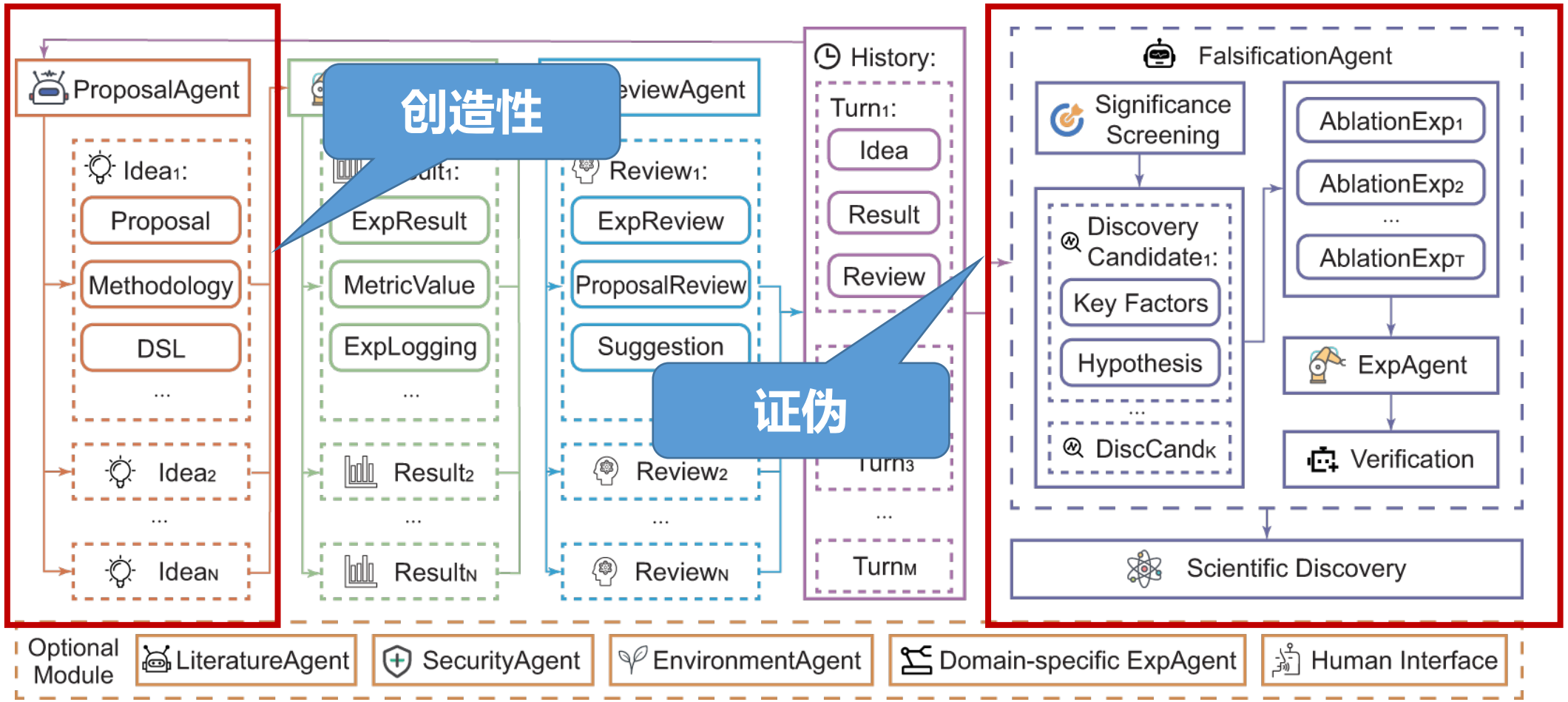
- **Optional Module：可集成其他模块实现补充功能。**



Liu et al., 2024. AIGS: Generating Science from AI-Powered Automated Falsification. arXiv:2411.11910.

