



# AIGS: 全自主AI科学发现探索

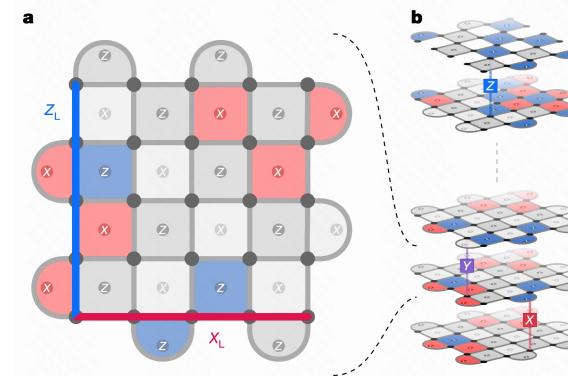
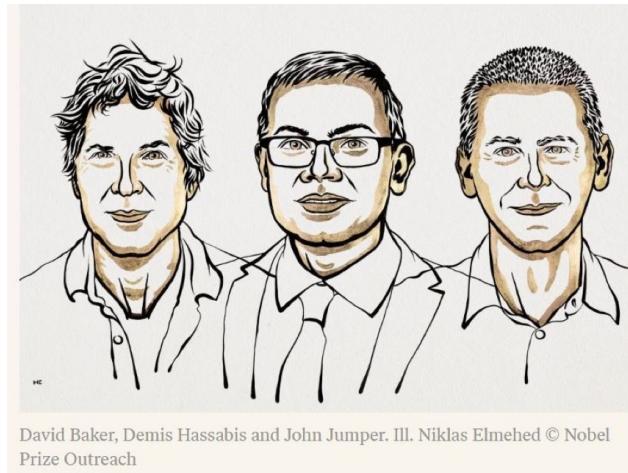
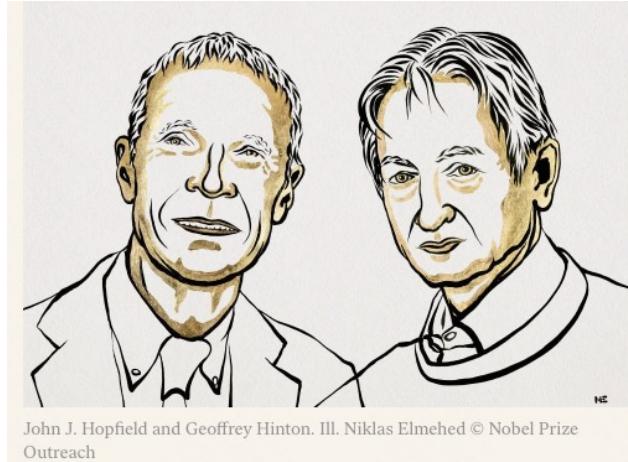
李鹏  
清华大学

# AI深刻改变科学的研究范式



清华大学  
Tsinghua University

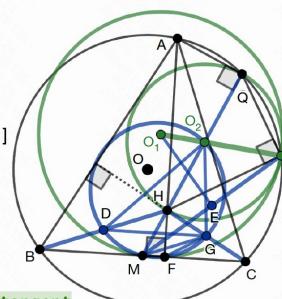
- 深度学习的发展赋能AI4Science，极大地推动了科学的发展。



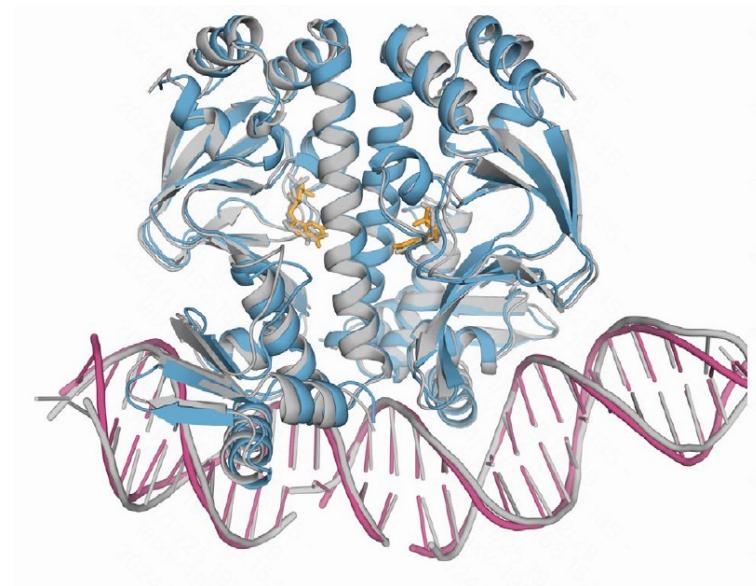
AlphaQuant

Solution

**Construct D: midpoint BH [a]**  
[a],  $O_2$  midpoint HQ  $\Rightarrow BQ \parallel O_2D$  [20]  
...  
**Construct G: midpoint HC [b]** ...  
 $\angle GMD = \angle GO_2D \Rightarrow M O_2 G D$  cyclic [26]  
...  
[a], [b]  $\Rightarrow BC \parallel DG$  [30]  
...  
**Construct E: midpoint MK [c]**  
..., [c]  $\Rightarrow \angle KFC = \angle KO_1E$  [104]  
...  
 $\angle FKO_1 = \angle FKO_2 \Rightarrow K_1 \parallel K_2$  [109]  
[109]  $\Rightarrow O_1 O_2$  collinear  $\Rightarrow (O_1)(O_2)$  tangent



