AI自主科学发现

李鹏

清华大学

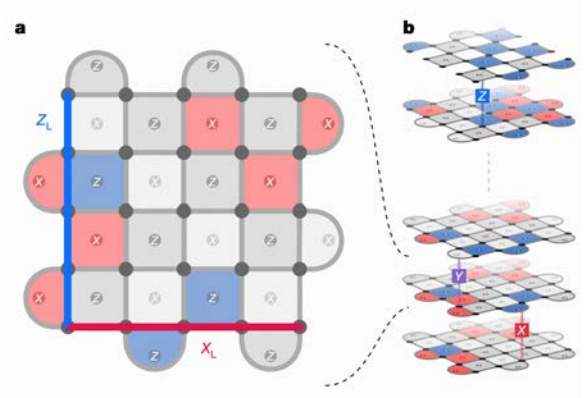
AI深刻改变科学研究范式

清華大學 Tsinghua University

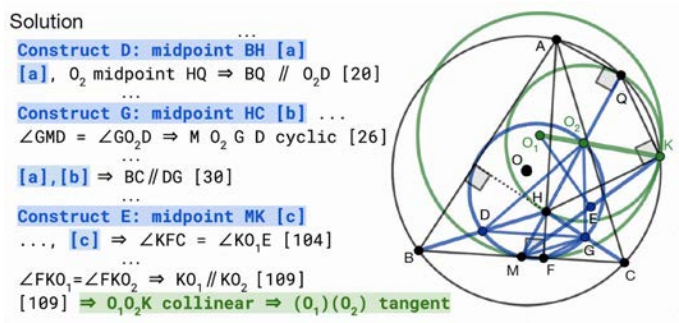
- **深度学习的发展赋能AI4Science，极大地推动了科学的发展。**

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

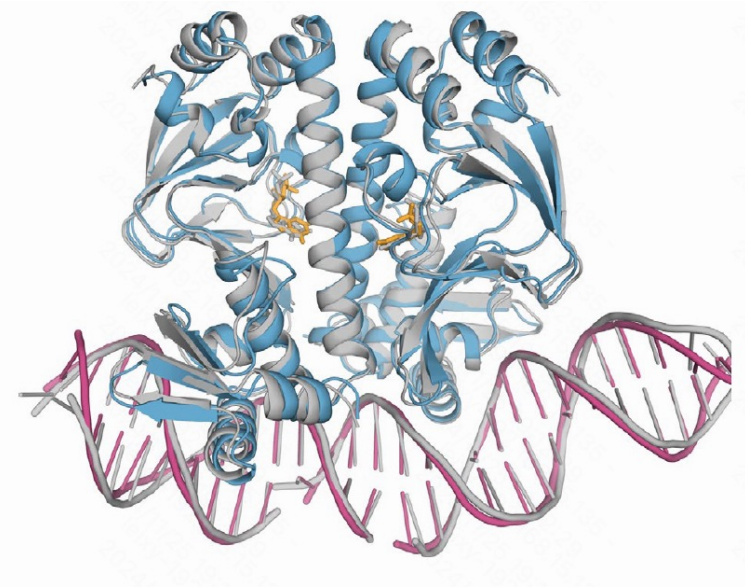
David Baker, Demis Hassabis and John Jumper. Ill. Niklas Elmehed © Nobel Prize Outreach