三原则：证伪和创造性

- 证伪和创造性主要由FalsificationAgent和ProposalAgent实现。

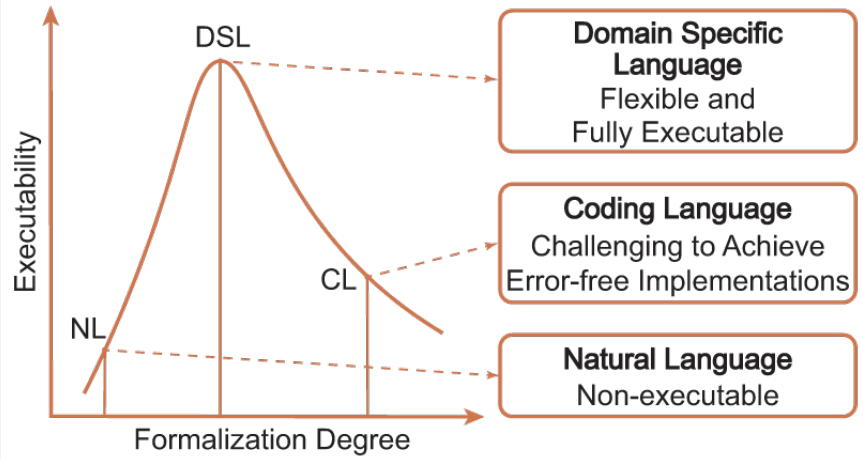


Liu et al., 2024. AIGS: Generating Science from AI-Powered Automated Falsification. arXiv:2411.11910.

三原则：可执行性

- 领域特定语言（DSL）：结合自然语言和代码语言的优势提升可执行性，**同时也赋予了系统更好的跨领域迁移性。**

可执行性



Domain Specific Language

Pre-Defined

Framework Settings:
- Method: Bootstrap / ...
- Benchmark: AlfWorld / ...

Available Functions:
- random_truncate(param)
- re_simulate(param)
- call_llm(param)
- ...

LLM Generated

```
{
  "Method": "Bootstrap",
  "Benchmark": "AlfWorld",
  "Pseudocode": "```python
new_traj = random_truncate()
new_traj = resimulate()
prompt = ''
new_traj = call_llm(prompt)```"
}
```

Natural Language

LLM Generated

Language agents can gather high-quality trajectories by adjusting actions in the original ReAct sequence and ...

Coding Language

LLM Generated

```
# main.py
from trajectory import \
    random_truncate, resimulate
from llm_handler import \
    call_llm

def main():
    traj = random_truncate()
    resimulated_traj = \
    resimulate(traj)
    ...

def generate_prompt(traj):
    ...

if __name__ == "__main__":
    main()
```

```
# trajectory.py
import random
...
def secondary_simulation(traj):
    return [step / (random. \
    randint(0, 5)) for step in traj]

# llm_handler.py
    ...
```

Runtime Logging

```
Traceback:
  ...
  File "trajectory.py", line 38 ...
    return [step / ...
ZeroDivisionError:
  division by zero
```

Liu et al., 2024. AIGS: Generating Science from AI-Powered Automated Falsification. arXiv:2411.11910.

实验设计

- 选择三种具有挑战性的任务：
  - 数据工程：该任务旨在过滤并提取高质量的数据子集；
  - 自指导对齐：该任务旨在迭代生成指令-响应数据集；
  - 语言建模：该任务旨在调整语言模型的结构和训练参数，以改进预训练效果。
- 评价：围绕证伪、创造性和可执行性对系统进行评估。

Liu et al., 2024. AIGS: Generating Science from AI-Powered Automated Falsification. arXiv:2411.11910.

实验结果：人工评价证伪过程

清華大學 Tsinghua University

- **Baby-AIGS能够通过证伪生成有效的科学发现。**

Metric	AVG	STD	P-Value	MIN	MAX
Importance Score ($0 \sim 2$)					
BABY-AIGS (Ours)	1.80	0.41	0.02	0.00	**2.00**
Top Conference	**2.00**	0.00	—	**2.00**	**2.00**
Consistency Score ($0 \sim 2$)					
BABY-AIGS (Ours)	1.00	0.86	0.00	0.00	**2.00**
Top Conference	**2.00**	0.00	—	**2.00**	**2.00**
Correctness Score ($0 \sim 2$)					
BABY-AIGS (Ours)	0.95	0.94	0.00	0.00	**2.00**
Top Conference	**2.00**	0.00	—	**2.00**	**2.00**
Overall Score ($0 \sim 2$)					
BABY-AIGS (Ours)	1.25	0.47	0.00	0.67	**2.00**
Top Conference	**2.00**	0.00	—	**2.00**	**2.00**

- **最优表现：系统能够识别可能与科学发现相关的重要因素并进行自主证伪。**
- **平均表现：系统的证伪过程显著低于顶级会议中现有文献的满意程度。**

Liu et al., 2024. AIGS: Generating Science from AI-Powered Automated Falsification. arXiv:2411.11910.