AlphaGeometry



<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>

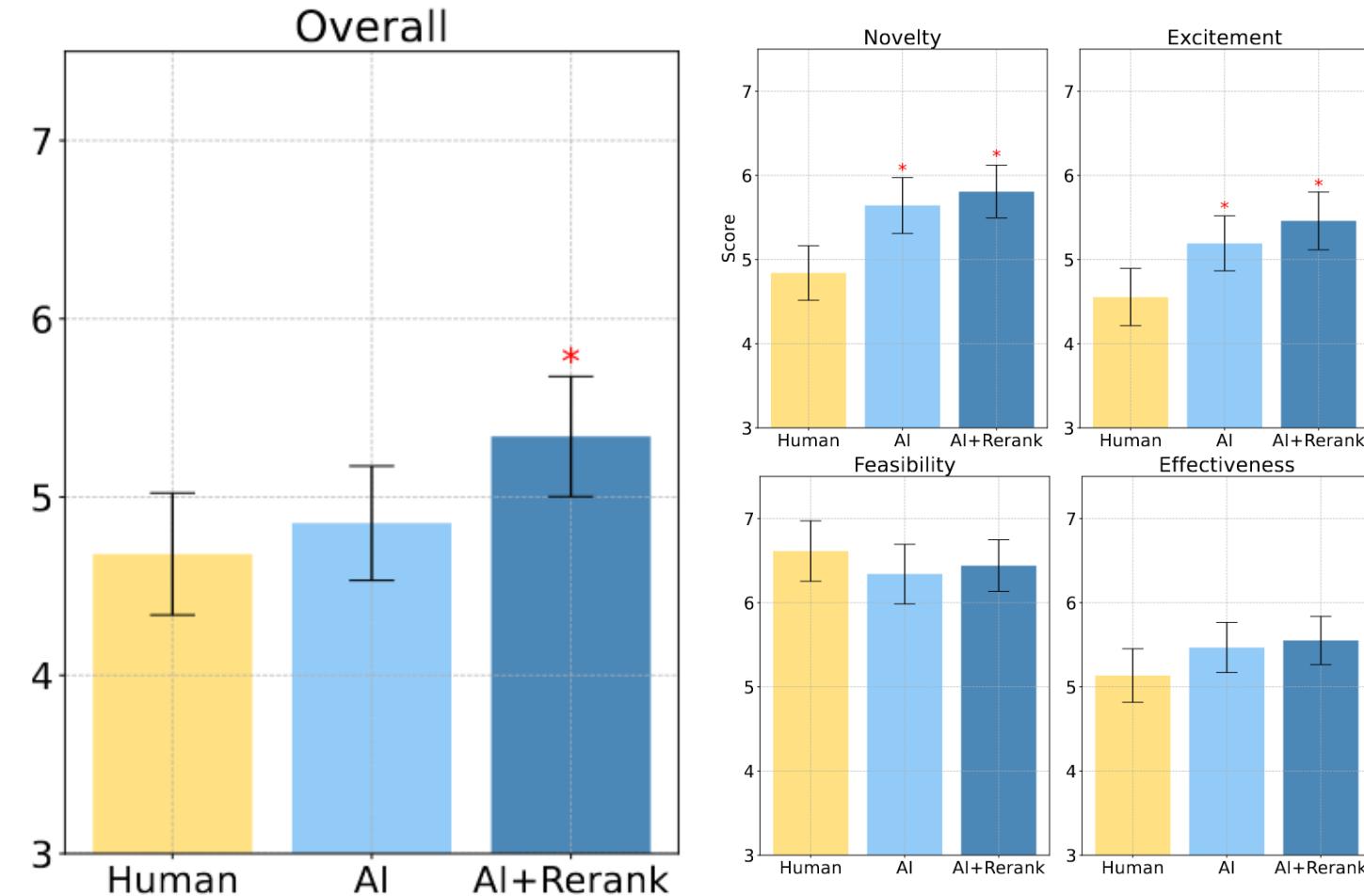
# AI展现出自主科学研究潜力



## ● 如何利用AI自主作出科学发现受到广泛关注，已展现出可观潜力。

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

- 考核了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

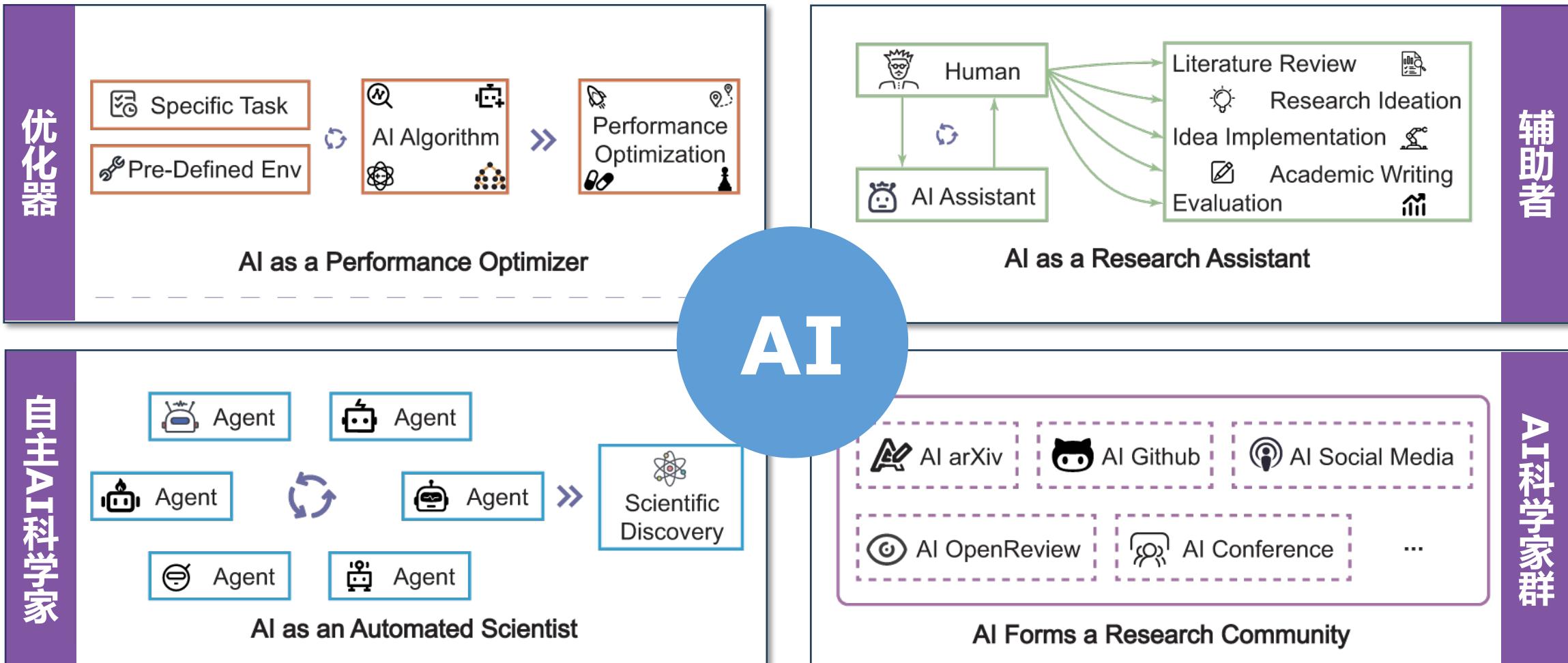


# AI参与科研的范式



清华大学  
Tsinghua University

- 根据AI在科研中的角色可以划分出 4 种AI参与科学的研究的范式。

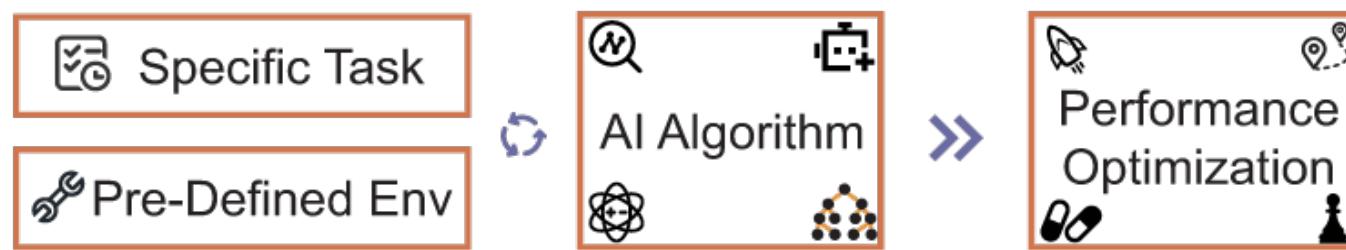


# 范式一：优化器

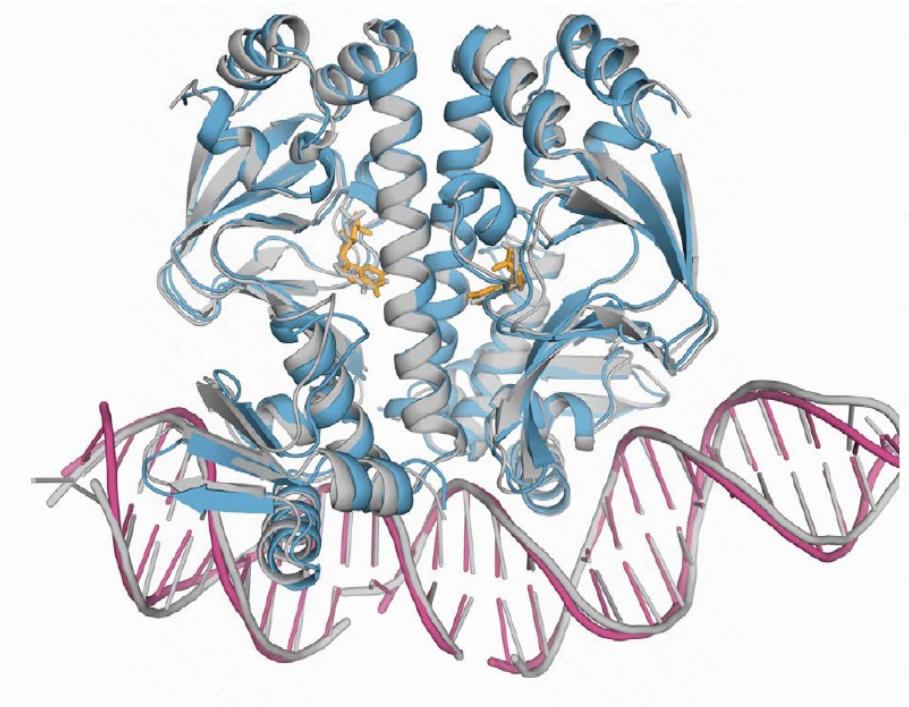
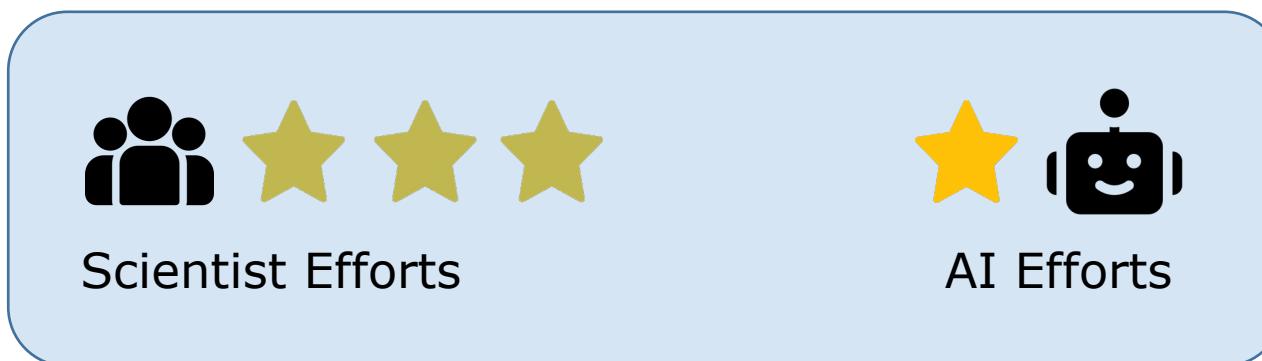


清华大学  
Tsinghua University

- AI在科研中的角色：专注于优化特定任务的表现。



AI as a Performance Optimizer

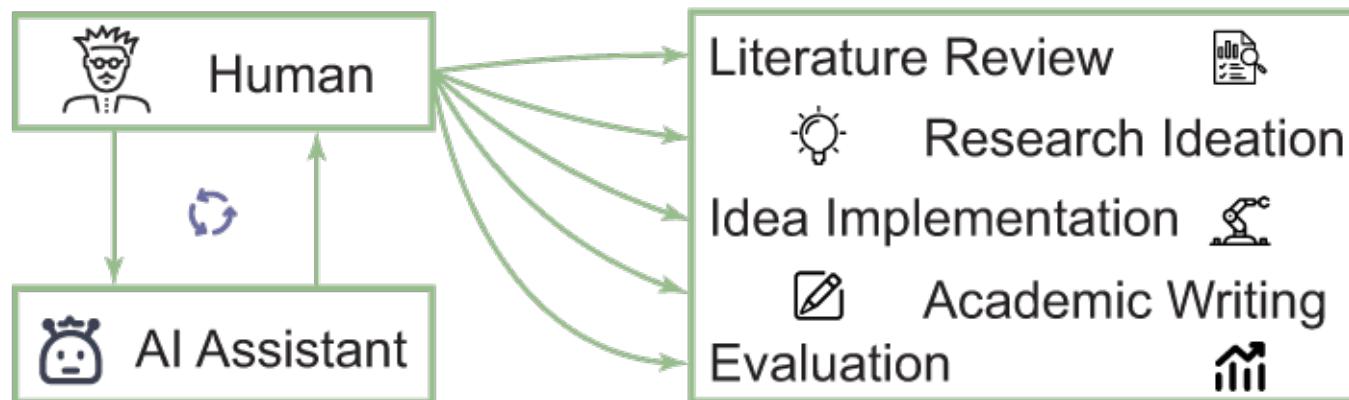


AlphaFold

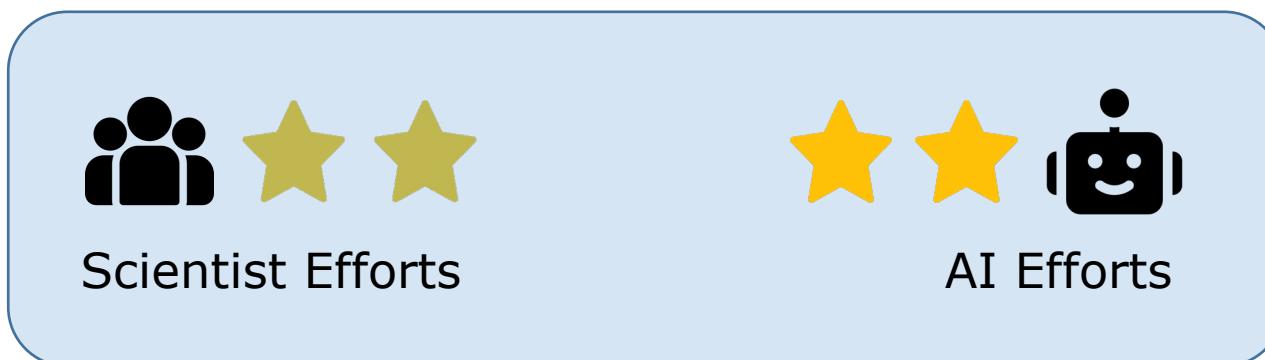
# 范式二：辅助者



● 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



```
1 import datetime
2
3 def parse_expenses(expenses_string):
4     """Parse the list of expenses and return the list of triples (date, amount, currency)
5     Ignore lines starting with #.
6     Parse the date using datetime.
7     Example expenses_string:
8         2023-01-02 -34.01 USD
9         2023-01-03 2.59 DKK
10        2023-01-03 -2.72 EUR
11
12    expenses = []
13
14    for line in expenses_string.splitlines():
15        if line.startswith("#"):
16            continue
17        date, value, currency = line.split(" ")
18        expenses.append((datetime.datetime.strptime(date, "%Y-%m-%d"),
19                         float(value),
20                         currency))
21    return expenses
22
23 expenses_data = '''2023-01-02 -34.01 USD
24 2023-01-03 2.59 DKK
25 2023-01-03 -2.72 EUR'''
```

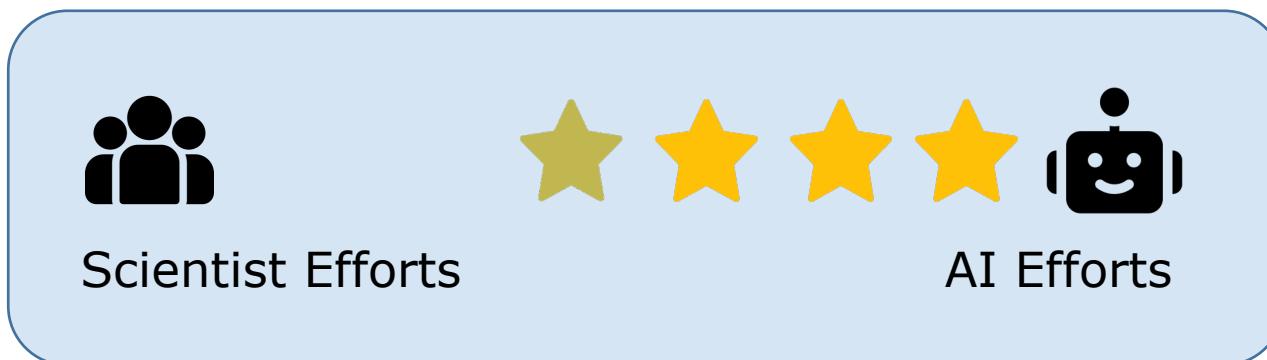
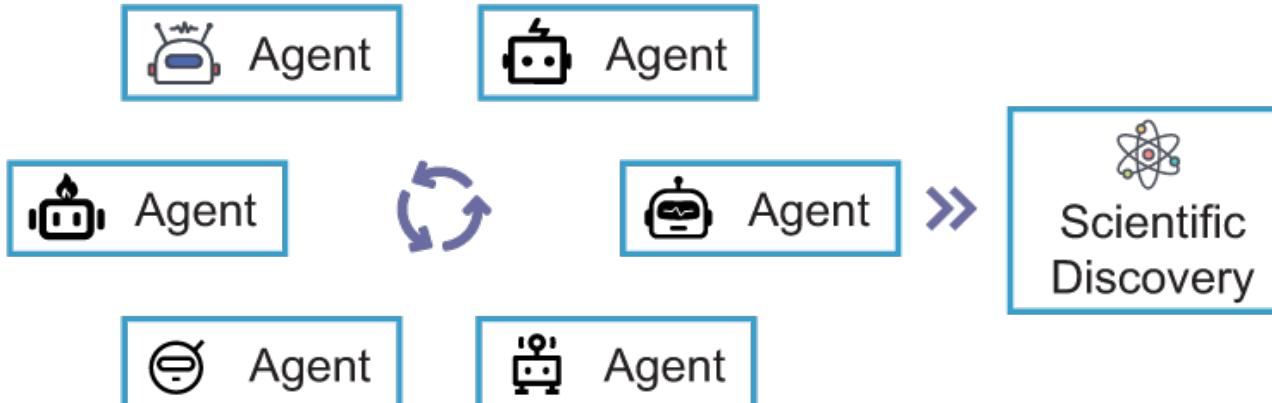
Github Copilot

# 范式三：自主AI科学家



清华大学  
Tsinghua University

- 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 in different types of low-dimensional spaces, 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.

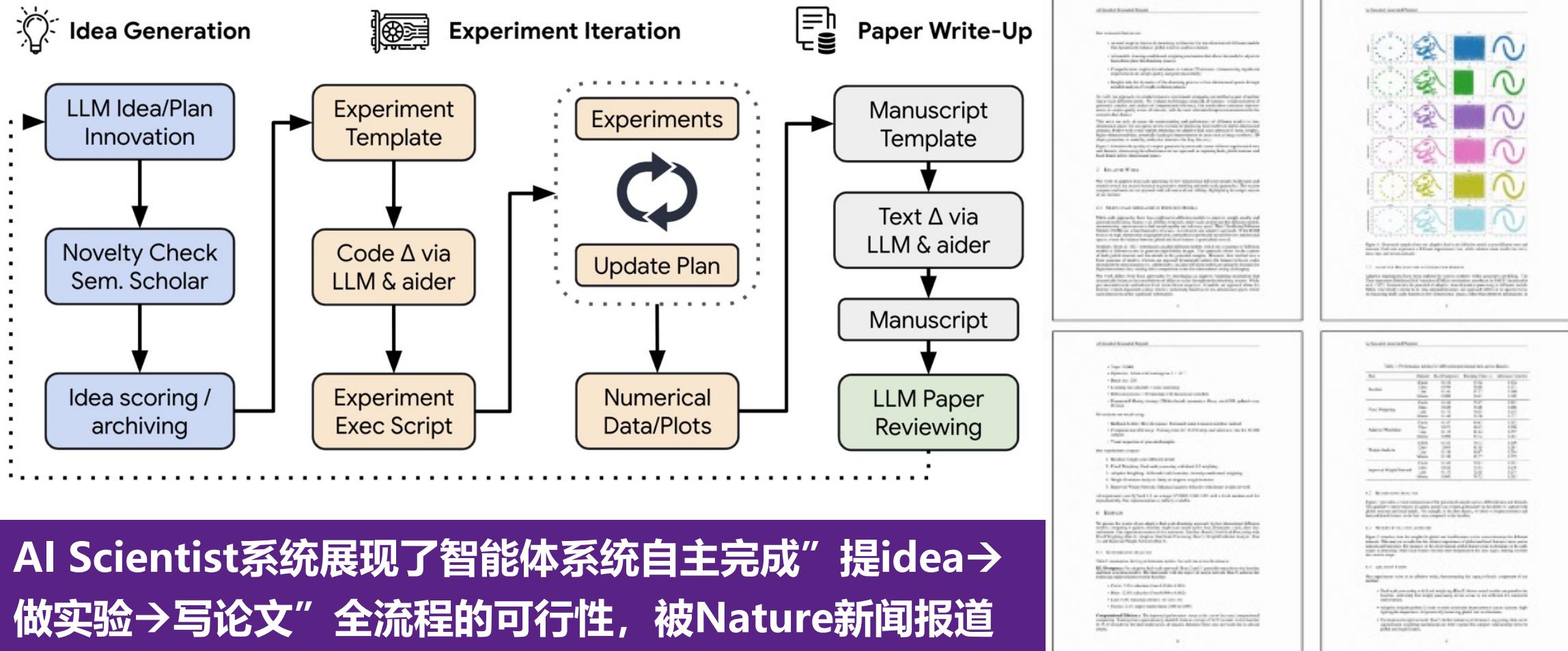
AI Generated Paper

# 范式三的初步探索



清华大学  
Tsinghua University

- 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自主科学研究充满挑战



## ● Case 分析：AI-Scientist 产生的研究学术价值较低。

UNLOCKING GROKKING: A COMPARATIVE STUDY OF WEIGHT INITIALIZATION STRATEGIES IN TRANSFORMER MODELS

We systematically compare five initialization methods (PyTorch default, Xavier, He, Orthogonal, and Kaiming Normal) across four arithmetic tasks in finite fields

difficulty generalize after prolonged training, remains poorly understood, hindering the development of efficient training strategies. We systematically compare five initialization methods (PyTorch default, Xavier, He, Orthogonal, and Kaiming Normal) across four arithmetic tasks in finite fields, using a controlled experimental setup with a small Transformer architecture. Our approach combines rigorous

Now, let's implement the He initialization for Run 2. We'll modify the `\_\_init\_weights` method in the `Transformer` class:

```
experiment.py
```python
<<<<< SEARCH
def __init_weights(self):
    for module in self.modules():
        if isinstance(module, (nn.Linear, nn.Embedding)):
            nn.init.xavier_uniform_(module.weight)
        elif isinstance(module, nn.LayerNorm):
            nn.init.constant_(module.weight, 1.0)
        ...