AlphaQuant



AlphaGeometry

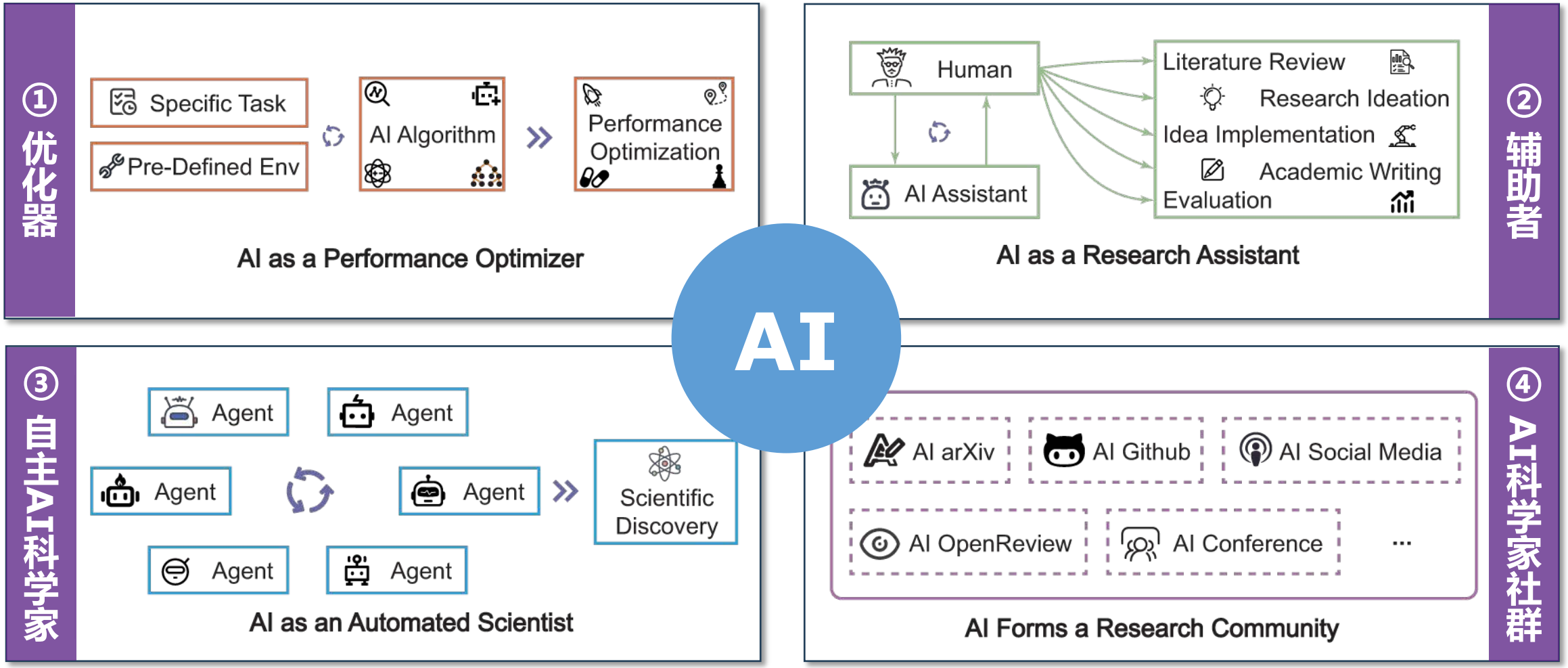


AlphaFold

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

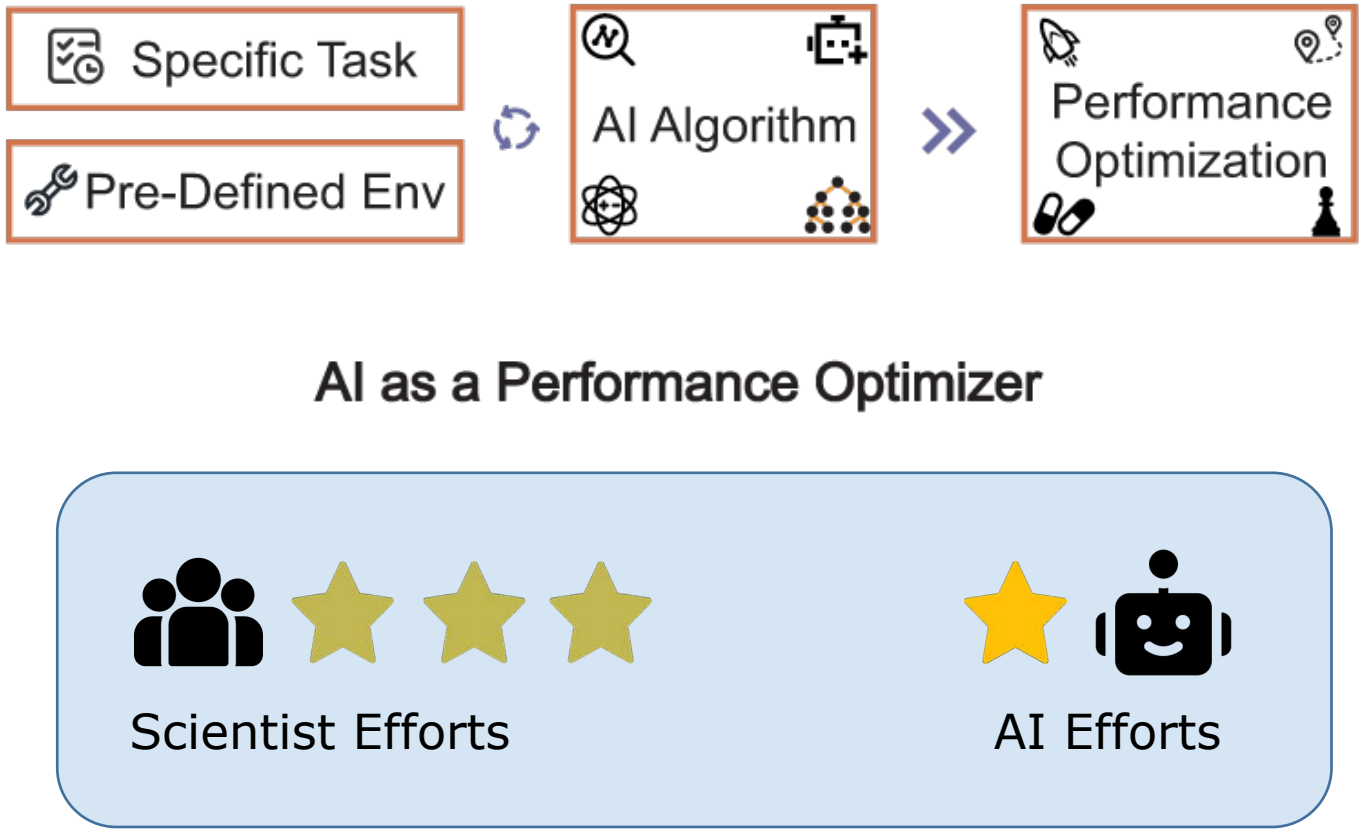
- 根据AI在科研中的角色可以划分出 4 种AI参与科学研究的范式。

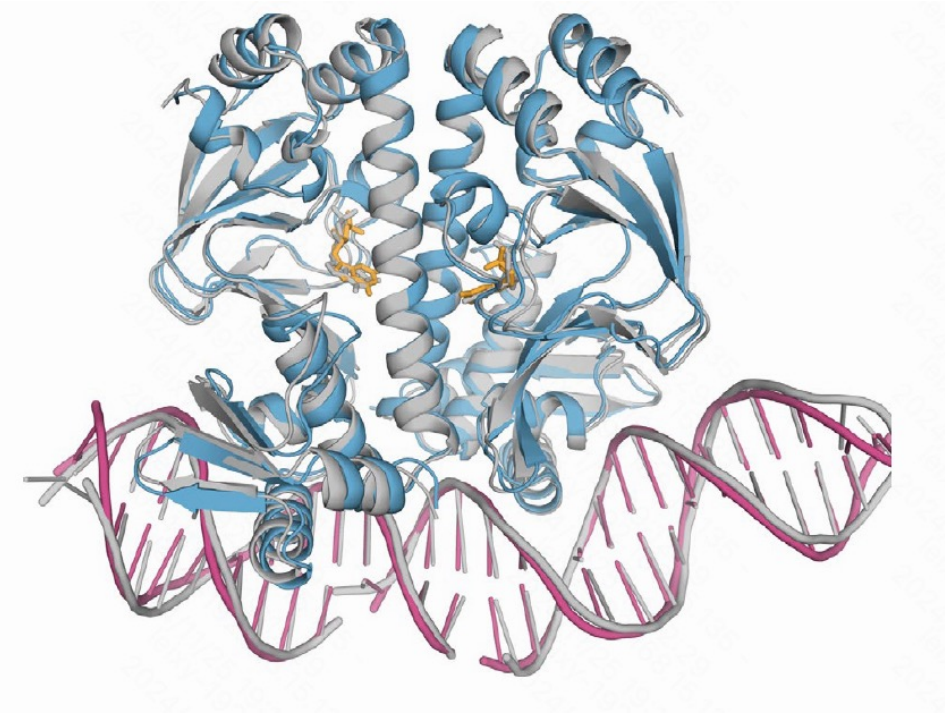


Liu et al., 2024. AIGS: Generating Science from AI-Powered Automated Falsification. arXiv:2411.11910.

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。



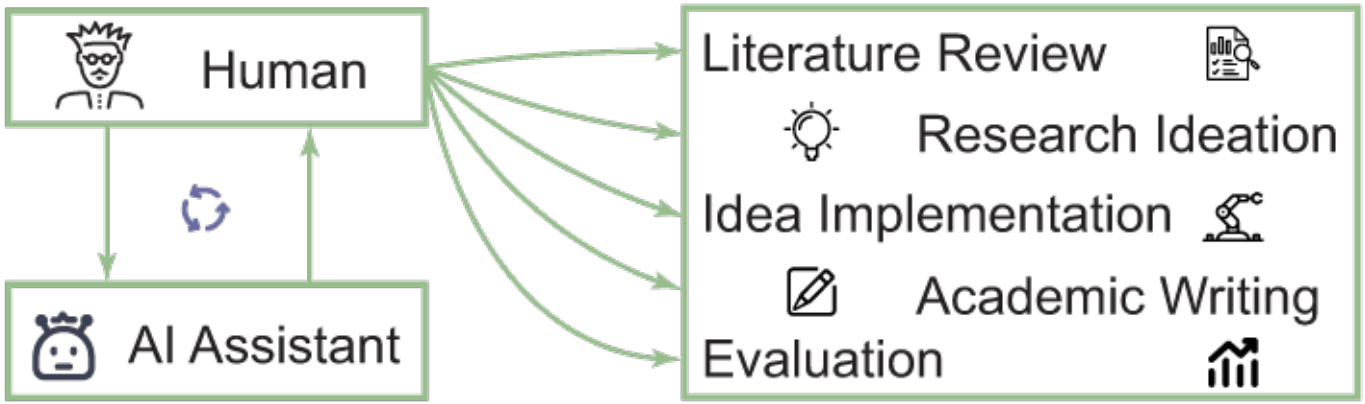


AlphaFold

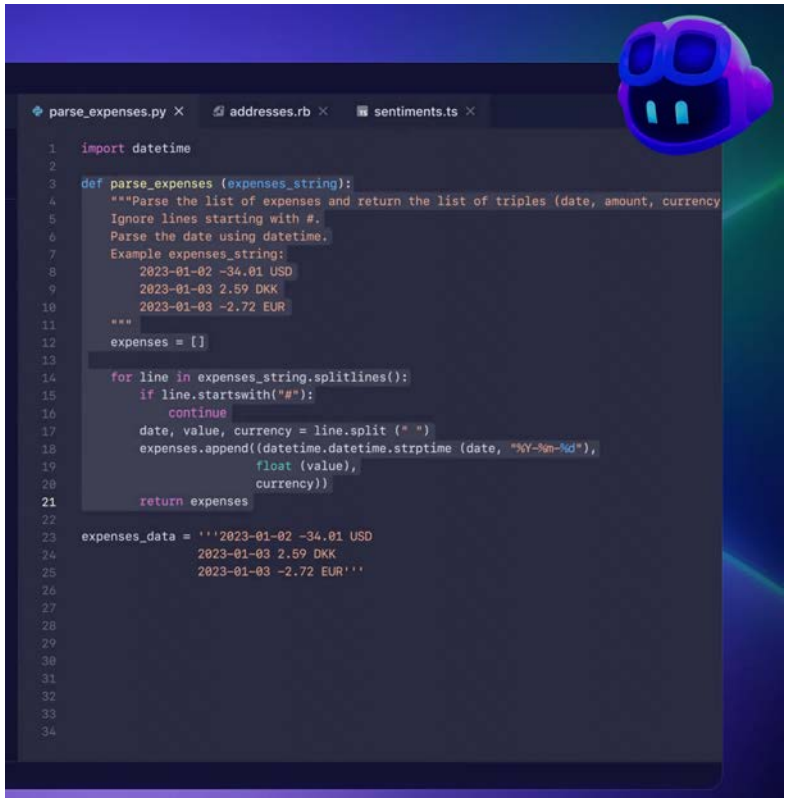
https://www.nature.com/articles/s41586-024-08148-8