实验结果：通过基准测试评估创造力

- **Baby-AIGS在想法生成和相应方法设计上优于基线方法。**

Method	MT-Bench ↑	
	15-shot ICL	SFT
Baseline (Turn 0)	4.18	4.53
AI Scientist	4.36	4.67
BABY-AIGS (Ours)	**4.51**	4.77
Top Conference	4.45	**5.01**

Methodology Summarization (Data Engineering)

1. Rate the response based on its contextual coherence, ensuring it logically follows the conversation.
2. Evaluate the relevance by checking if the answer stays on-topic with minimal digression.
3. Check for logical reasoning in explanations, ensuring the response is not just factual but also thoughtful.
4. Consider if the complexity and detail match the question's requirements, avoiding oversimplification.
5. Finally, evaluate the tone for politeness, clarity, and natural conversational flow.

- **Baby-AIGS 相比基线系统表现出优势，证明了丰富的反馈机制更能有效激发模型的创造力。**
- **目前 Baby-AIGS 的结果不及顶会论文水准。**

Liu et al., 2024. AIGS: Generating Science from AI-Powered Automated Falsification. arXiv:2411.11910.

实验结果：通过基准测试评估创造力

- **其他两项任务上的结果如下，与基线系统相比具有优势。**

Method	MT-Bench ↑
Baseline (Turn 0)	2.45
BABY-AIGS (Ours)	**3.26**

Methodology Summarization (Self-Instruct Alignment)

Make the instruction to cover different scenarios if it lacks specificity, clearer if ambiguous, aligned with natural conversations, and to contain a diverse range of task types if it lacks variety.

Method	Perplexity ↓		
	shakespeare_char	enwik8	text8
Baseline (Turn 0)	**1.473**	1.003	0.974
BABY-AIGS (Ours)	1.499	**0.984**	**0.966**

Methodology Summarization (Language Modeling)

Reduce the dropout rate with more attention heads to increase model expressiveness. And implement a cyclical learning rate and adjust the weight decay to regularize the model.

Liu et al., 2024. AIGS: Generating Science from AI-Powered Automated Falsification. arXiv:2411.11910.