>>>> REPLACE
...
if isinstance(module, (nn.Linear, nn.Embedding)):
    nn.init.xavier_uniform_(module.weight)
elif isinstance(module, nn.LayerNorm):
    nn.init.constant_(module.weight, 1.0)
    nn.init.constant_(module.bias, 0.0)
```

- 表面上看是研究 Grokking，其实只是更换了模型参数初始化的方法
- 实际的代码修改仅是局部对于初始化函数调用的更换



## ● Case 分析：AI-Scientist 产生的实验设计可执行性较差。

### Algorithm 1 Experimental Procedure

```
1: for each task  $T_i \in \mathcal{T}$  do
2:   for each initialization strategy  $S_j \in \mathcal{S}$  do
3:     for  $k = 1$  to 3 do                                ▷ Three runs per configuration
4:       Initialize  $\theta_0 \sim P_j(\theta)$ 
```

### Compare performance of different initialization strategies for task $T_i$

```
9:   Calculate  $S_{99}$  for this run
10:  end for
11:  Compute mean and standard error of metrics across the three runs
12: end for
13: Compare performance of different initialization strategies for task  $T_i$ 
14: end for
15: Analyze trends and patterns across tasks and initialization strategies
```

## ● 方法伪代码中包含模糊不清的“分析”和“比较”等操作，可执行性不足

# AI自主科学研究充满挑战



清华大学  
Tsinghua University

## ● Case 分析：虽然AI-Scientist 产生的论文图文并茂，但难于被信任。

AI Scientist Generated Paper

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 and local features in the denoising process. By separating the denoising process into two parallel paths, one focusing on global features and another on local features, our model can effectively capture both types of features simultaneously. We propose a novel adaptive weighting mechanism that allows the model to adjust its focus throughout the denoising process.

Our main contributions are:

- A novel adaptive dual-scale denoising architecture for low-dimensional diffusion models that can effectively capture both global and local features.
- A flexible, iterative-conditioned weighting mechanism that allows the model to adjust its focus throughout the denoising process.
- Competitive empirical evaluations on various 2D datasets, demonstrating significant improvements over state-of-the-art methods.
- Insights into the dynamics of the denoising process in low-dimensional space through detailed analysis of weight evolution metrics.

AI Scientist Generated Paper

Our main contributions are:

- A novel adaptive dual-scale denoising architecture for low-dimensional diffusion models that can effectively capture both global and local features.
- A flexible, iterative-conditioned weighting mechanism that allows the model to adjust its focus throughout the denoising process.
- Competitive empirical evaluations on various 2D datasets, demonstrating significant improvements over state-of-the-art methods.
- Insights into the dynamics of the denoising process in low-dimensional space through detailed analysis of weight evolution metrics.

AI Scientist Generated Paper

AI Scientist Generated Paper

In this experiment, we use the MNIST dataset and a 2D diffusion model to demonstrate the performance of our proposed dual-scale denoising approach. Our results show that our method achieves better denoising performance than baseline methods, particularly for noisy images.

Our main contributions are:

- An adaptive dual-scale denoising architecture for low-dimensional diffusion models that can effectively capture both global and local features.
- A flexible, iterative-conditioned weighting mechanism that allows the model to adjust its focus throughout the denoising process.
- Competitive empirical evaluations on various 2D datasets, demonstrating significant improvements over state-of-the-art methods.
- Insights into the dynamics of the denoising process in low-dimensional space through detailed analysis of weight evolution metrics.

AI Scientist Generated Paper

We focus on adaptive diffusion models for 2D datasets. Let  $\mathcal{X} \subset \mathbb{R}^d$  be the data space, and  $\mathcal{F}_{\text{diff}}(\mathcal{X})$  be the space of diffusion models over  $\mathcal{X}$ . Our goal is to learn a generative model that samples from a distribution  $p_{\text{model}}(\cdot | \mathcal{X})$  closely approximating  $p_{\text{data}}(\cdot | \mathcal{X})$ .

The diffusion process is defined by a sequence of latent variables  $x_0, \dots, x_T$ , where  $x_0 \sim p_{\text{init}}(\cdot)$  and  $x_T = \mathcal{X}$  are the sequence of randomly noise vectors of  $\mathcal{X}$ . The forward process is defined as:

$$(x_t)_{t=1}^T = f(x_{t-1}) \sqrt{1 - \alpha_t} + \bar{x}_t, \quad (1)$$

where  $\alpha_t = 1 - \beta_t$  is the noise schedule.

The reverse process, parameterized by a neural network  $\phi$ , is defined as:

$$p(x_{t-1} | x_t) = \mathcal{N}(x_{t-1}, \phi(x_t, \beta_t)), \quad (2)$$

In low-dimensional space, each dimension carries significant information about the overall structure of the data. This presents a unique challenge: the model must simultaneously update both global features and local features to maintain balance. Our proposed adaptive dual-scale denoising approach does not specifically address the challenge of balancing global structure and local details, which is the primary focus of this work.

Our approach extends this line of research by introducing an adaptive dual-scale method specifically designed for low-dimensional data. Unlike existing methods such as DualScaleDiffusion or TADDM, which use a standard diffusion model architecture, our method explicitly models the relationship between global and local features through dual-scale architecture and adaptive weighting mechanisms.

In summary, our adaptive dual-scale denoising approach for low-dimensional diffusion models addresses the challenge of balancing global and local features in low-dimensional space. It is particularly well-suited to the challenges of low-dimensional data, such as images and 3D point clouds. Our results demonstrate that our approach outperforms state-of-the-art dual-scale approaches and makes it particularly suited for capturing complex low-dimensional data features. We believe that this work will contribute to the broader field of generative modeling and encourage others to continue exploring the effectiveness of our approach in this specific problem setting.

AI Scientist Generated Paper

where  $W \in \mathbb{R}^{n \times d}$  represents the weight for the global and local features at timestamp  $t$ . The weight network  $\Phi$  is implemented as:

$$\text{MLP}_{\Phi}(x_t) = \text{Linear}(\text{LayerNorm}(\text{Dropout}(\text{MLP}(\text{MLP}_{\Phi}(x_t))))) \quad (7)$$

This design allows for complex weight computation, enabling smooth adaptation of the global and local features across different dimensions. The use of LayerNorm and residual block further improves numerical stability and training efficiency. In addition, we propose a learning rate scheduler for the weight network to ensure balanced learning.

### 4.3. COMBINED DENOISING PROCESS

The final denoising process is a weighted combination of the global and local features:

$$x_t = (1 - w_t) \cdot g_t(x_{t-1}) + w_t \cdot l_t(x_{t-1}), \quad (8)$$

where  $w_t \in [0, 1]$  are the components of  $W$ . The combination allows the model to leverage both global and local features in its predictions, with the balance dynamically adjusted based on the current iteration.

### 4.4. TRAINING PROCESS

We train our model using the same objective as in the original DDPM He et al. (2020):

$$\mathcal{L} = \mathbb{E}_{(x, t)} [\|x_t - \phi(x_{t-1}, \beta_t)\|^2] \quad (9)$$

# “它是由AI产生的”是影响您判断的根本性因素吗？

Figure 2: Evolution of global and local feature weights across iterations for different datasets. The x-axis represents iterations (0 to 20) and the y-axis represents the weight for global (red) and local (blue) features at the start (iteration 0) and end (iteration 20) of the denoising process. The colors indicate the weight magnitude, with red for global and blue for local features.

Figure 3: Histograms showing the weight for global and local features across iterations for different datasets. The x-axis represents the weight value and the y-axis represents the frequency. The figure shows four histograms: Circle (red), Line (red), Line (blue), and Line (global). All histograms are centered around 0.5, indicating that the weights are balanced across iterations.

RESULTS

We evaluate our adaptive dual-scale denoising approach on four 2D datasets: circle, line, and spiral. The results show that our approach outperforms state-of-the-art denoising methods across all three datasets.

EXPERIMENTAL SETUP

We evaluate our adaptive dual-scale denoising approach on four 2D datasets: circle, line, and spiral. The results show that our approach outperforms state-of-the-art denoising methods across all three datasets.