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

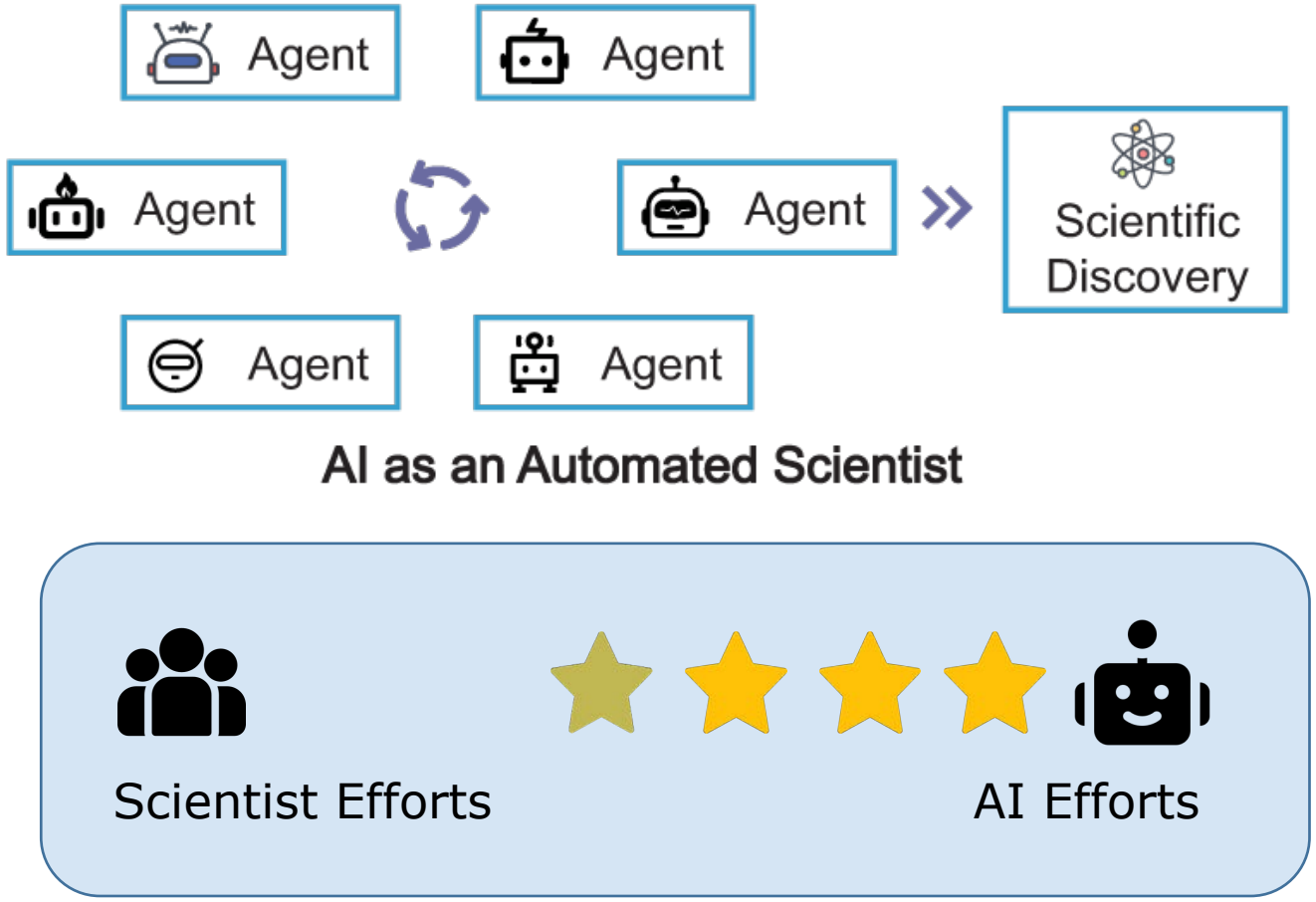


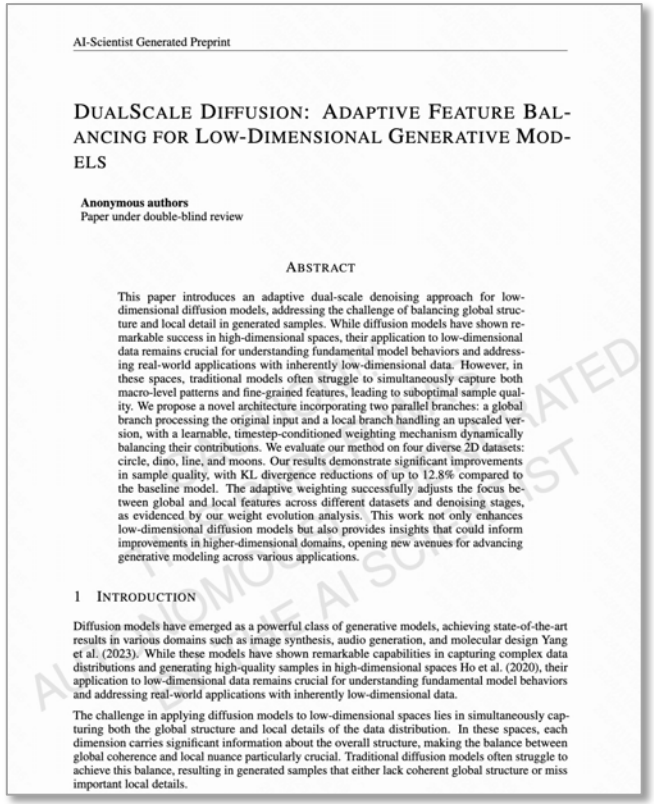
Github Copilot

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

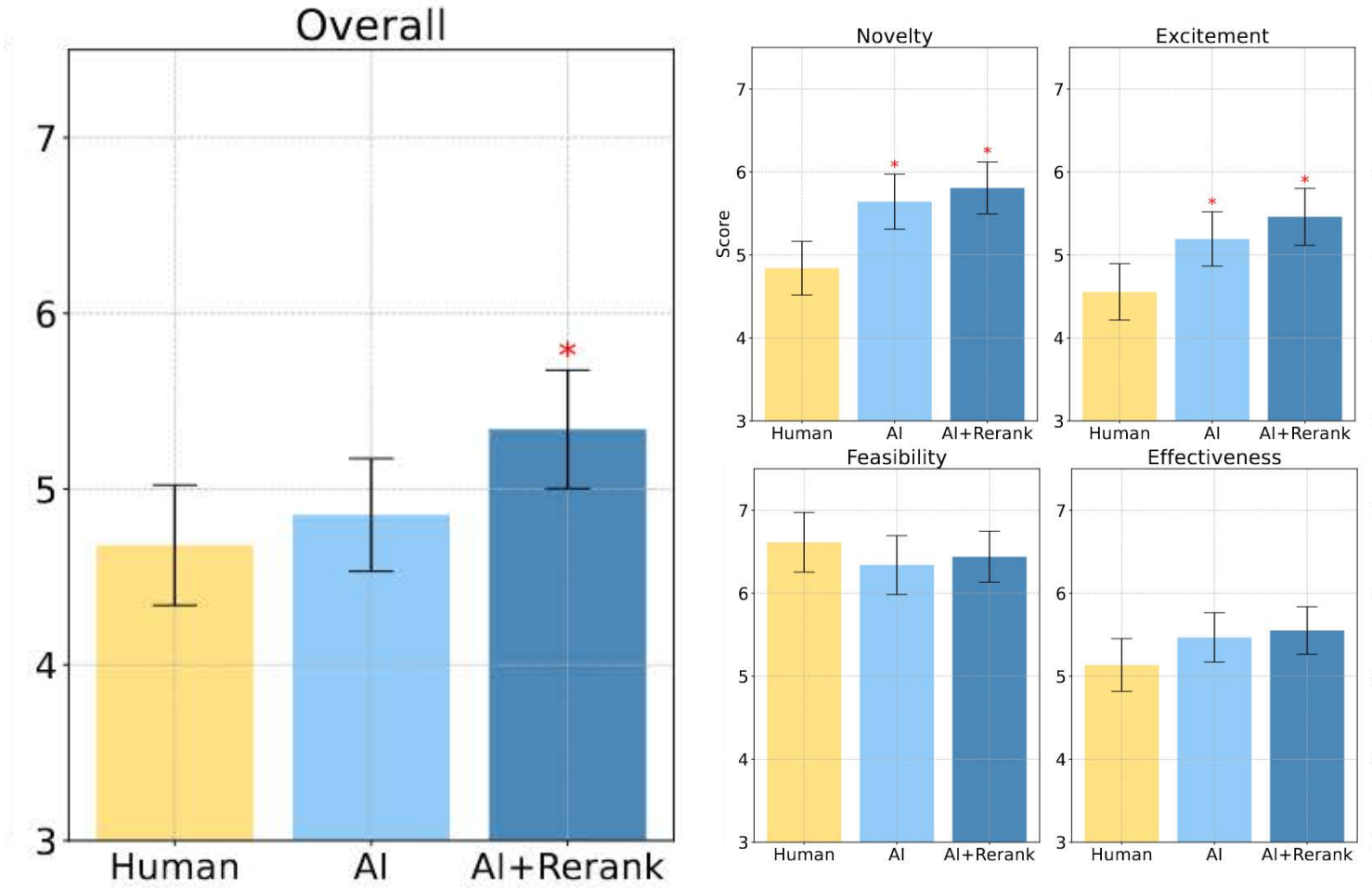
Lu et al., 2024. The AI Scientist: Towards Fully Automated Open-ended Scientific Discovery. arXiv:2408.06292.

范式三的初步探索

- **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

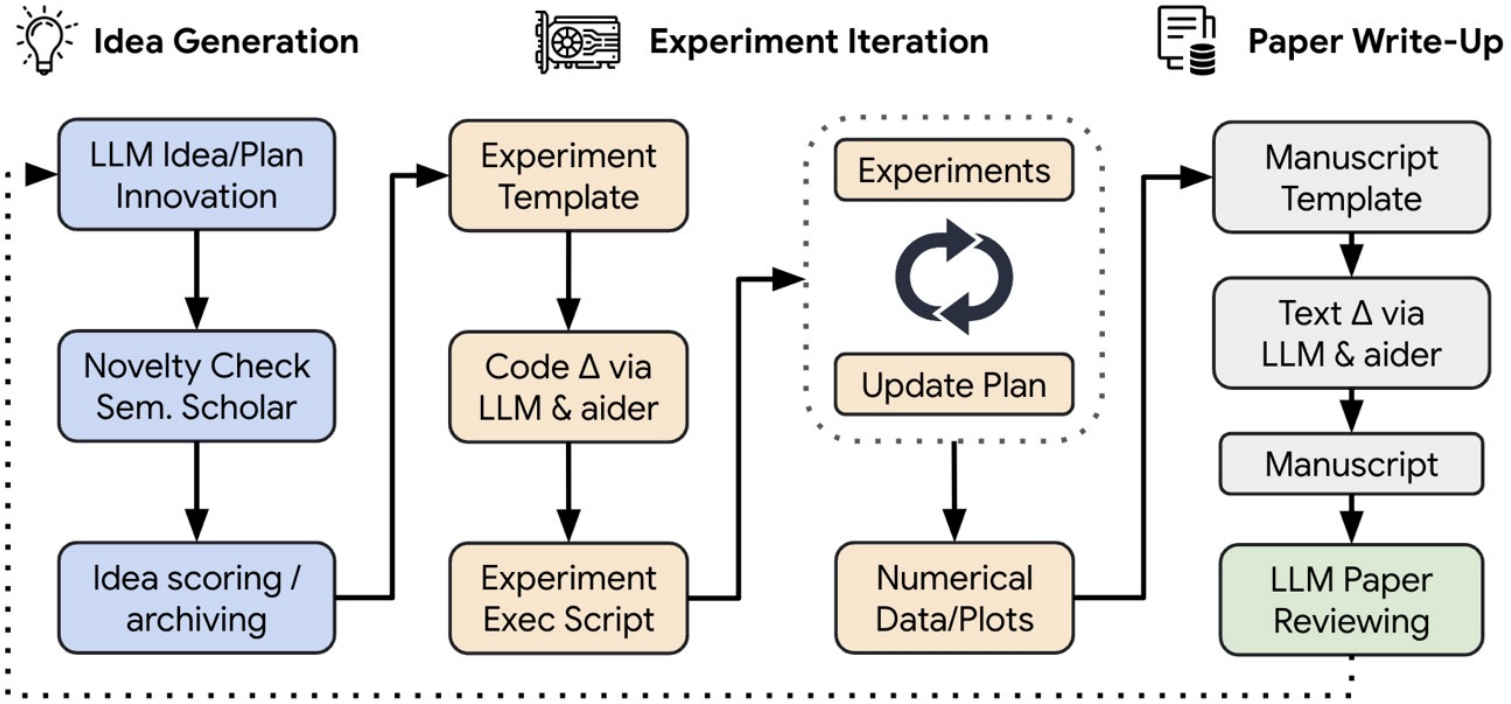
- ❑ 考查了7个NLP主题
- ❑ 约束在提示学习子领域内
- ❑ 系统无需对idea进行验证
- ❑ 人类被试提供的未必是其个人想到的最优idea