实验结果：通过成功率比较评估可执行性

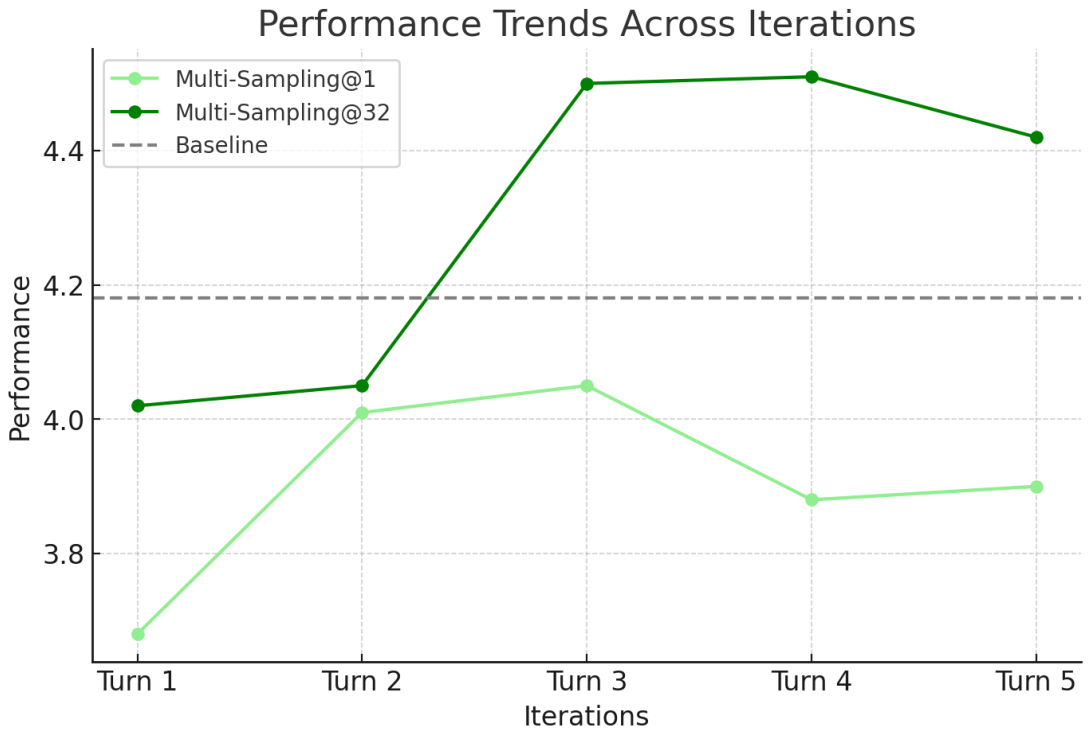
- **Baby-AIGS 在可执行性上显著优于现有系统。**

Method	Experiment Success Rate (Exp. SR)	Overall Success Rate (Overall SR)
AI Scientist	44.8%	29.2%
Baby-AIGS (Ours)	**Almost 100%**	**Almost 100%**

- **定量分析表明，Baby-AIGS 的可执行性优于基线系统。**
- **目前系统将生成的想法转化为实验结果及最终科学发现的成功率接近100%。**
- **这一高可执行性归因于我们引入的领域特定语言（DSL）。**