Table 1 summarizes the key performance metrics for each of the datasets.

**K1 Strength:** Our adaptive dual-scale denoising approach (Line) achieves higher K1 strength compared to the adaptive weight network (Line) that is achieved by the following improvements over the baseline:

- Circle: 2.8% improvement (Line vs. 0.54)
- Line: 1.0% improvement (Line vs. 0.54)
- Line: 0.9% improvement (Line vs. 0.55)
- Line: 3.7% improvement (Line vs. 0.60)

**Quantitative Analysis:** The improved performance comes at the cost of increased computational complexity. Our model architecture requires a bit more memory and computation time compared to the baseline.

**Qualitative Analysis:** Our experiments serve as an ablation study, demonstrating the impact of each component of our model.

- Circle: A simple denoising model.
- Line: A model with both global and local features.
- Line: A dual-scale structure.
- Line: Our dual-scale denoising approach.

Our model architecture is implemented in PyTorch 1.9.0.

**Training parameters:**

1. Baseline: High-scale diffusion model

2. Fixed Weighting: Dual-scale processing with fixed 0.5 weighting

3. Adaptive Weighting: Dual-scale with iterative, temporally-conditioned weighting

4. Weight Network Approach: Dual-scale with adaptive weight network

All experiments run PyTorch 1.9 on a single NVIDIA V100 GPU with a batch size of 64 for 100 epochs. Our implementation is publicly available.

RESULTS

We present the results of our adaptive dual-scale denoising approach for low-dimensional diffusion models, comparing it against a baseline single-scale model across four 2D datasets: circle, line, spiral, and circle. The results show that our adaptive dual-scale denoising approach significantly outperforms the baseline model across all datasets, particularly for the spiral dataset.

EXPERIMENTAL SETUP

We evaluate our adaptive dual-scale denoising approach on four 2D datasets: circle, line, and spiral. The results show that our approach outperforms state-of-the-art denoising methods across all three datasets.

Table 1 summarizes the key performance metrics for each of the datasets.

**K1 Strength:** Our adaptive dual-scale denoising approach (Line) achieves higher K1 strength compared to the adaptive weight network (Line) that is achieved by the following improvements over the baseline:

- Circle: 2.8% improvement (Line vs. 0.54)
- Line: 1.0% improvement (Line vs. 0.54)
- Line: 0.9% improvement (Line vs. 0.55)
- Line: 3.7% improvement (Line vs. 0.60)

**Quantitative Efficiency:** The improved performance comes at the cost of increased computational complexity. Our model architecture requires a bit more memory and computation time compared to the baseline.

**Qualitative Analysis:** Our experiments serve as an ablation study, demonstrating the impact of each component of our model.

- Circle: A simple denoising model.
- Line: A model with both global and local features.
- Line: A dual-scale structure.
- Line: Our dual-scale denoising approach.

Our model architecture is implemented in PyTorch 1.9.0.

**Training parameters:**

Figure 2: Evolution of global and local feature weights across iterations for different datasets. The x-axis represents iterations (0 to 20) and the y-axis represents the weight for global (red) and local (blue) features at the start (iteration 0) and end (iteration 20) of the denoising process. The colors indicate the weight magnitude, with red for global and blue for local features.

Figure 3: Histograms showing the weight for global and local features across iterations for different datasets. The x-axis represents the weight value and the y-axis represents the frequency. The figure shows four histograms: Circle (red), Line (red), Line (blue), and Line (global). All histograms are centered around 0.5, indicating that the weights are balanced across iterations.

RESULTS

We evaluate our adaptive dual-scale denoising approach on four 2D datasets: circle, line, and spiral. The results show that our approach outperforms state-of-the-art denoising methods across all three datasets.

EXPERIMENTAL SETUP

We evaluate our adaptive dual-scale denoising approach on four 2D datasets: circle, line, and spiral. The results show that our approach outperforms state-of-the-art denoising methods across all three datasets.