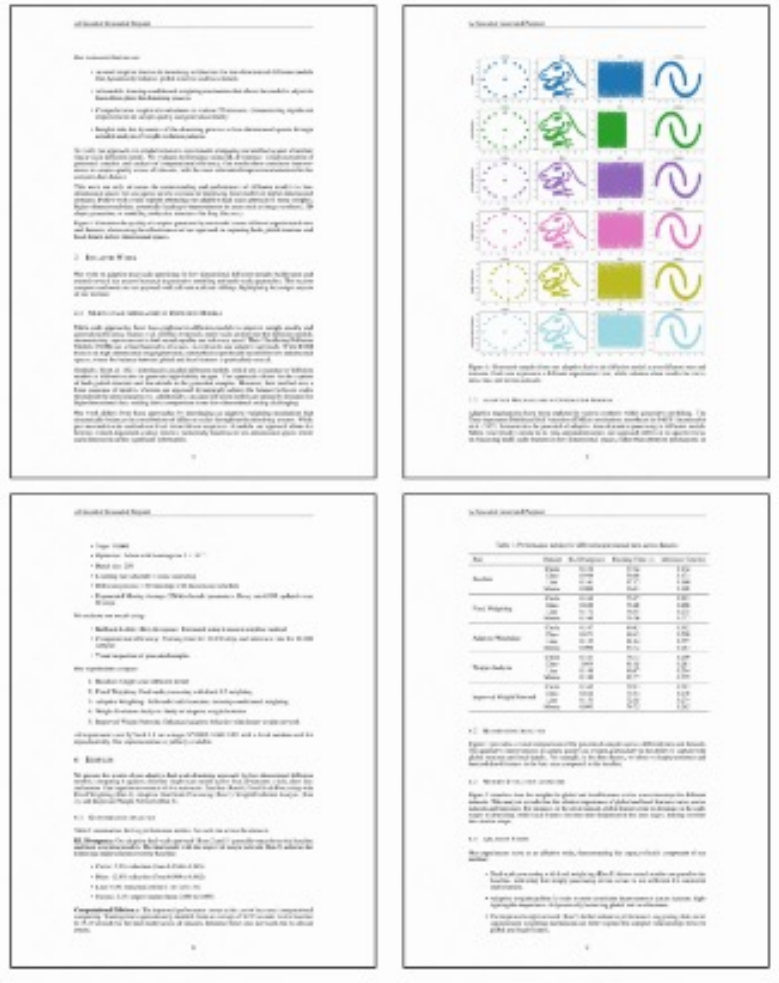


Si et al., 2024. Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers. arXiv:2409.04109.

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

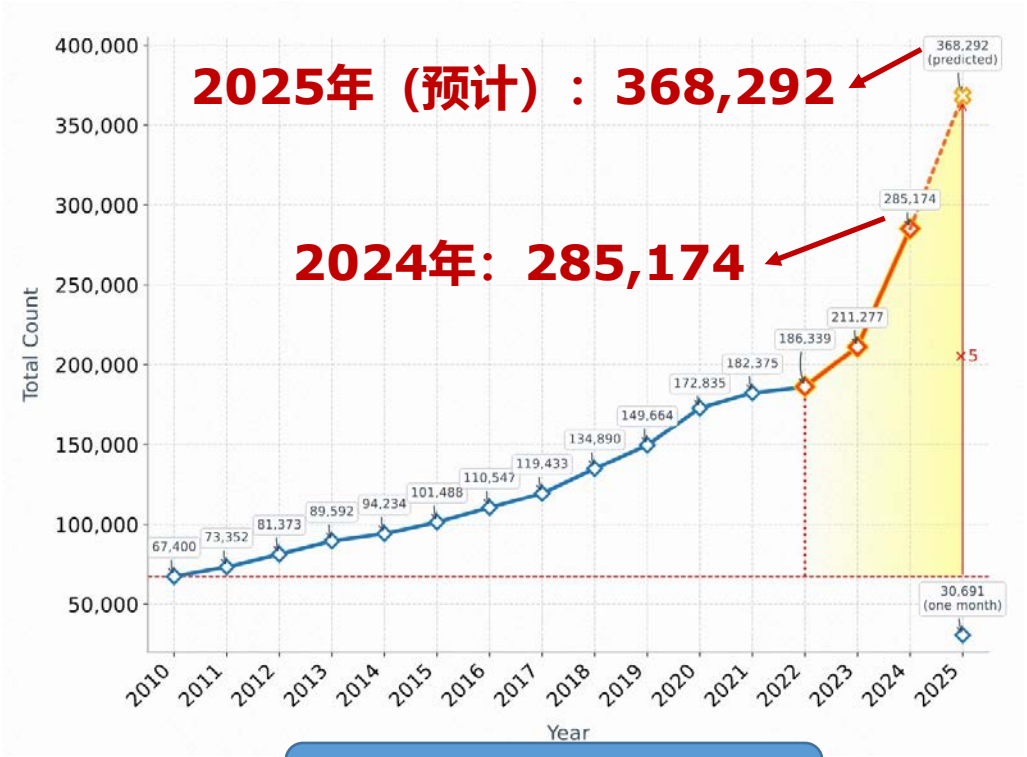
Lu et al., 2024. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. arXiv:2408.06292.

文献调研

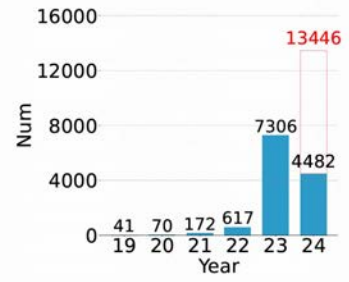
文献调研是科学发现的基础

清華大學 Tsinghua University

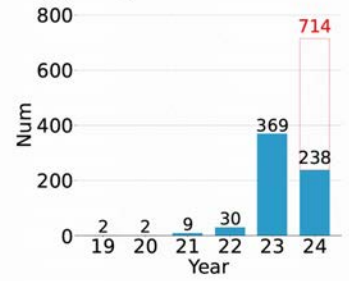
- **随着论文数量的指数型膨胀，文献调研愈发费时费力，利用智能体自动化文献调研具有很高的应用价值。**



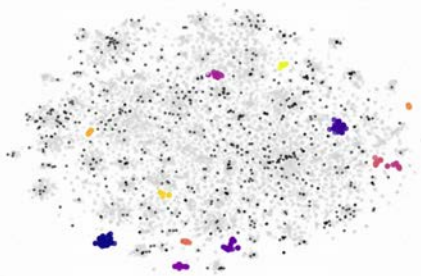
ArXiv Submissions

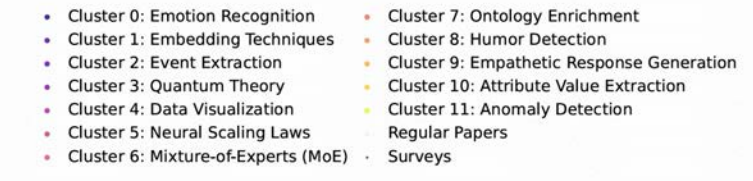


(a) #Paper about LLM.



(b) #Survey about LLM.





(c) T-SNE visualization of surveys and papers about LLMs. Clusters represent groups of papers identified through clustering, which currently lack comprehensive survey coverage.