实验分析：多采样帮助提高创造力

- **Baby-AIGS 采取了多采样的方法来提高模型的创造力。**

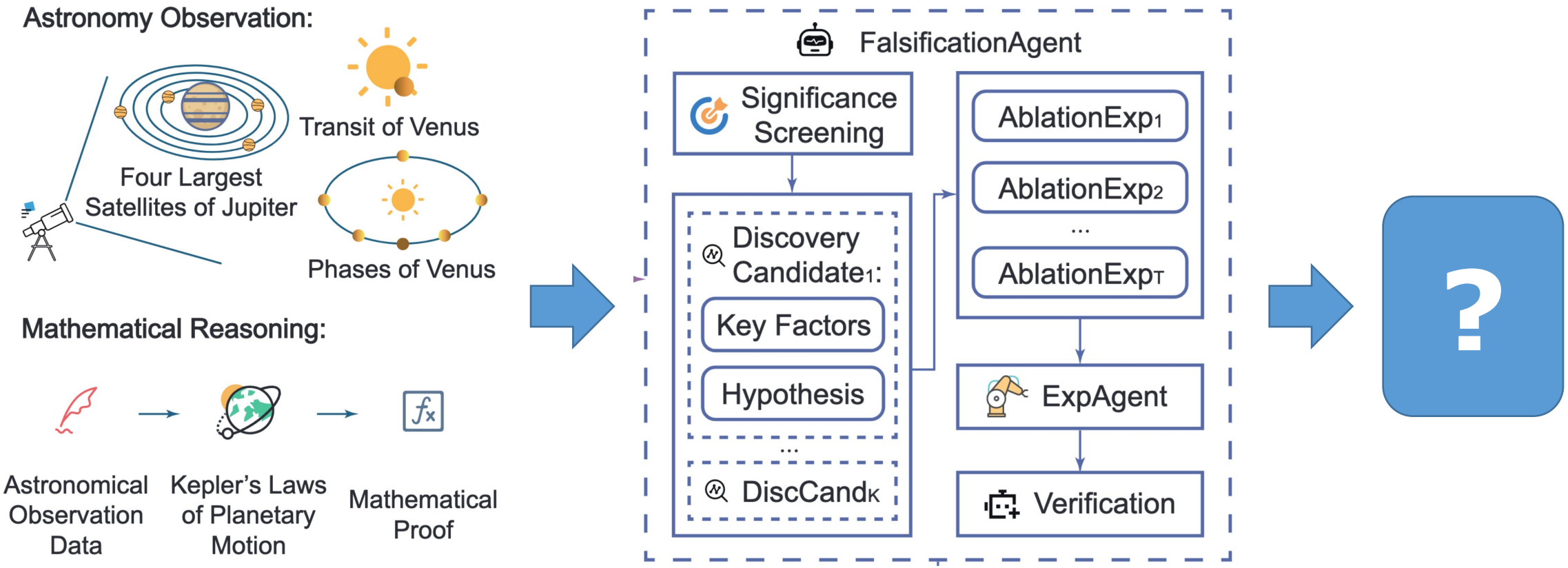


- **一种基于搜索的 scaling inference compute 方法。**
- **更好的创造力表现一定程度上归功于多采样搜索。**

Liu et al., 2024. AIGS: Generating Science from AI-Powered Automated Falsification. arXiv:2411.11910.