Table 1 summarizes the key performance metrics for each of the datasets.

**K1 Strength:** Our adaptive dual-scale denoising approach (Line) achieves higher K1 strength compared to the adaptive weight network (Line) that is achieved by the following improvements over the baseline:

- Circle: 2.8% improvement (Line vs. 0.54)
- Line: 1.0% improvement (Line vs. 0.54)
- Line: 0.9% improvement (Line vs. 0.55)
- Line: 3.7% improvement (Line vs. 0.60)

**Quantitative Efficiency:** The improved performance comes at the cost of increased computational complexity. Our model architecture requires a bit more memory and computation time compared to the baseline.

**Qualitative Analysis:** Our experiments serve as an ablation study, demonstrating the impact of each component of our model.

- Circle: A simple denoising model.
- Line: A model with both global and local features.
- Line: A dual-scale structure.
- Line: Our dual-scale denoising approach.

Our model architecture is implemented in PyTorch 1.9.0.

**Training parameters:**

Figure 2: Evolution of global and local feature weights across iterations for different datasets. The x-axis represents iterations (0 to 20) and the y-axis represents the weight for global (red) and local (blue) features at the start (iteration 0) and end (iteration 20) of the denoising process. The colors indicate the weight magnitude, with red for global and blue for local features.

Figure 3: Histograms showing the weight for global and local features across iterations for different datasets. The x-axis represents the weight value and the y-axis represents the frequency. The figure shows four histograms: Circle (red), Line (red), Line (blue), and Line (global). All histograms are centered around 0.5, indicating that the weights are balanced across iterations.

RESULTS

We evaluate our adaptive dual-scale denoising approach on four 2D datasets: circle, line, and spiral. The results show that our approach outperforms state-of-the-art denoising methods across all three datasets.

EXPERIMENTAL SETUP

We evaluate our adaptive dual-scale denoising approach on four 2D datasets: circle, line, and spiral. The results show that our approach outperforms state-of-the-art denoising methods across all three datasets.

Table 1 summarizes the key performance metrics for each of the datasets.

**K1 Strength:** Our adaptive dual-scale denoising approach (Line) achieves higher K1 strength compared to the adaptive weight network (Line) that is achieved by the following improvements over the baseline:

- Circle: 2.8% improvement (Line vs. 0.54)
- Line: 1.0% improvement (Line vs. 0.54)
- Line: 0.9% improvement (Line vs. 0.55)
- Line: 3.7% improvement (Line vs. 0.60)

**Quantitative Efficiency:** The improved performance comes at the cost of increased computational complexity. Our model architecture requires a bit more memory and computation time compared to the baseline.

**Qualitative Analysis:** Our experiments serve as an ablation study, demonstrating the impact of each component of our model.

- Circle: A simple denoising model.
- Line: A model with both global and local features.
- Line: A dual-scale structure.
- Line: Our dual-scale denoising approach.

Our model architecture is implemented in PyTorch 1.9.0.

**Training parameters:**

Figure 2: Evolution of global and local feature weights across iterations for different datasets. The x-axis represents iterations (0 to 20) and the y-axis represents the weight for global (red) and local (blue) features at the start (iteration 0) and end (iteration 20) of the denoising process. The colors indicate the weight magnitude, with red for global and blue for local features.

Figure 3: Histograms showing the weight for global and local features across iterations for different datasets. The x-axis represents the weight value and the y-axis represents the frequency. The figure shows four histograms: Circle (red), Line (red), Line (blue), and Line (global). All histograms are centered around 0.5, indicating that the weights are balanced across iterations.

RESULTS

We evaluate our adaptive dual-scale denoising approach on four 2D datasets: circle, line, and spiral. The results show that our approach outperforms state-of-the-art denoising methods across all three datasets.

EXPERIMENTAL SETUP