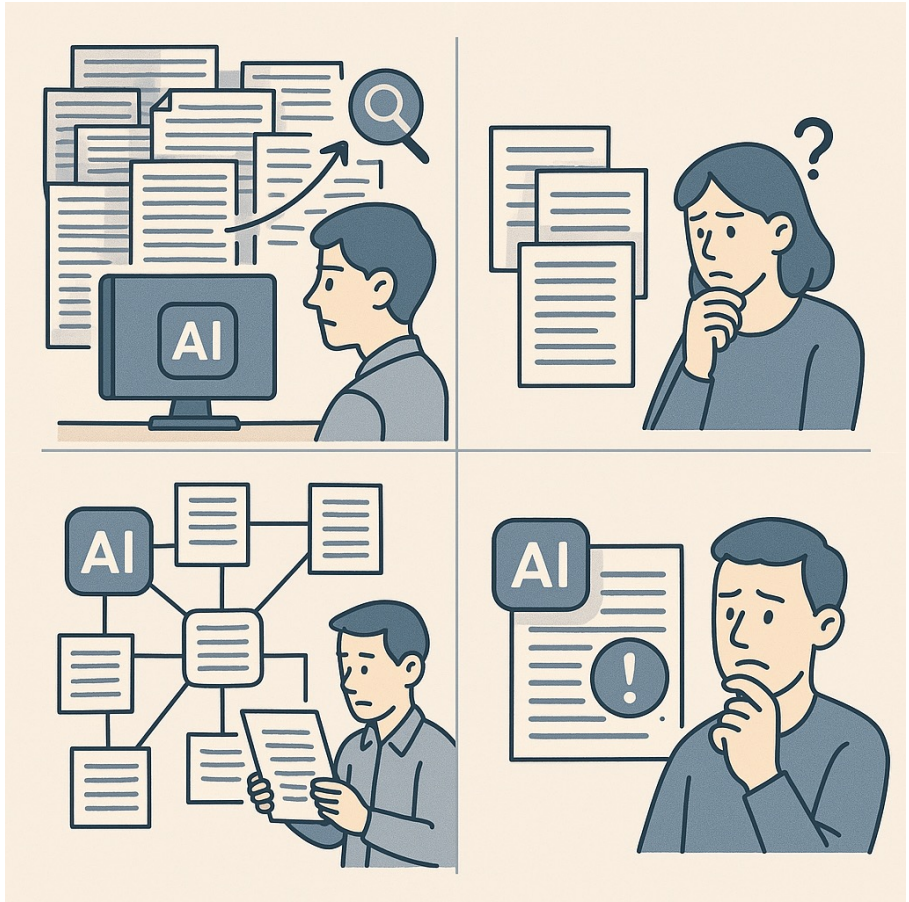
LLM-Related Papers

Liang et al., 2025. SurveyX: Academic survey automation via large language models. arXiv:2502.14776.
Wang et al., 2024. Autosurvey: Large language models can automatically write surveys. NeurIPS 2024.