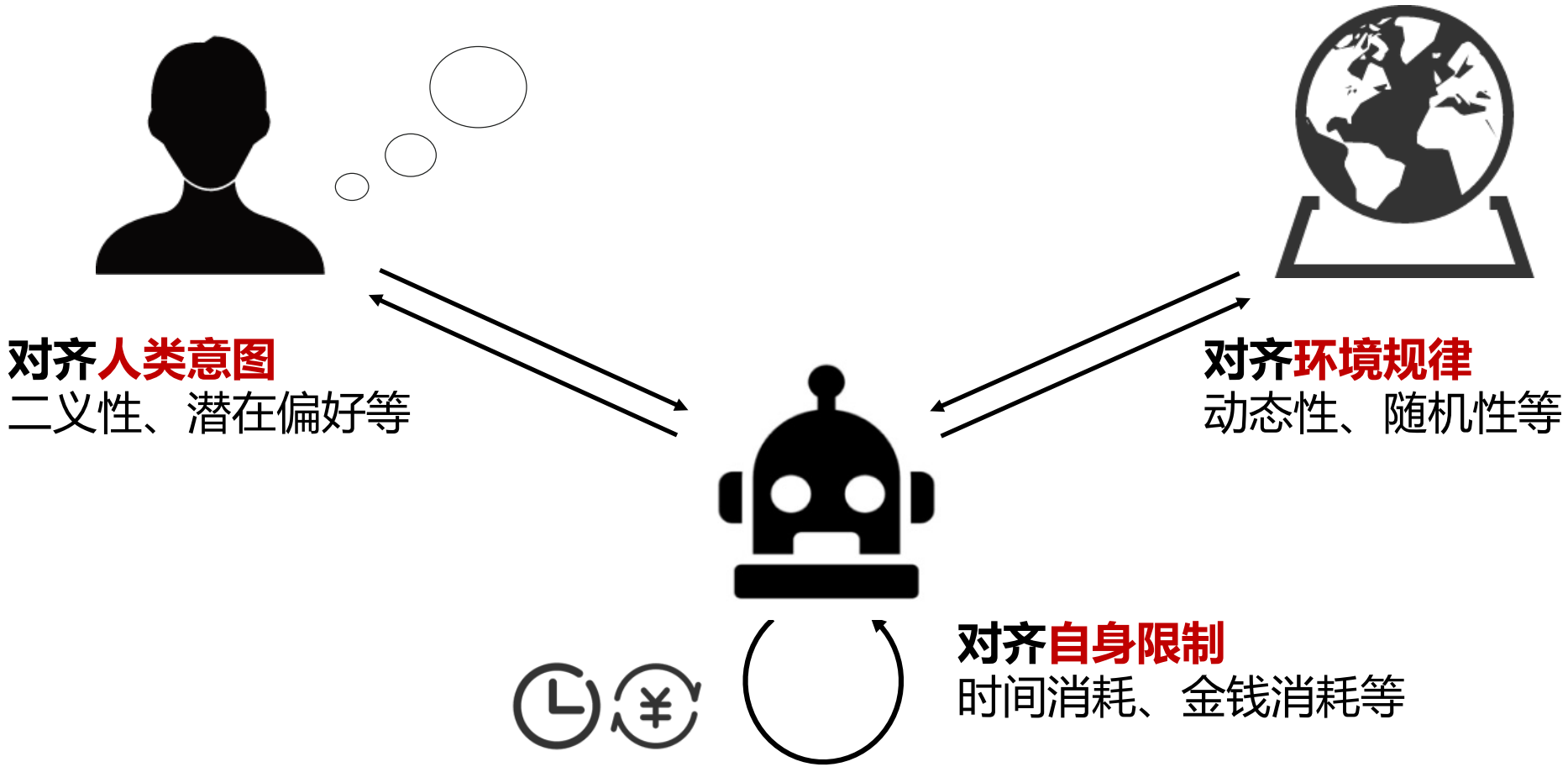
展望：需要对AI自主科学本身进行证伪

- 我们需要进一步设计针对 AI 科研系统的证伪流程。



展望：AIGS 需要遵循 UA² 对齐原则

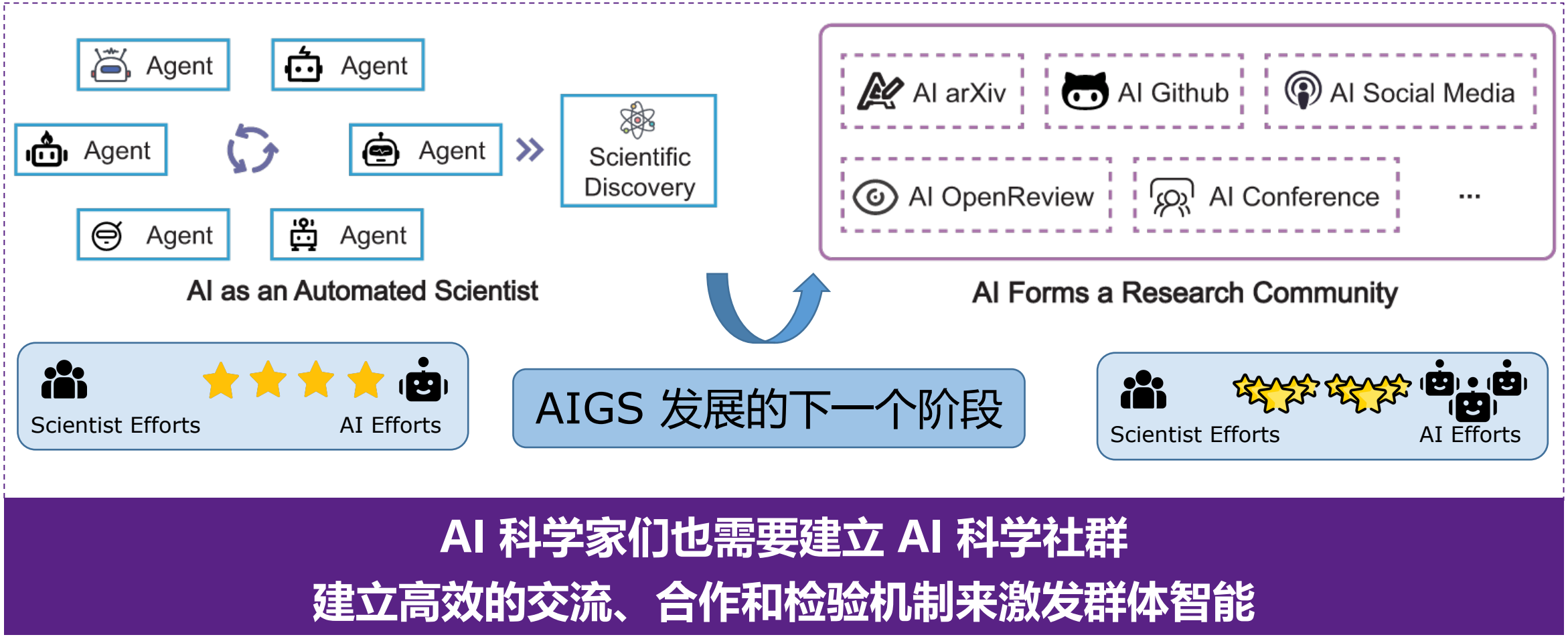
- AIGS 智能体需要和自身、人类、环境统一对齐。



Yang et al., 2024. Towards Unified Alignment Between Agents, Humans, and Environment. ICML 2024.