We evaluate our adaptive dual-scale denoising approach on four 2D datasets: circle, line, and spiral. The results show that our approach outperforms state-of-the-art denoising methods across all three datasets.

Table 1 summarizes the key performance metrics for each of the datasets.

**K1 Strength:** Our adaptive dual-scale denoising approach (Line) achieves higher K1 strength compared to the adaptive weight network (Line) that is achieved by the following improvements over the baseline:

- Circle: 2.8% improvement (Line vs. 0.54)
- Line: 1.0% improvement (Line vs. 0.54)
- Line: 0.9% improvement (Line vs. 0.55)
- Line: 3.7% improvement (Line vs. 0.60)

**Quantitative Efficiency:** The improved performance comes at the cost of increased computational complexity. Our model architecture requires a bit more memory and computation time compared to the baseline.

**Qualitative Analysis:** Our experiments serve as an ablation study, demonstrating the impact of each component of our model.

- Circle: A simple denoising model.
- Line: A model with both global and local features.
- Line: A dual-scale structure.
- Line: Our dual-scale denoising approach.

Our model architecture is implemented in PyTorch 1.9.0.

**Training parameters:**

Figure 2: Evolution of global and local feature weights across iterations for different datasets. The x-axis represents iterations (0 to 20) and the y-axis represents the weight for global (red) and local (blue) features at the start (iteration 0) and end (iteration 20) of the denoising process. The colors indicate the weight magnitude, with red for global and blue for local features.

Figure 3: Histograms showing the weight for global and local features across iterations for different datasets. The x-axis represents the weight value and the y-axis represents the frequency. The figure shows four histograms: Circle (red), Line (red), Line (blue), and Line (global). All histograms are centered around 0.5, indicating that the weights are balanced across iterations.

RESULTS

We evaluate our adaptive dual-scale denoising approach on four 2D datasets: circle, line, and spiral. The results show that our approach outperforms state-of-the-art denoising methods across all three datasets.

EXPERIMENTAL SETUP

We evaluate our adaptive dual-scale denoising approach on four 2D datasets: circle, line, and spiral. The results show that our approach outperforms state-of-the-art denoising methods across all three datasets.

Table 1 summarizes the key performance metrics for each of the datasets.

**K1 Strength:** Our adaptive dual-scale denoising approach (Line) achieves higher K1 strength compared to the adaptive weight network (Line) that is achieved by the following improvements over the baseline:

- Circle: 2.8% improvement (Line vs. 0.54)
- Line: 1.0% improvement (Line vs. 0.54)
- Line: 0.9% improvement (Line vs. 0.55)
- Line: 3.7% improvement (Line vs. 0.60)

**Quantitative Efficiency:** The improved performance comes at the cost of increased computational complexity. Our model architecture requires a bit more memory and computation time compared to the baseline.

**Qualitative Analysis:** Our experiments serve as an ablation study, demonstrating the impact of each component of our model.

- Circle: A simple denoising model.
- Line: A model with both global and local features.
- Line: A dual-scale structure.
- Line: Our dual-scale denoising approach.

Our model architecture is implemented in PyTorch 1.9.0.

**Training parameters:**

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

11

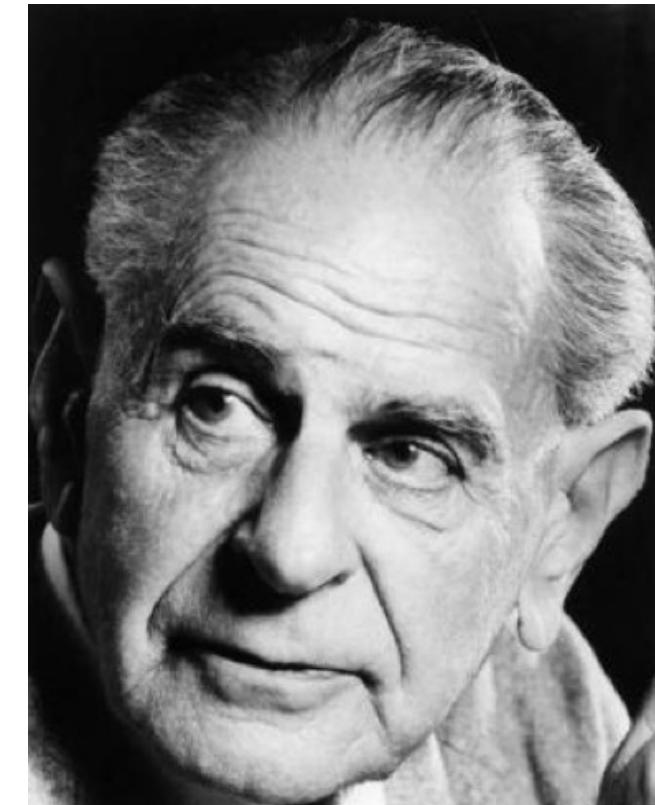
# 回顾科学研究的核心原则



清华大学  
Tsinghua University

## ● 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

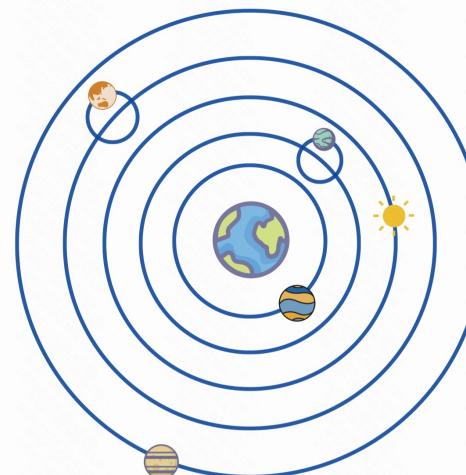
# 科学研究的核心是证伪



清华大学  
Tsinghua University

- 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

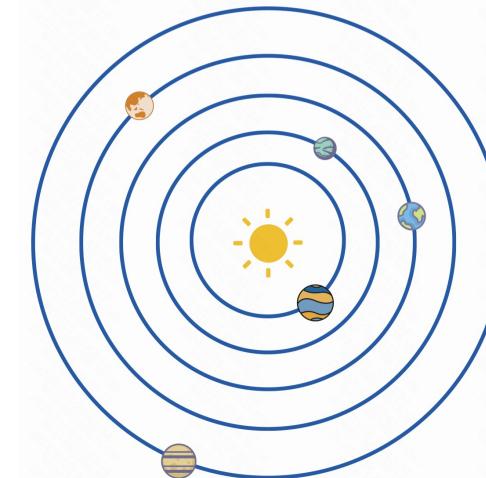


Geocentric Theory:

- 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.
- ...

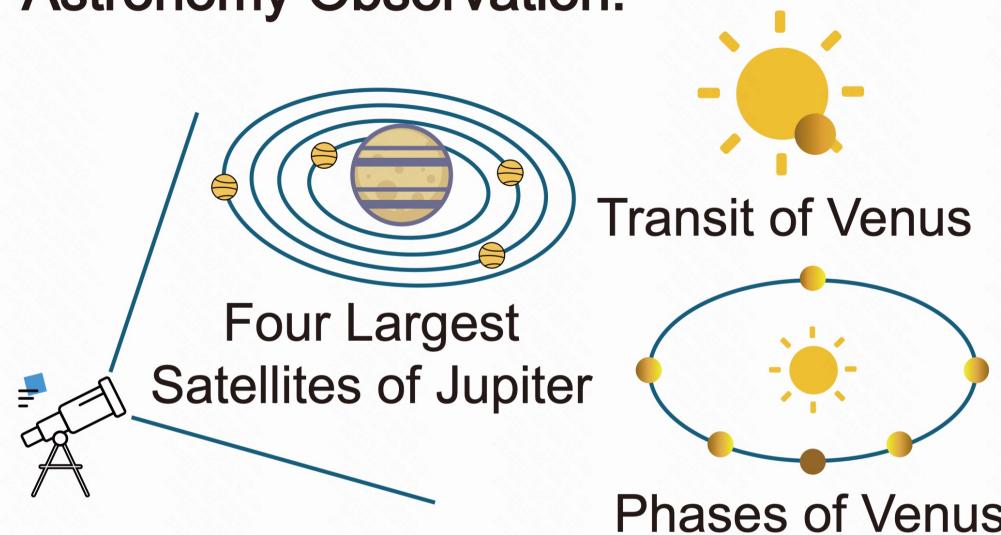
# 科学研究的核心是证伪



清华大学  
Tsinghua University

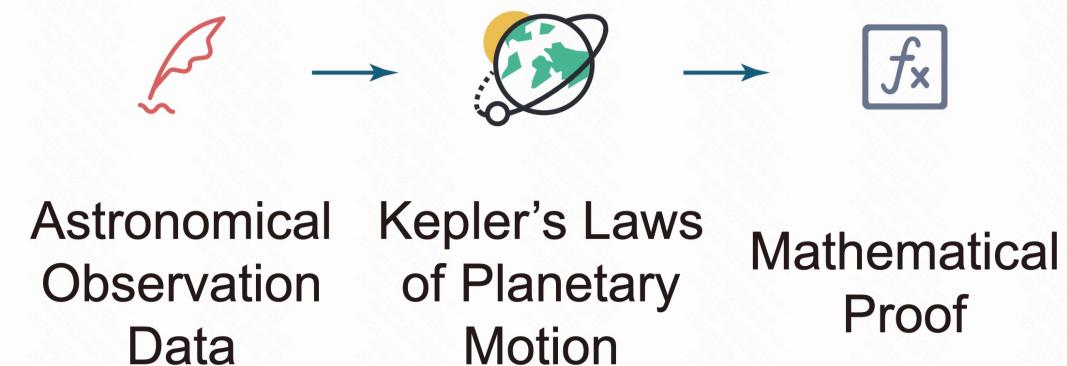
- 在科学研究的过程中，什么是证伪 (Falsification)?
- 需要设计实验证明假设正确或错误：

Astronomy Observation:



设计观察性实验证假说

Mathematical Reasoning:



数学形式化推理验证假说

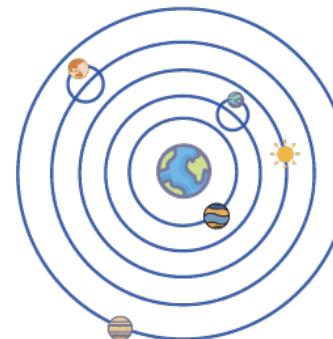
# 科学研究的核心是证伪



清华大学  
Tsinghua University