自动化文献调研的核心挑战

- **AI自动化文献调研的核心挑战：**

文献膨胀
搜索复杂度显著增加



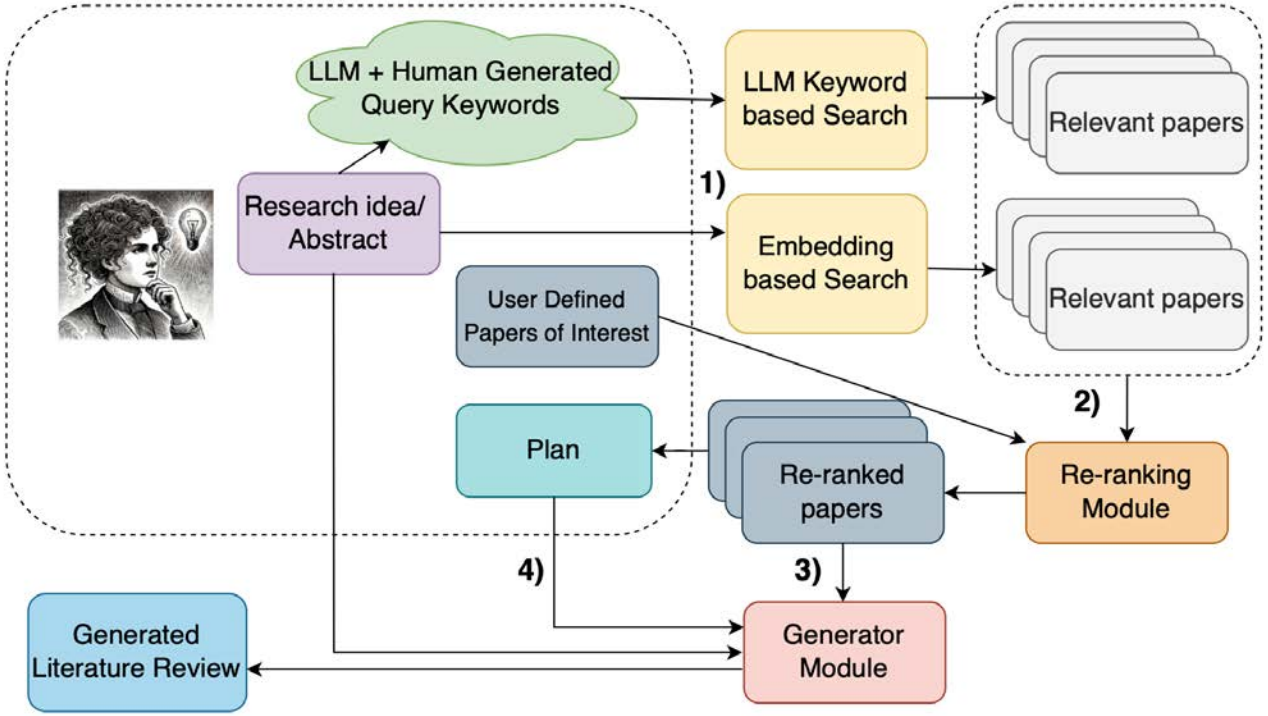
文献理解需要专业知识
准确召回和排序难度大

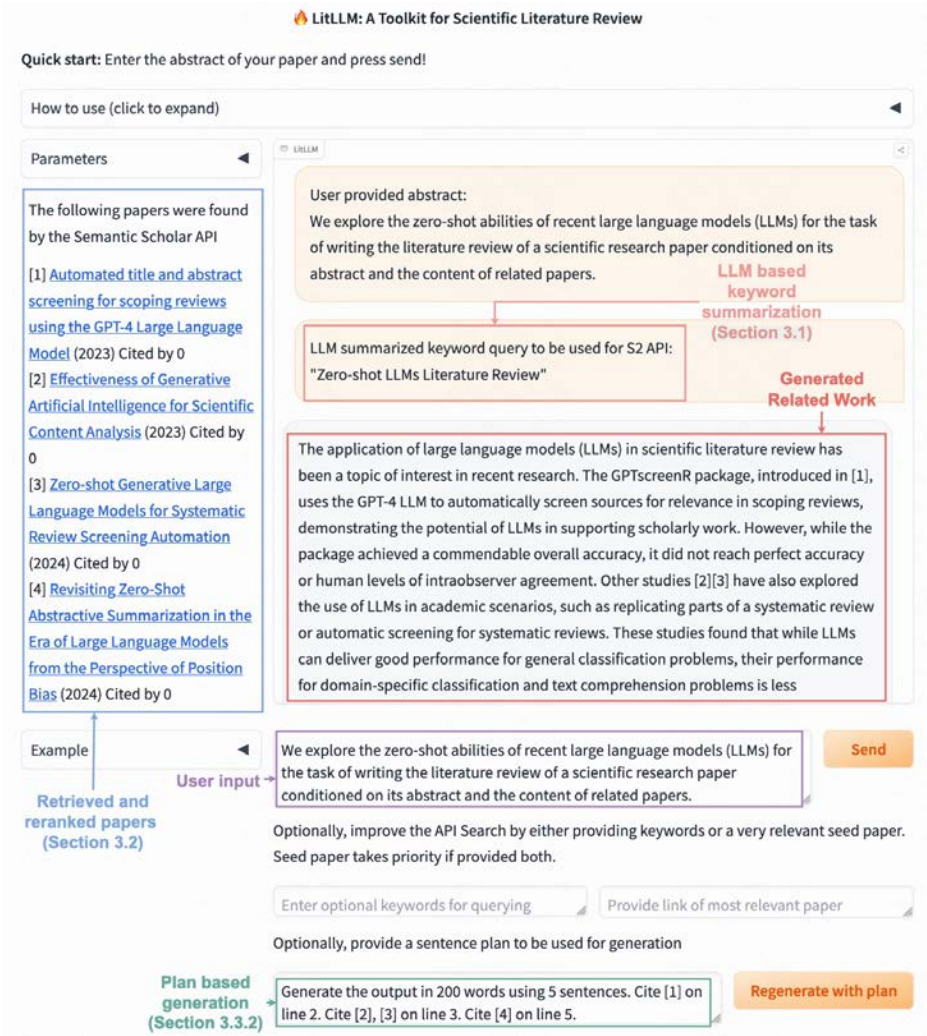
对大量召回文献信息
的综合整理和利用

幻觉性问题
信息不可靠

利用智能体技术自动化文献调研

- 大模型作为辅助接入学术搜索引擎
  - 用户意图理解
  - 整合大规模学术信息

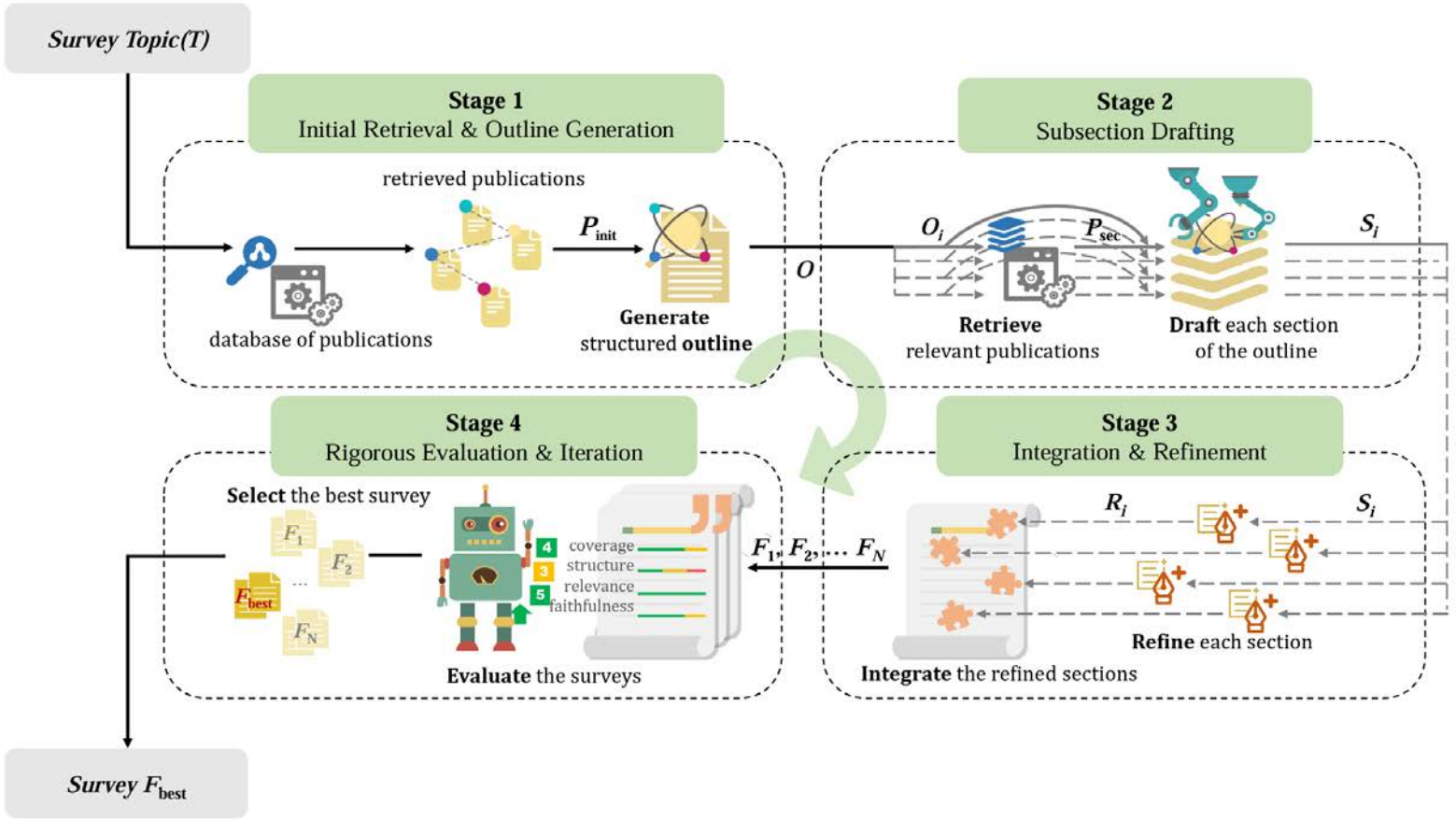




Agarwal et al., 2024. LitLLM: A toolkit for scientific literature review. arXiv:2402.01788.

利用智能体技术自动化文献调研

- 随着基础模型能力的提升，文献调研智能体逐渐从概念走向应用。



完善大模型写作框架
- 先写大纲
- 分段撰写
- 反思优化

提高综述的写作质量

Wang et al., 2024. AutoSurvey: Large language models can automatically write surveys. NeurIPS 2024.

利用智能体技术自动化文献调研

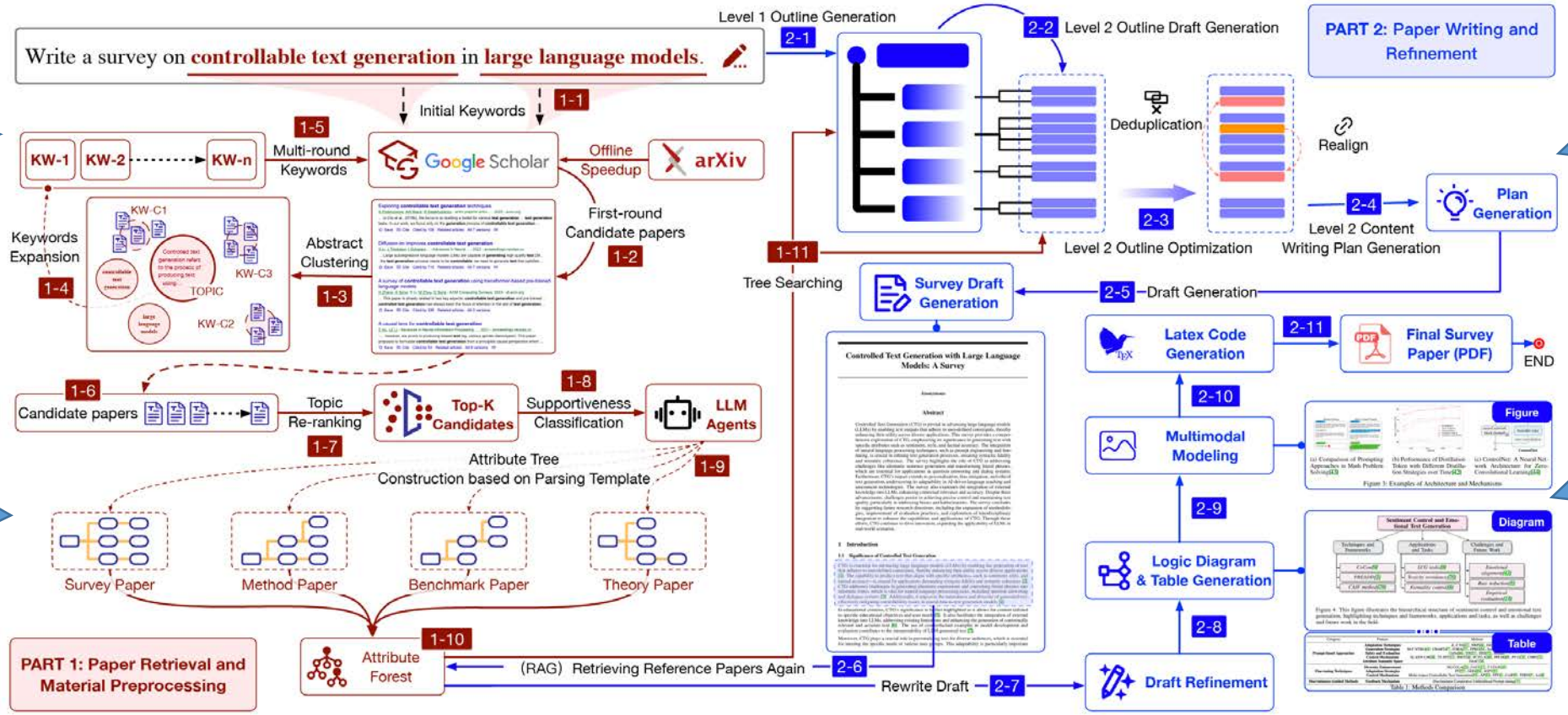
- **随着基础模型能力的提升，文献调研智能体逐渐从概念走向应用。**

多词多轮
文献召回

架构完善
写作流程

结构化
文献解析

多模态
报告生成



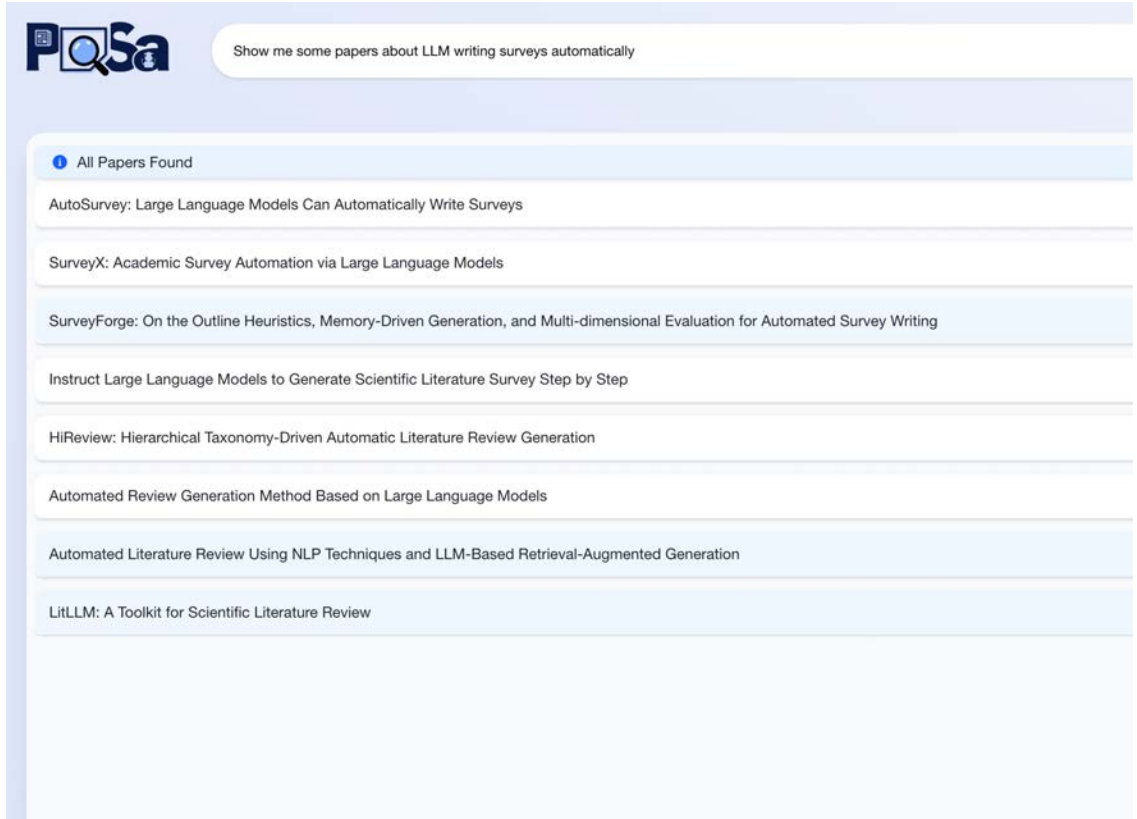
系统复杂度上升，提高文献收集的准确度和综述撰写质量

Liang et al., 2025. SurveyX: Academic survey automation via large language models. arXiv:2502.14776.