展望：建立 AI 科学家社群

- 建立AI科学家社群，催化产生跨学科重大科学发展。



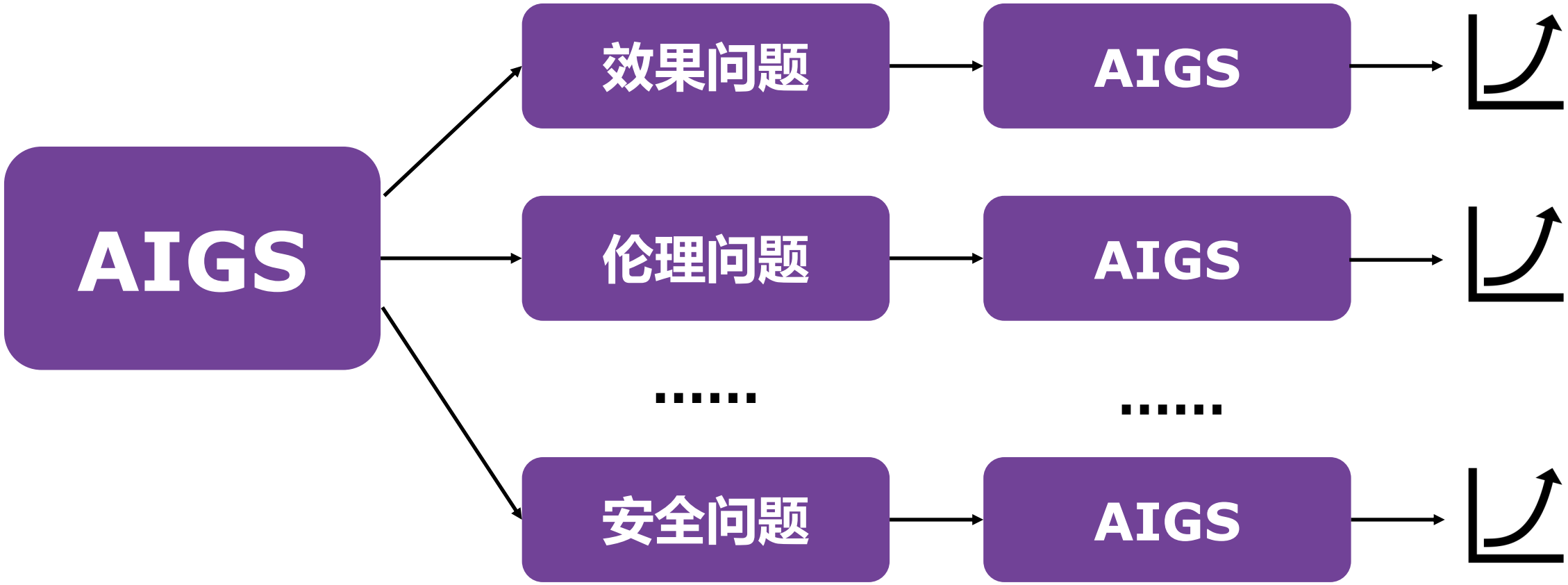
AI 科学家们也需要建立 AI 科学社群

建立高效的交流、合作和检验机制来激发群体智能

Liu et al., 2024. AIGS: Generating Science from AI-Powered Automated Falsification. arXiv:2411.11910.

展望：AIGS for AIGS

- 利用AIGS来提升AIGS、解决AIGS自身面临的挑战。



Liu et al., 2024. AIGS: Generating Science from AI-Powered Automated Falsification. arXiv:2411.11910.

主要成员

清華大學 Tsinghua University

李鹏

刘子君

刘铠铭

朱奕祺

刘洋

雷轩宇

杨宗瀚

张真赫

总结

- 提出AIGS概念，即AI全自主作出的符合科学标准的科学发现。
- 实现AIGS的核心原则是保证系统具备证伪能力、创造性和可执行性：
  - 证伪 (Falsification) 是科学研究的核心；
  - 创造性 (Creativity) 的想法是科学研究的起点；
  - 可执行性 (Executability) 构成了证伪的基础。

扫码关注AIGS进展

- 实现Baby-AIGS系统，具备自主证伪能力，初步验证AIGS的可行性。
- 展望：未来符合AIGS定义的自主科研系统将成为推动科学发展重要力量。

谢谢