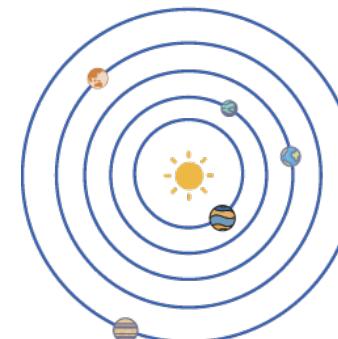
## ● 只有经过严格的证伪过程，才能产生可靠的科学发现。

### Hypothesis



Geocentric Theory:

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

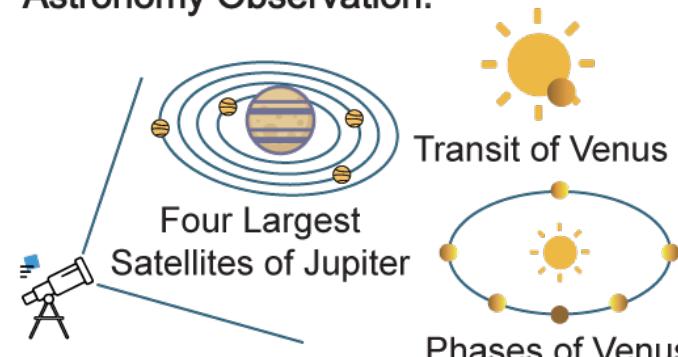


Heliocentric Theory:

- 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

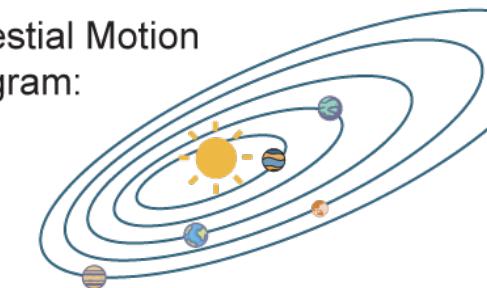
Final Discovery:

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

→ Geocentric Theory **✗**

→ Heliocentric Theory **✓**

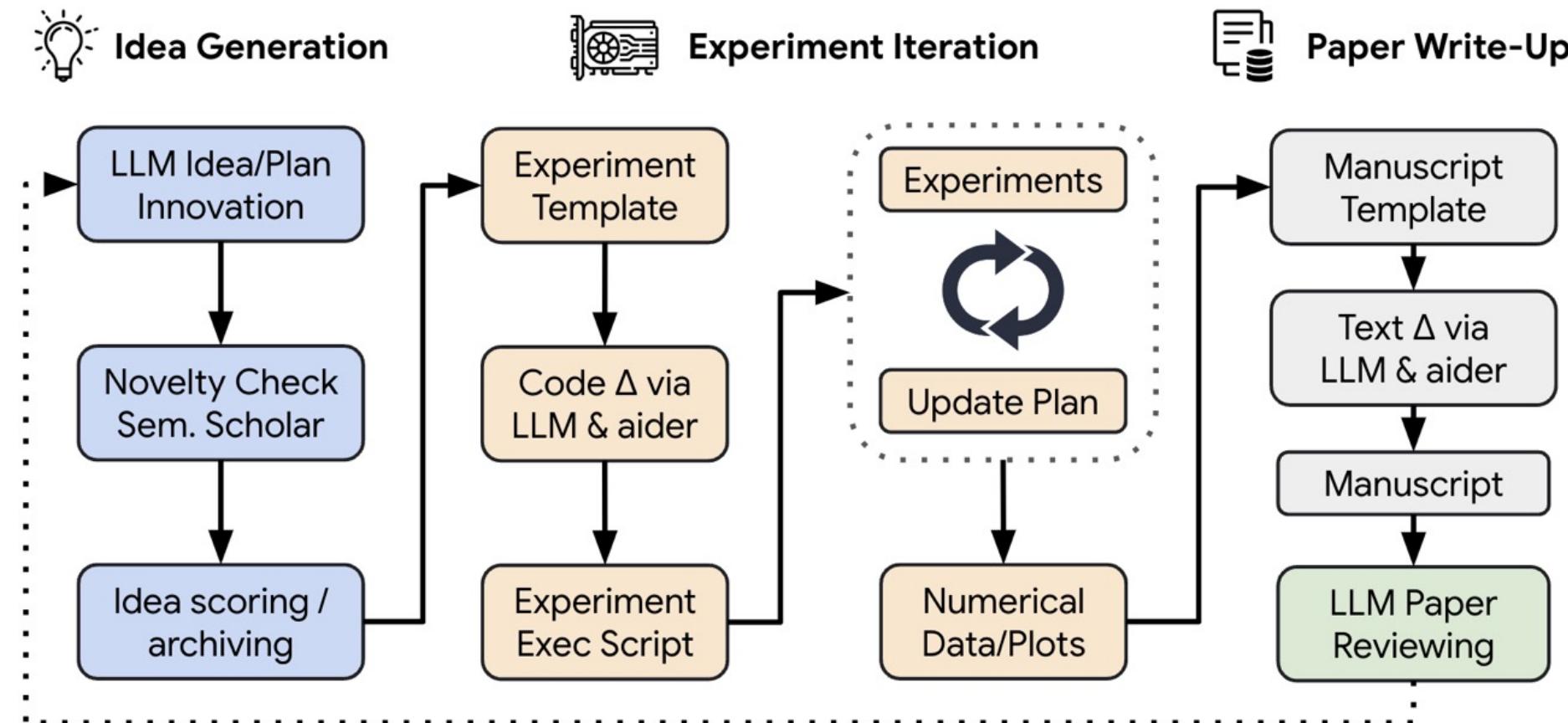
→ Celestial Motion Diagram:



# 目前的自主科研系统缺少自主证伪



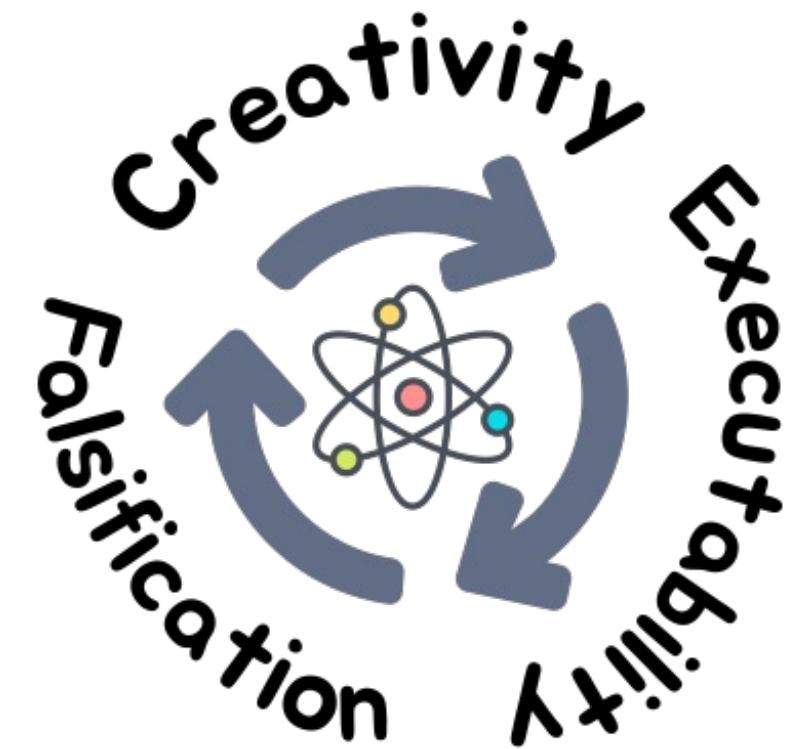
- 带着证伪的视角审视目前的科研智能体架构：目前的 AI 全流程科研系统缺少自主证伪 (Falsification) 过程。



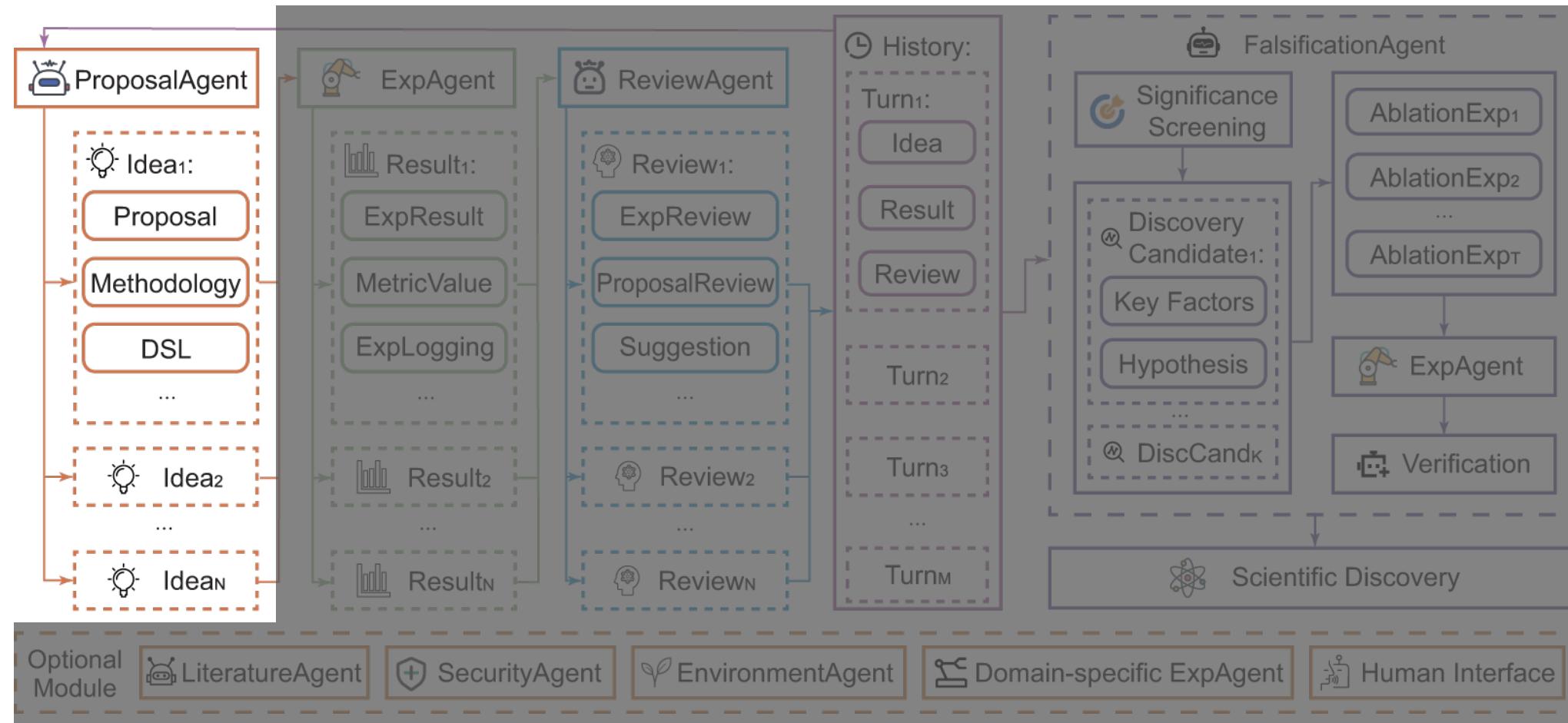
# AIGS及核心原则



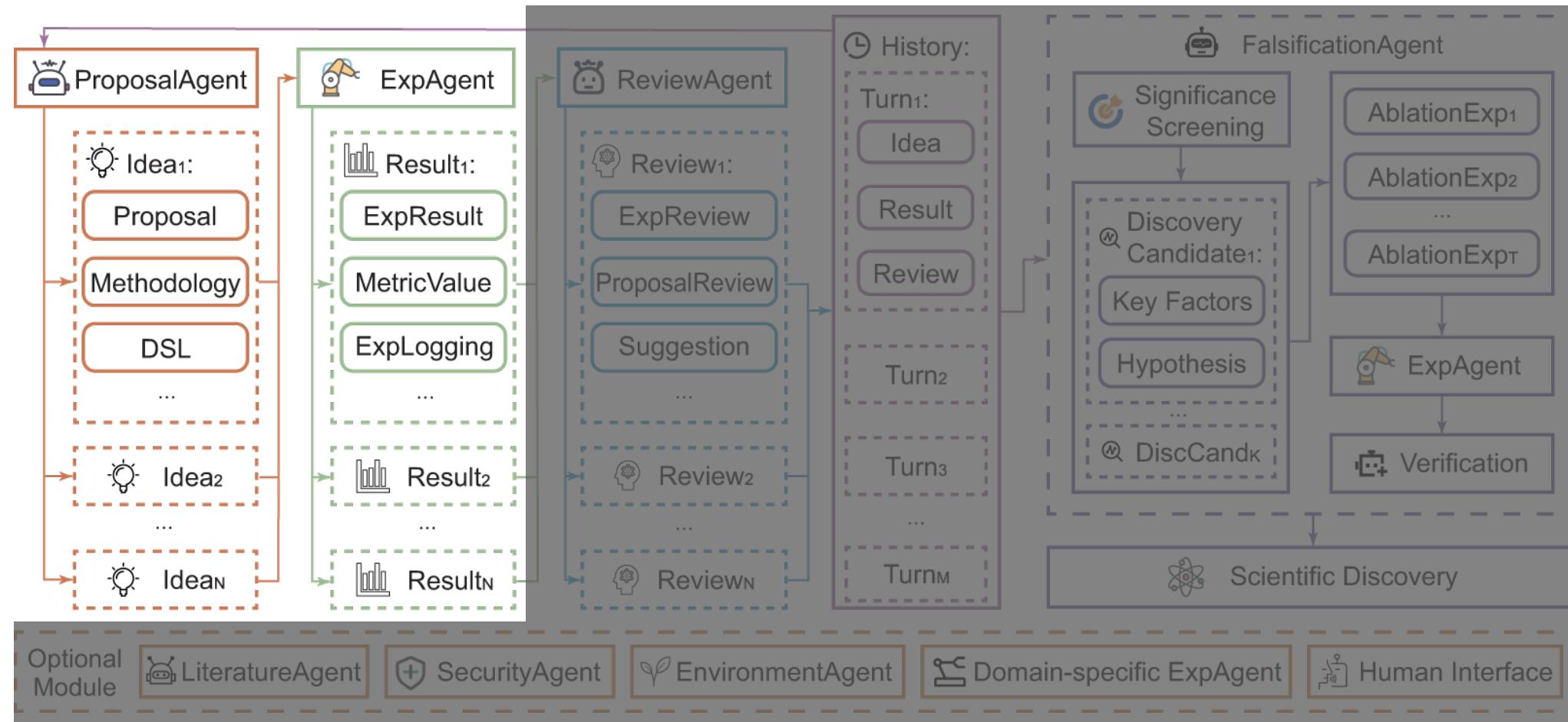
- AIGS (AI-Generated Science) : AI全自主作出的符合科学标准的科学发现。
- AIGS的实现需基于 3 个核心原则：
  - 证伪 (Falsification) 是科学研究的核心
    - 设计并执行实验以验证或反驳假设
  - 创造性 (Creativity) 的想法是科学研究的起点
    - 科学发现需要不断提出新的假设
  - 可执行性 (Executability) 构成了证伪的基础
    - 科学假设需要可执行的实验来验证



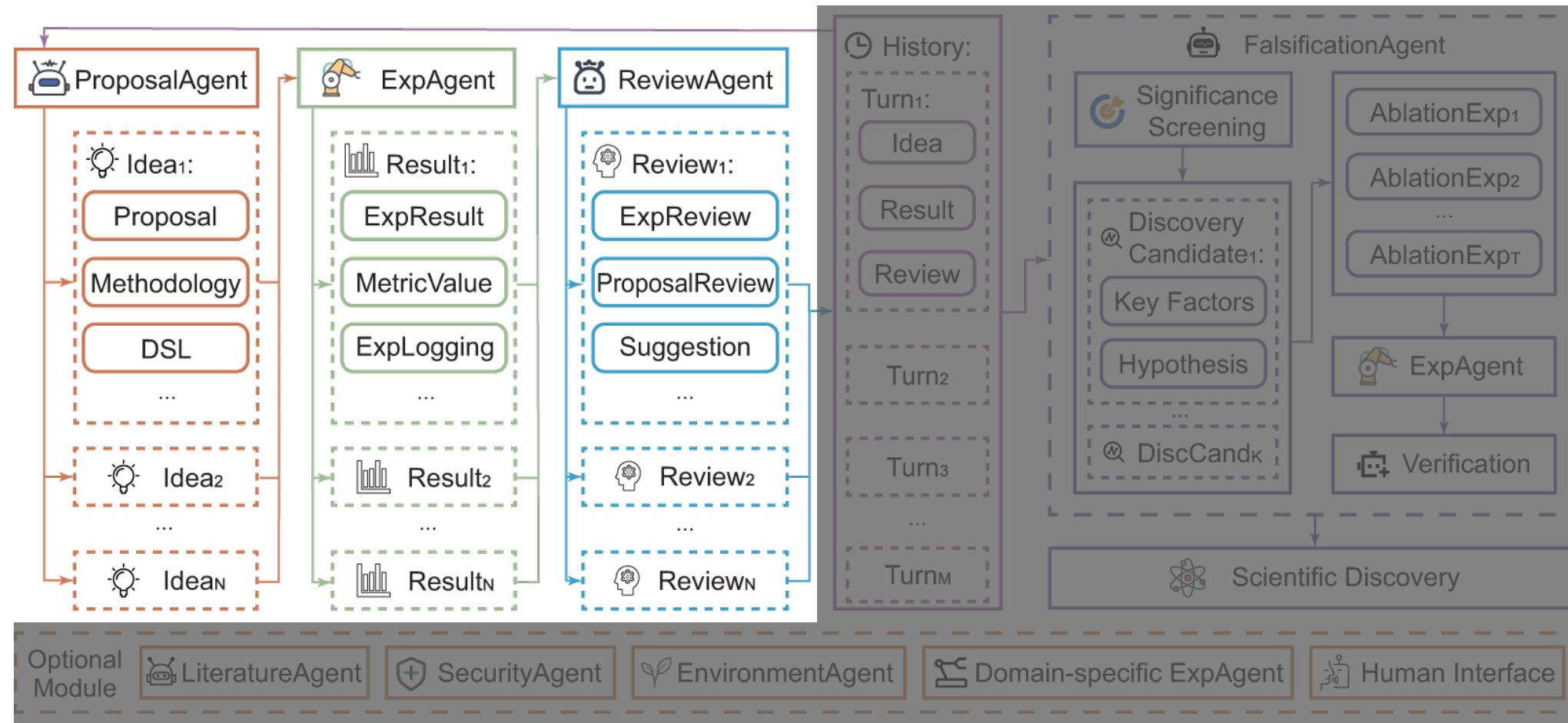
- **ProposalAgent:** 提出具有创造力的科研想法。



- **ExpAgent:** 实现想法并且进行科学实验。

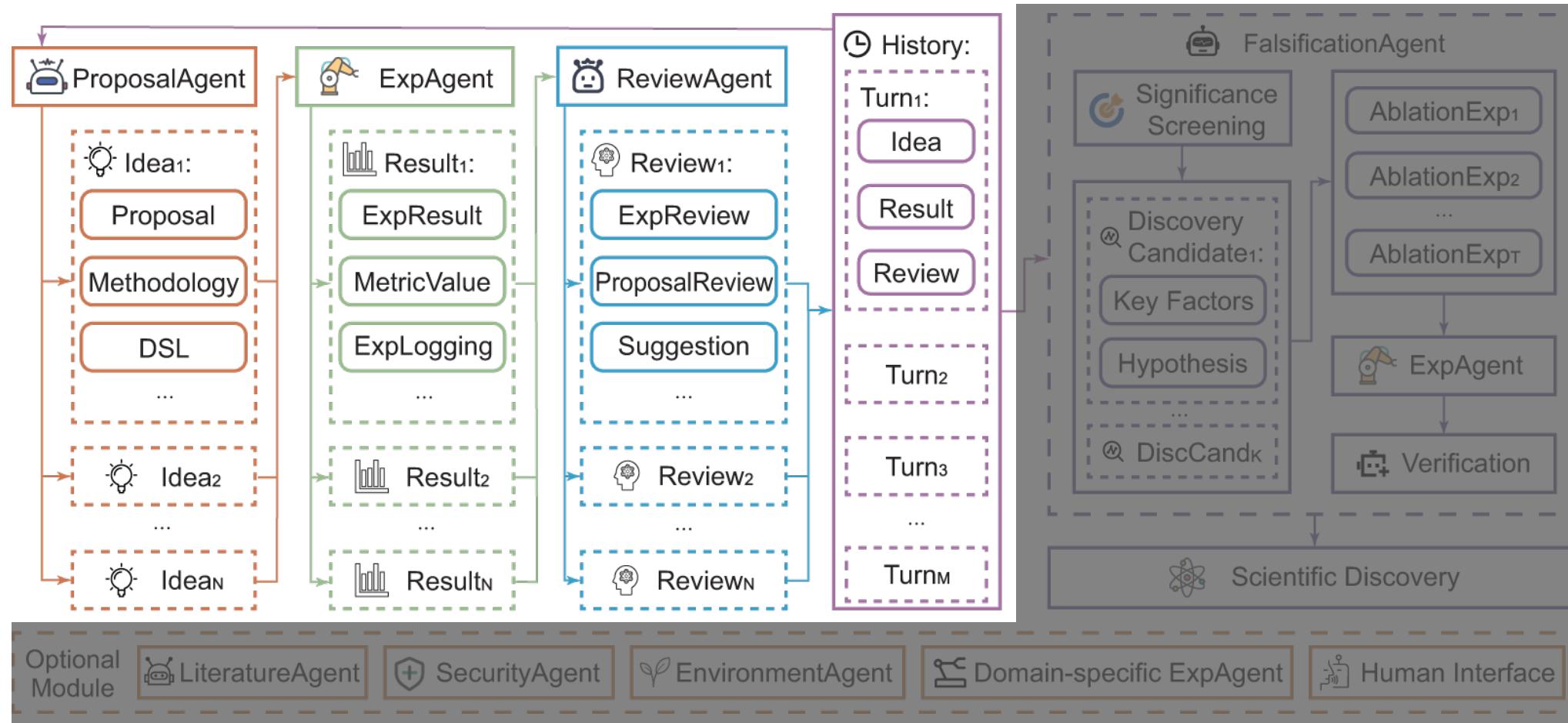


- ReviewAgent: 根据实验结果对想法提出建设性反馈。

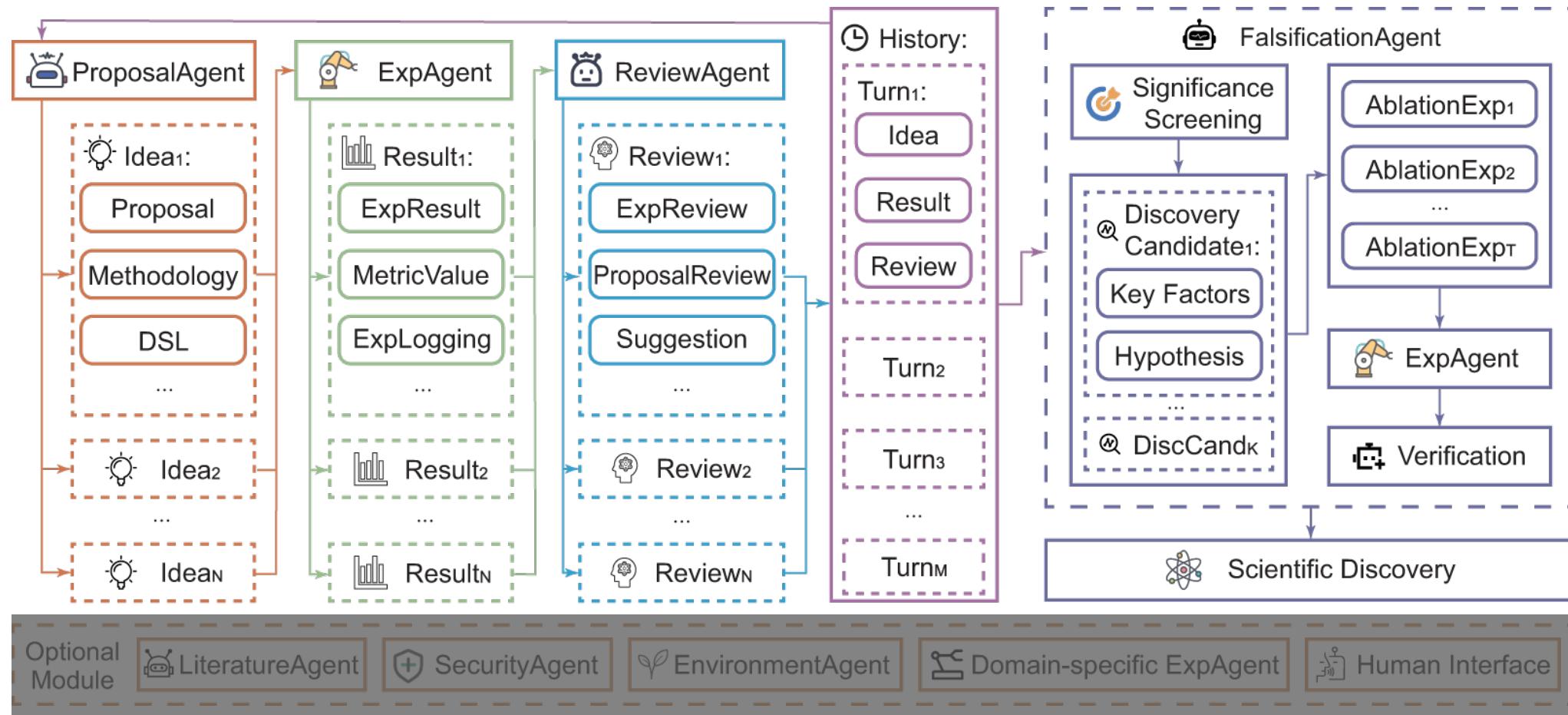




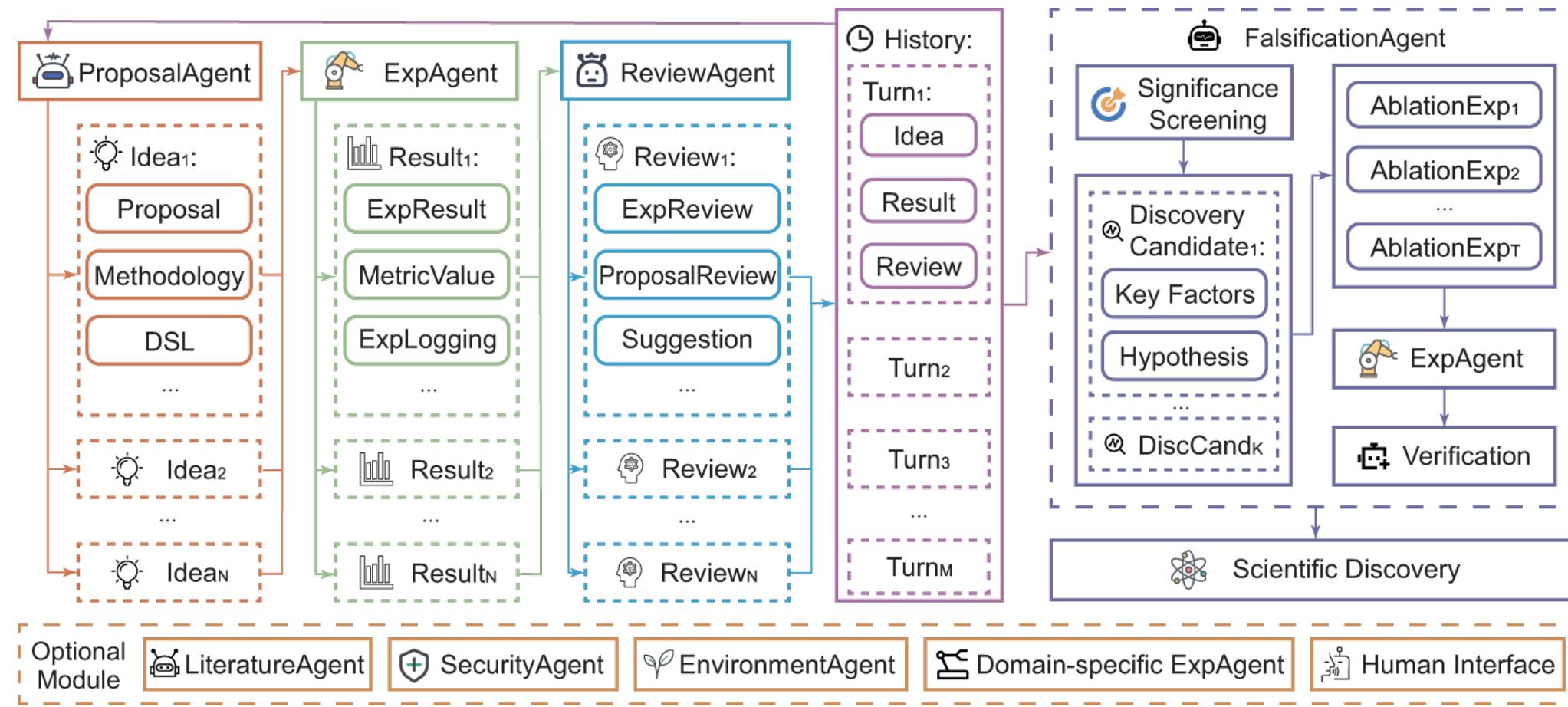
- Iteration: 循环迭代以提高实验表现。



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



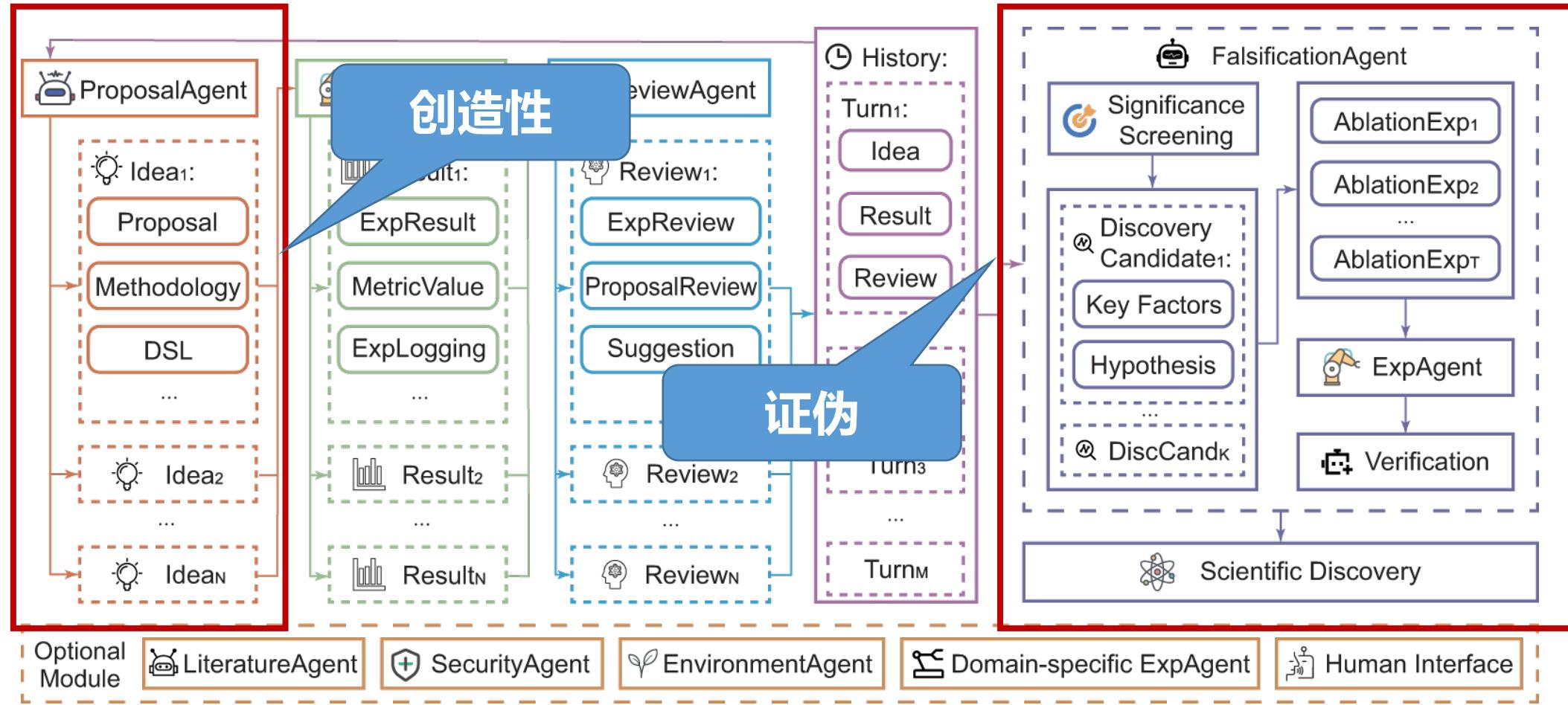
- Optional Module: 可集成其他模块实现补充功能。



# 三原则：证伪和创造性



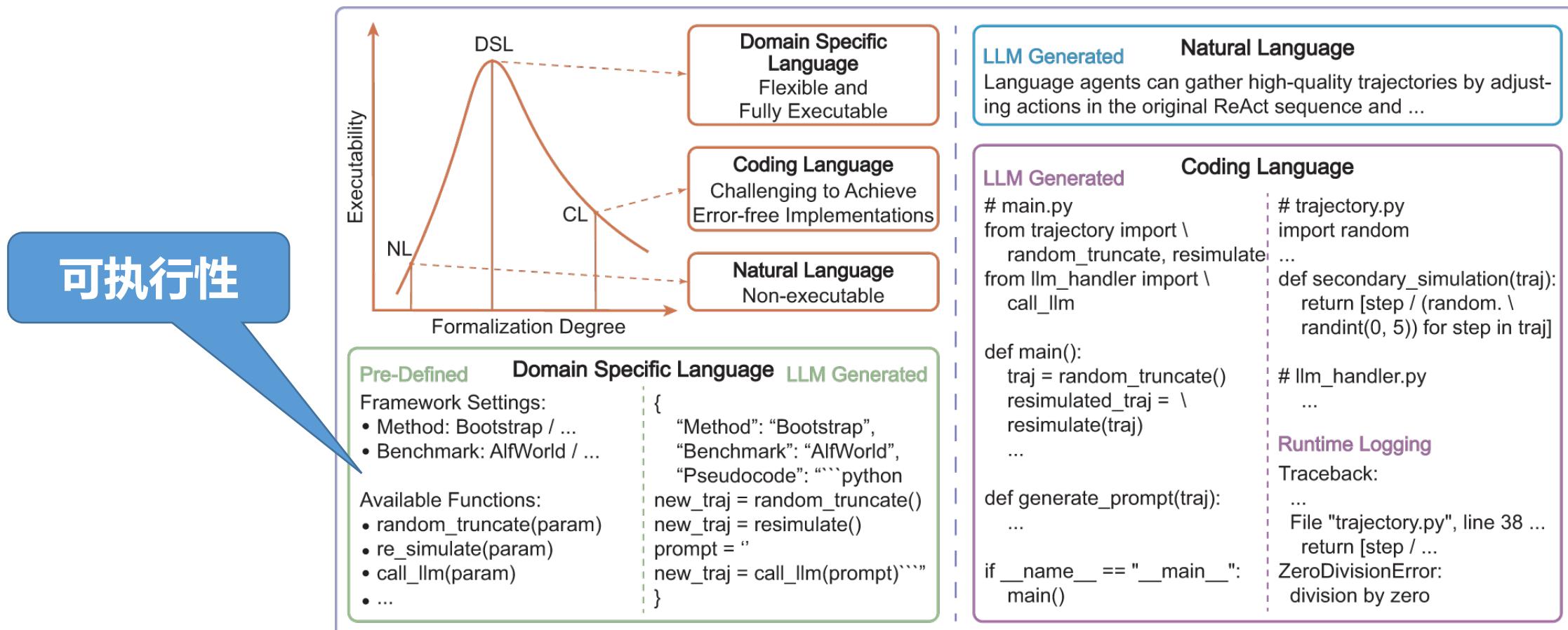
- 证伪和创造性主要由FalsificationAgent和ProposalAgent实现。



# 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。



# 实验设计



清华大学  
Tsinghua University

- 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。
- 实验指标上：我们基于 Baby-AIGS 的三大核心原则（证伪性、创造性和可执行性）进行评估。

# 实验结果：人工评价证伪过程



- Baby-AIGS能够通过证伪生成有效的科学发现。

Metric	Avg	Std	P-Value	Min	Max
<b>Importance Score (0 ~ 2)</b>					
BABY-AIGS (Ours)	1.80	0.41	0.02	0.00	<b>2.00</b>
Top Conference	<b>2.00</b>	0.00	—	<b>2.00</b>	<b>2.00</b>
<b>Consistency Score (0 ~ 2)</b>					
BABY-AIGS (Ours)	1.00	0.86	0.00	0.00	<b>2.00</b>
Top Conference	<b>2.00</b>	0.00	—	<b>2.00</b>	<b>2.00</b>
<b>Correctness Score (0 ~ 2)</b>					
BABY-AIGS (Ours)	0.95	0.94	0.00	0.00	<b>2.00</b>
Top Conference	<b>2.00</b>	0.00	—	<b>2.00</b>	<b>2.00</b>
<b>Overall Score (0 ~ 2)</b>					
BABY-AIGS (Ours)	1.25	0.47	0.00	0.67	<b>2.00</b>
Top Conference	<b>2.00</b>	0.00	—	<b>2.00</b>	<b>2.00</b>

- 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。

# 实验结果：通过基准测试评估创造力



清华大学  
Tsinghua University

- 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
<b>BABY-AIGS (Ours)</b>	<b>4.51</b>	4.77
Top Conference	4.45	<b>5.01</b>

## 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 baseline 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。

# 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
<b>BABY-AIGS (Ours)</b>	<b>3.26</b>

## 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)	<b>1.473</b>	1.003	0.974
<b>BABY-AIGS (Ours)</b>	1.499	<b>0.984</b>	<b>0.966</b>

## 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.

# 实验结果：通过成功率比较评估可执行性



清华大学  
Tsinghua University

- Baby-AIGS 在可执行性上显著优于现有系统。

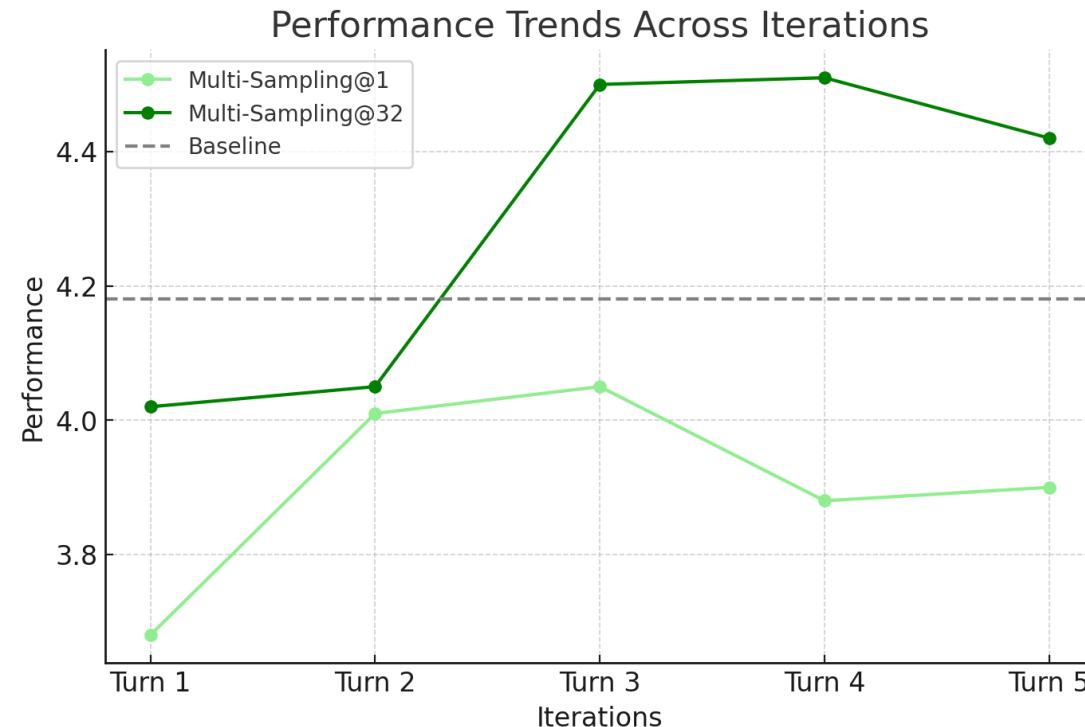
Method	Experiment Success Rate (Exp. SR)	Overall Success Rate (Overall SR)
AI Scientist	44.8%	29.2%
<b>Baby-AIGS (Ours)</b>	<b>Almost 100%</b>	<b>Almost 100%</b>

- 定量分析表明，Baby-AIGS 的可执行性优于 baseline。
- 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。

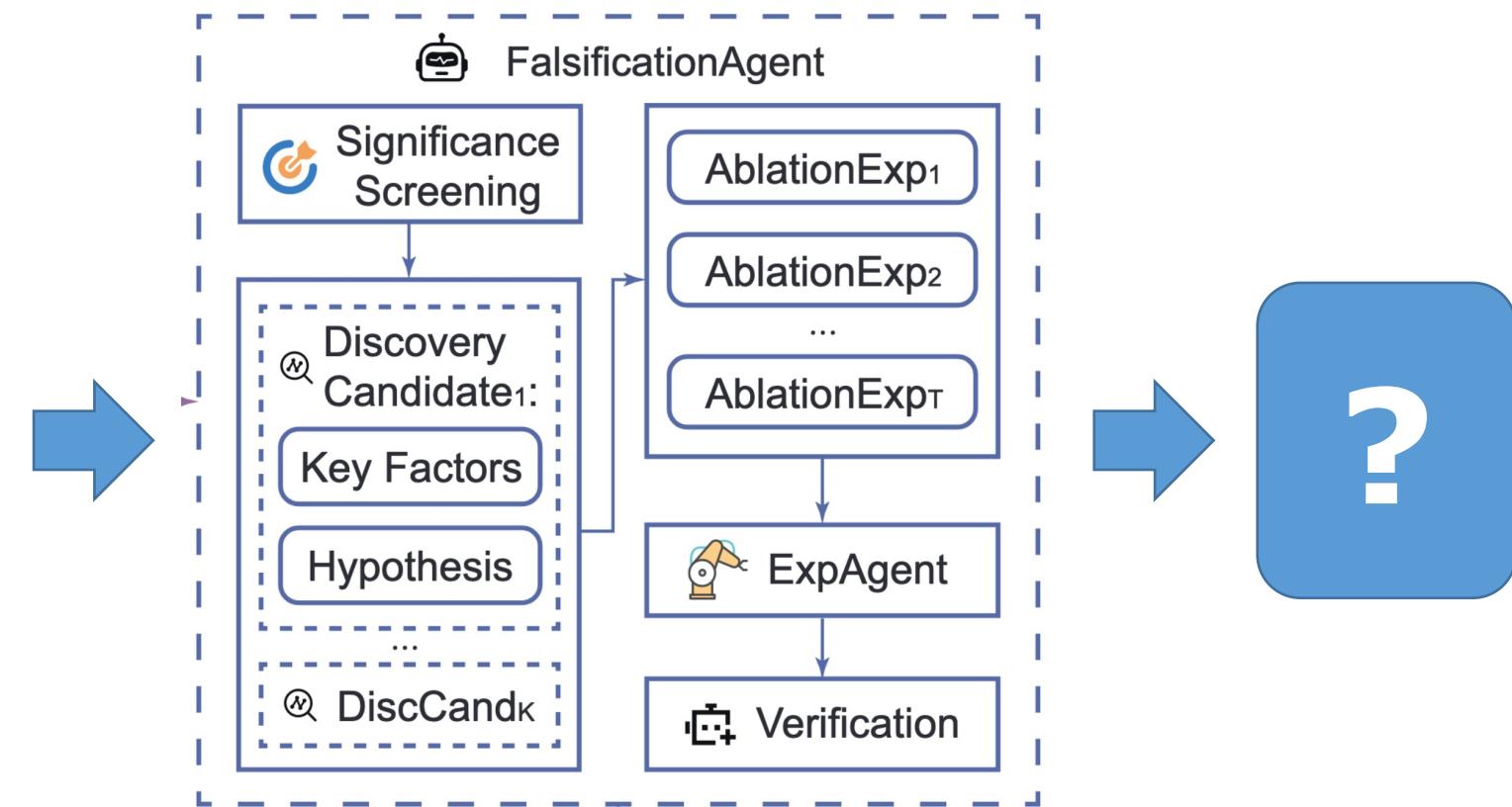
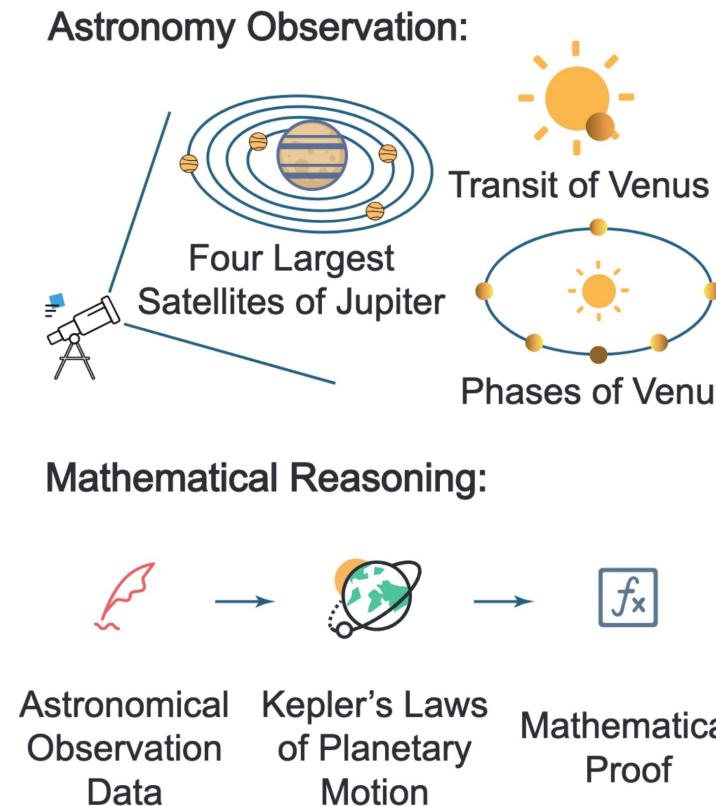


- 一种基于搜索的 scaling inference compute 方法。
- 更好的创造力表现一定程度上归功于多采样搜索。

# 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。

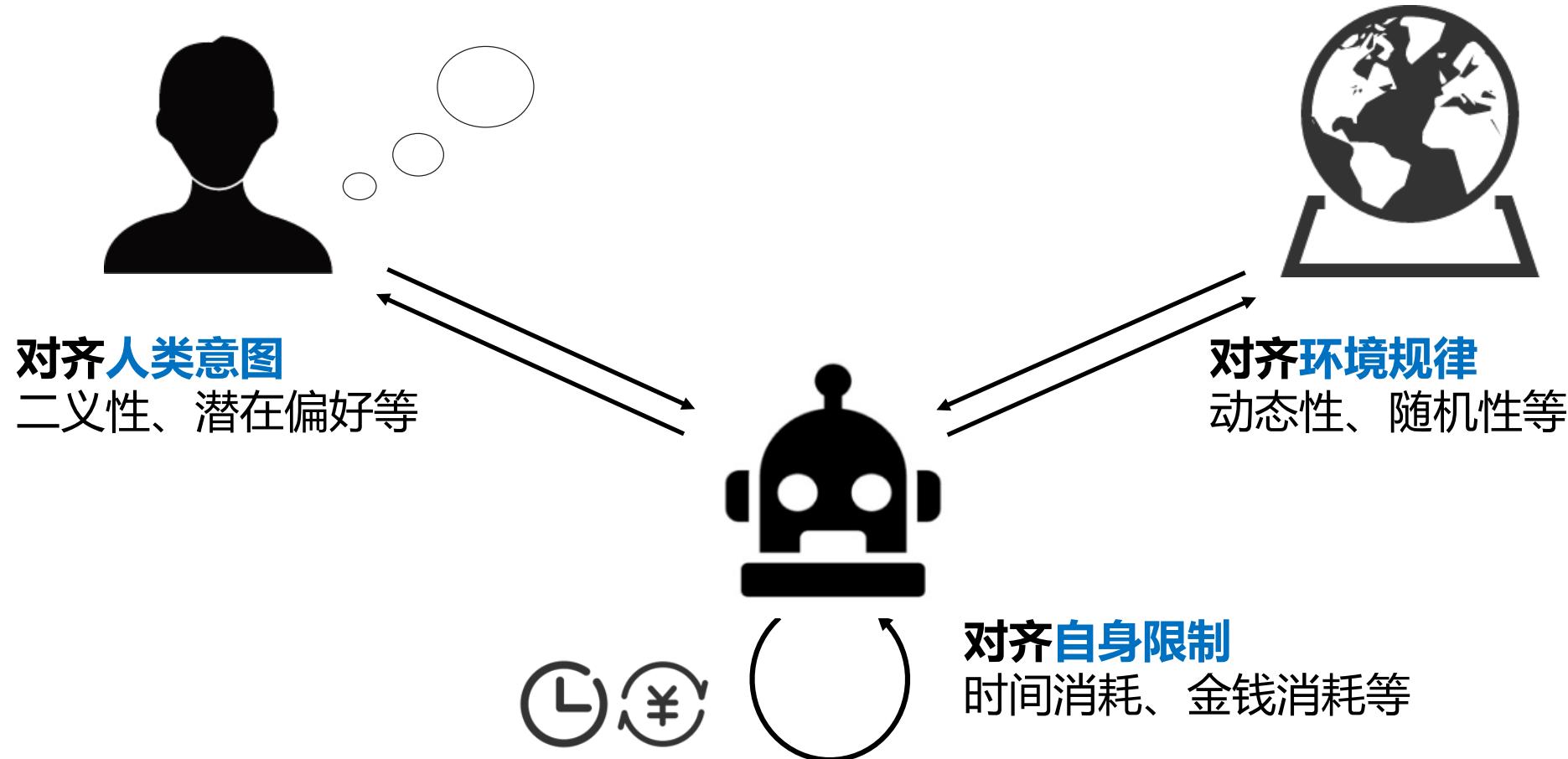


# 展望：AIGS 需要遵循 UA<sup>2</sup> 对齐原则



清华大学  
Tsinghua University

- AIGS 智能体需要和自身、人类、环境统一对齐。

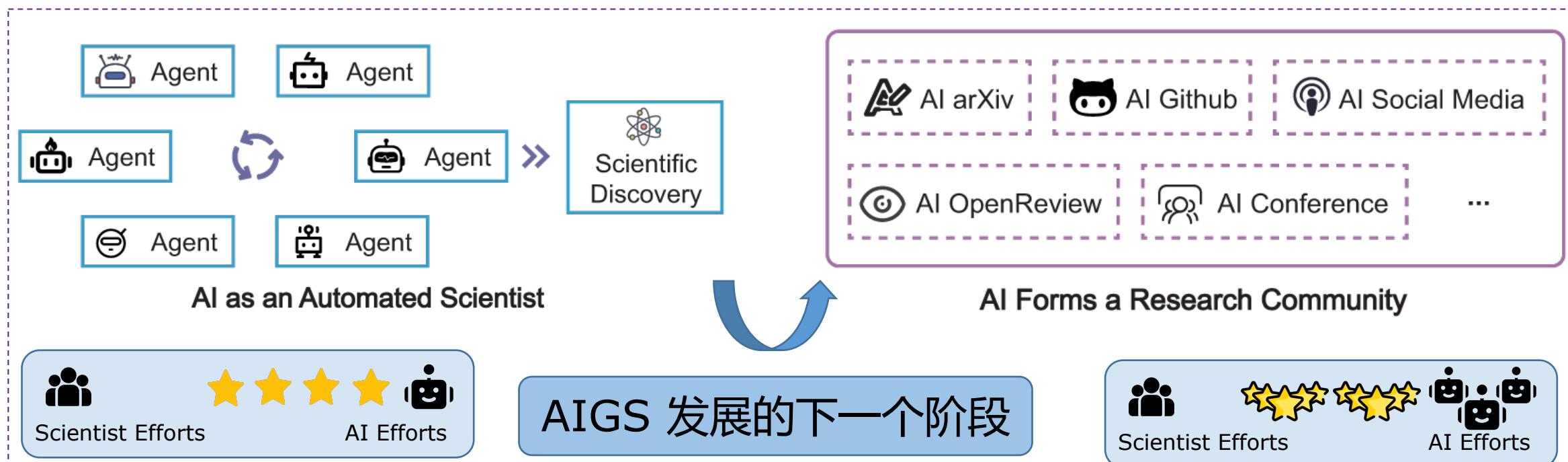


# 展望：建立 AI 科学家社群



清华大学  
Tsinghua University

## ● 建立AI科学家社群，催化产生跨学科重大科学发展。

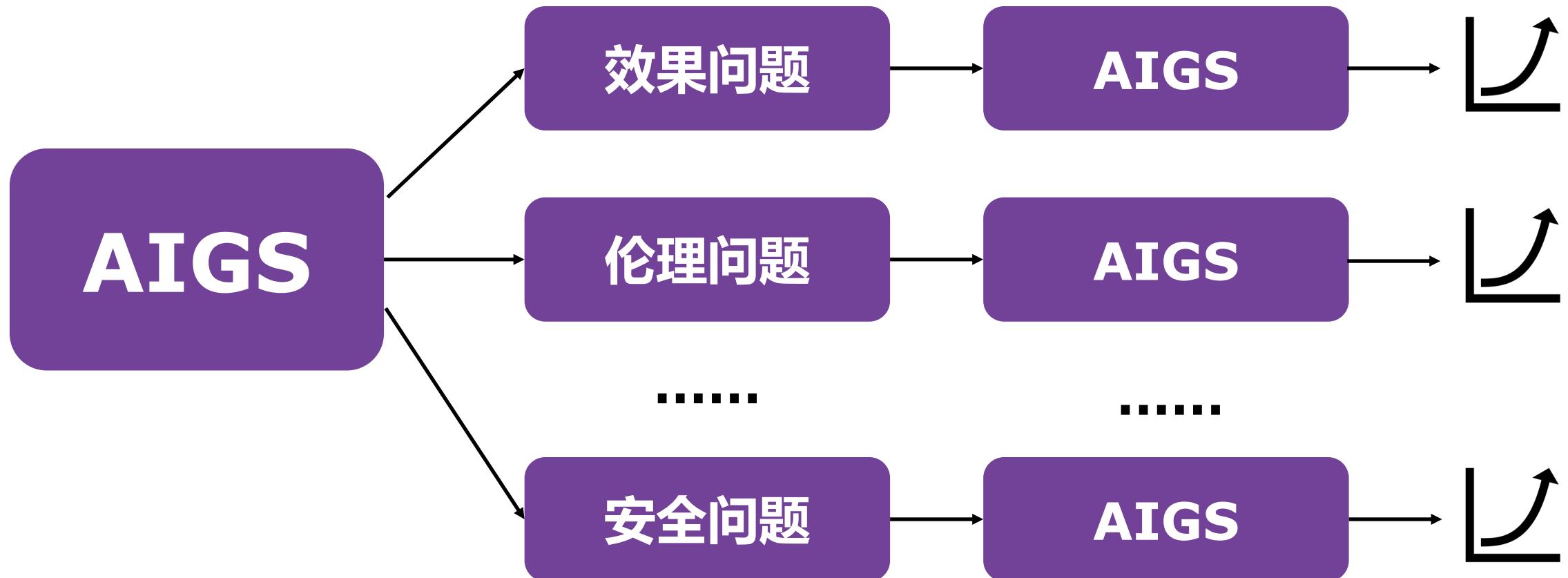


AI 科学家们也需要建立 AI 科学社群  
建立高效的交流、合作和检验机制来激发群体智能

# 展望：AIGS for AIGS



- 利用AIGS来提升AIGS、解决AIGS自身面临的挑战。



# 主要成员



李鹏



刘子君



刘铠铭



朱奕祺



刘洋



雷轩宇



杨宗瀚



张真赫

# 总结



清华大学  
Tsinghua University

- 提出AIGS (AI-Generated Science) , 即AI全自主作出的符合科学标准的科学发现。
- AIGS的实现需要基于 3 个核心原则：
  - 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。
- 提出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