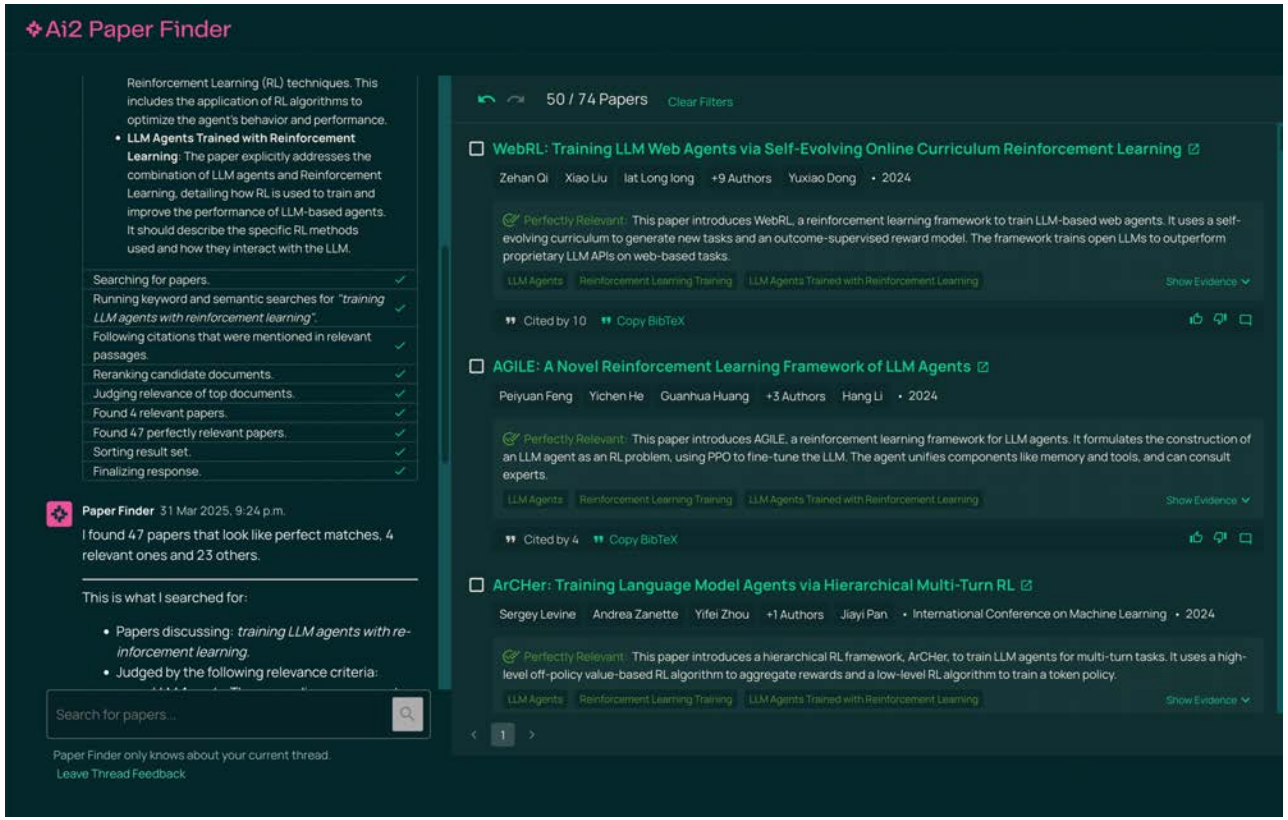
产品形态的 Paper Search Agent

Tsinghua University

- **专注于论文检索的信息搜集智能体已达到商业化水平。**



Paper Search Agent

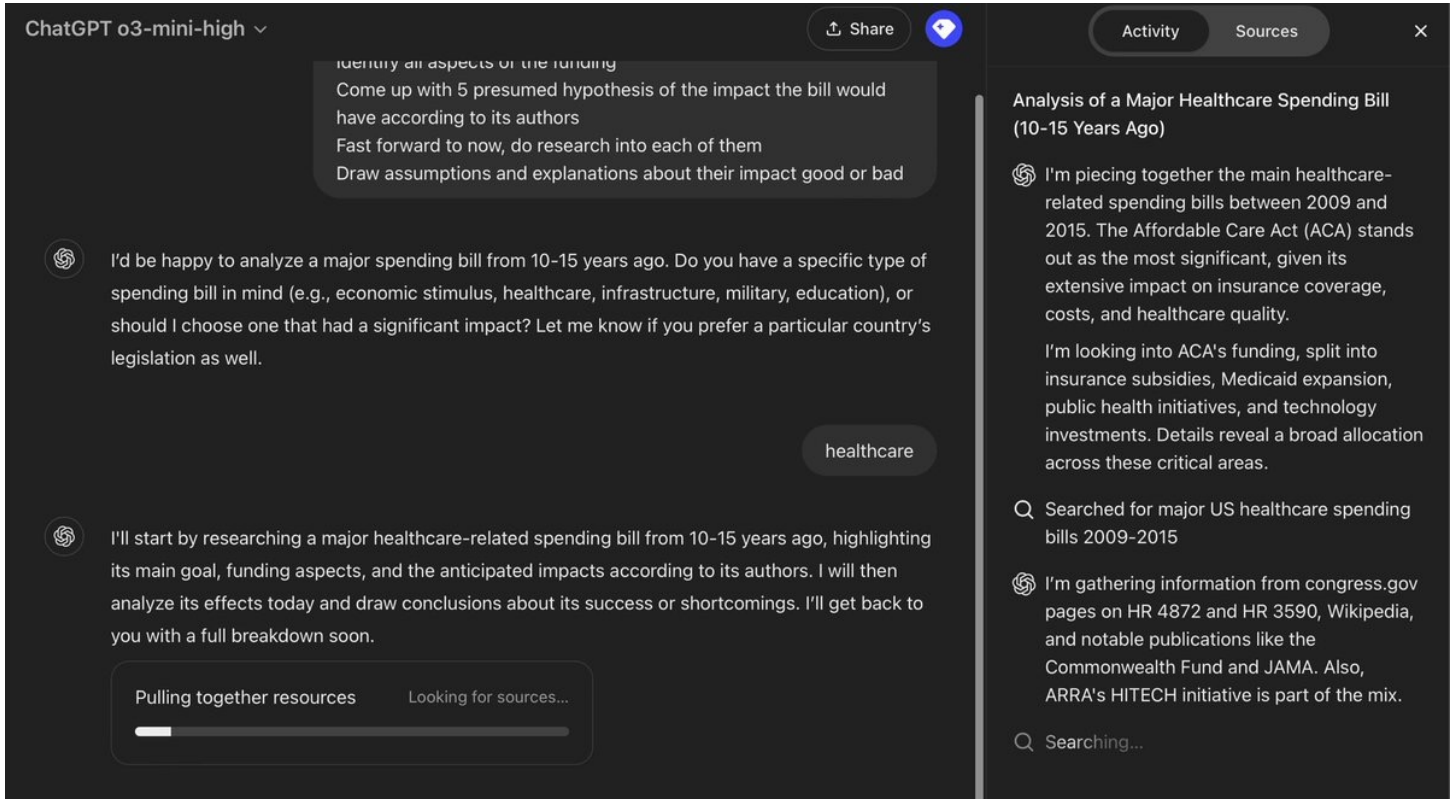


AI2 Paper Finder

https://pasa-agent.ai ； https://paperfinder.allen.ai/chat.

终极形态？OpenAI Deep Research

- **利用推理模型汇集海量信息，完成多步调研任务。**



Model	Accuracy (%)
GPT-4o	3.3
Grok-2	3.8
Claude 3.5 Sonnet	4.3
Gemini Thinking	6.2
OpenAI o1	**9.1**
DeepSeek-R1*	9.4
OpenAI o3-mini (medium)*	10.5
OpenAI o3-mini (high)*	13.0
OpenAI deep research**	**26.6**

* Model is not multi-modal, evaluated on text-only subset.
**with browsing + python tools

Complex Problem Solving

Humanity's Last Exams

https://x.com/danshipper/status/1886203397004783996 ； https://openai.com/index/introducing-deep-research/.

提出idea

科学研究中的Ideation

- 在科学研究过程中，什么是Ideation?
  - 人类在阅读过程中时常会注意关键部分，那类似的注意力机制是否能有效缓解语言模型的灾难性遗忘?
  - 细菌能够记住病毒的DNA片段并用酶入侵DNA，那人类是否能设计类似的系统对任意基因进行切割?
  - 社交平台的推荐算法不断地推荐相似的内容，如果信息获取越来越单一，人们是否会变得更极端、更难以理解不同群体?
- Ideation的过程是对某一领域存在的问题提出潜在的解决方案。

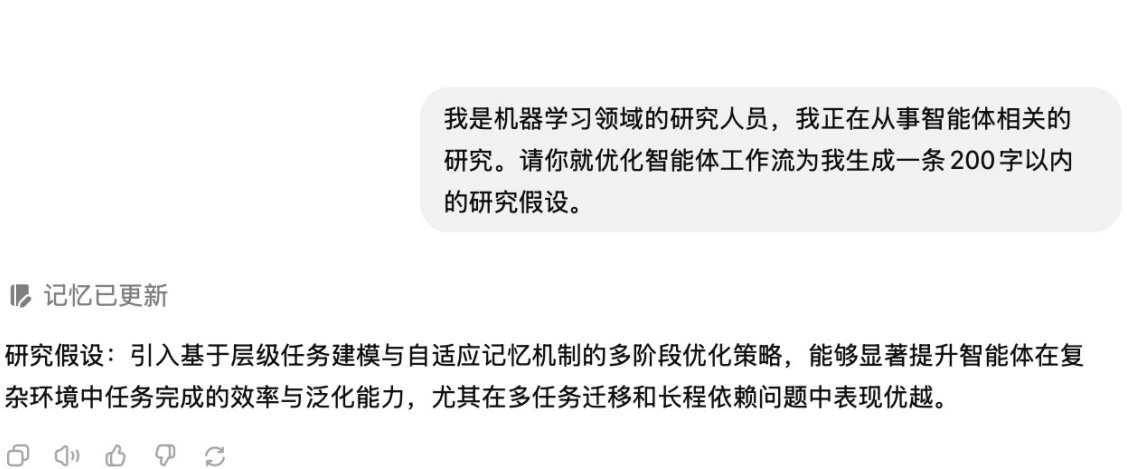
AI用于生成科学假设

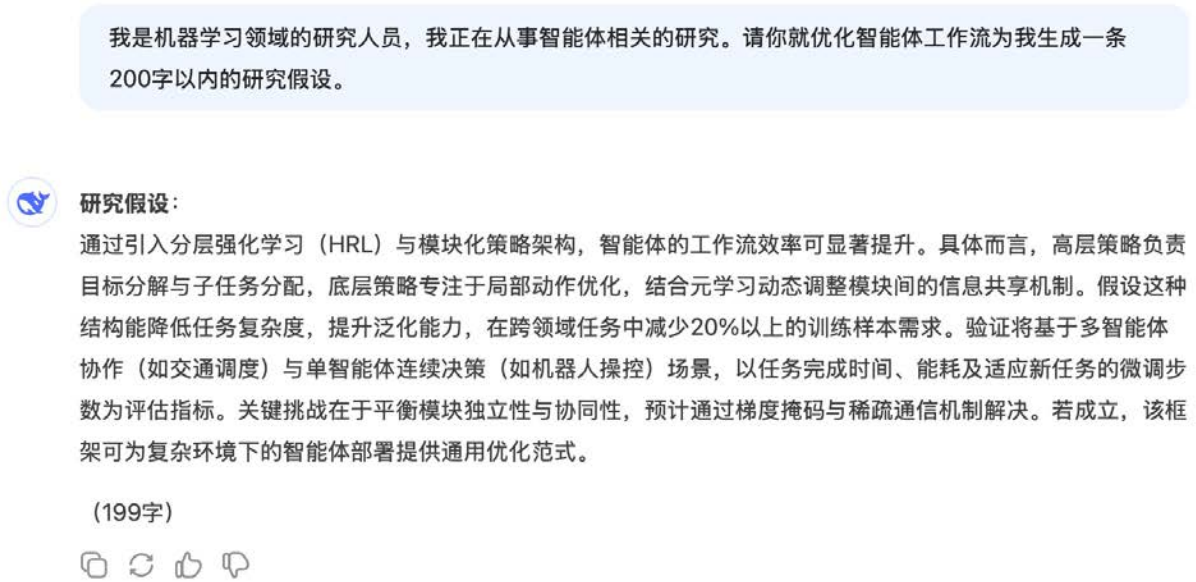
清華大學
Tsinghua University

- 早在四十多年前，文本挖掘技术就被用于生成假设。
  - 若已有的研究成果表明，A能够导致B，且B能够导致C，我们则可以得出如下的假设：A可能会导致C的发生。
  - 一款名为Arrowsmith的软件能够在已发表论文的集合中寻找这种间接联系。该软件曾提出一个假设：鱼油能够降低血液黏稠度，可能有助于治疗雷诺综合征，而随后的实验证实了这一假设是正确的。
- 然而这一系统存在许多缺陷：需要基于人类预先发现的规律、使用大量文献数据、生成的假设仅适用于特定场景等。
- 大语言模型的出现很大程度上弥补了这些缺陷。

LLM时代的Ideation

- 大语言模型经过大规模语料训练，涵盖不同学科领域的专业知识。
- 在单一领域上，大语言模型同样可能具备人类专家并不了解的知识。
- 我们可以直接prompt大语言模型生成研究假设。





LLM时代的Ideation

- **研究者早在2023年就开始尝试利用大语言模型生成研究假设，但效果受限于当时模型的能力。**

Type	Content
Input (Dong et al., 2022)	*seed term*: **speech unit boundaries** ; *context* (abridged): ... generate partial sentence translation given a streaming speech input. existing approaches ... break the acoustic units in speech, as boundaries between acoustic units in speech are not even. ...
Initial idea	A pause prediction model to identify **speech unit boundaries** ...
Iteration 1	A method that leverages acoustic and linguistic features to predict **speech unit boundaries** dynamically, ensuring smooth transitions ... differs from the existing research as it combines both acoustic properties and linguistic context ... adapting to variations in speaker characteristics, speaking styles, and languages.
Iteration 2	A novel method called Adaptive **Speech Unit Boundary** Detection (ASUBD) ... a combination of attention mechanisms to focus on relevant acoustic and linguistic features and reinforcement learning to guide the system to make optimal predictions of unit boundaries based on previous decisions...
Ground Truth	... an efficient monotonic segmentation module ... accumulate acoustic information incrementally and detect proper **speech unit boundaries**.

Wang et al., 2023. SciMON : Scientific Inspiration Machines Optimized for Novelty. ACL 2024.

LLM时代的Ideation

- 研究者早在2023年就开始尝试利用大语言模型生成研究假设，但效果受限于当时模型的能力。

Type	Content
Input (Dong et al., 2022)	*seed term*: **speech unit boundaries** ; *context* (abridged): ... generate partial sentence translation given a st … ndaries between a…
Initial idea	A…
Iteration 1	A… e… li… … ies dynamically, ic properties and … nguages.
Iteration 2	A… n… s… … ation of attention … ning to guide the
Ground Truth	... an efficient monotonic segmentation module ... accumulate acoustic information incrementally and detect proper **speech unit boundaries**.

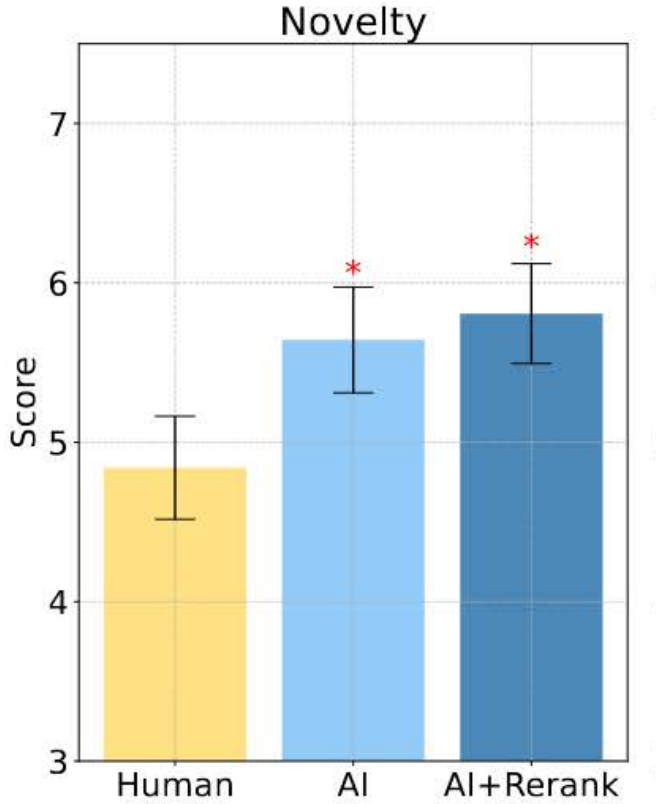
Type 相较于Ground Truth	Meditron	+SN	+CT	+KG
Helpful(%)	35	80	60	50
Unhelpful(%)	65	20	40	50
vs. GT(%)	30	45	50	35

Wang et al., 2023. SciMON : Scientific Inspiration Machines Optimized for Novelty. ACL 2024.

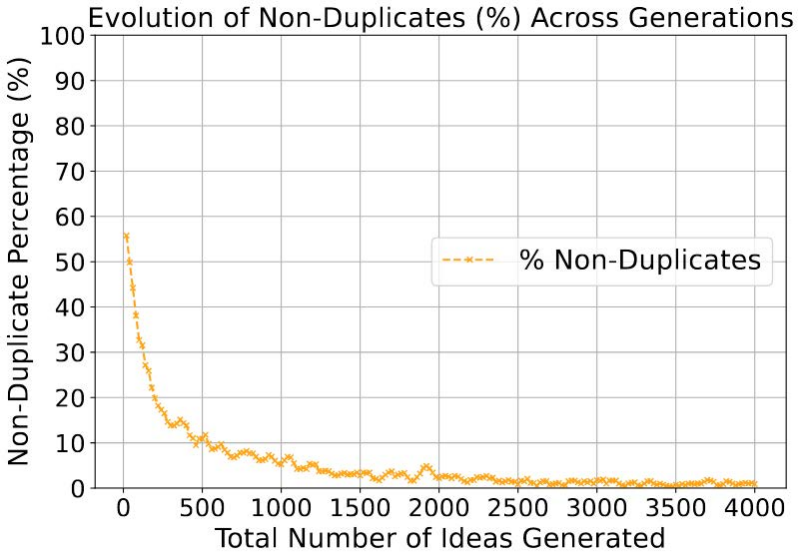
LLM时代的Ideation

- **大模型展现出了生成高质量idea的潜力，但仍有很大挑战。**

一定条件下可以生成创新性优于人类的idea



大模型生成idea的多样性存在明显不足



通过增加大模型生成idea的数量来提升不同idea数量的效率低下

大模型难于稳定评判其所生成idea的质量

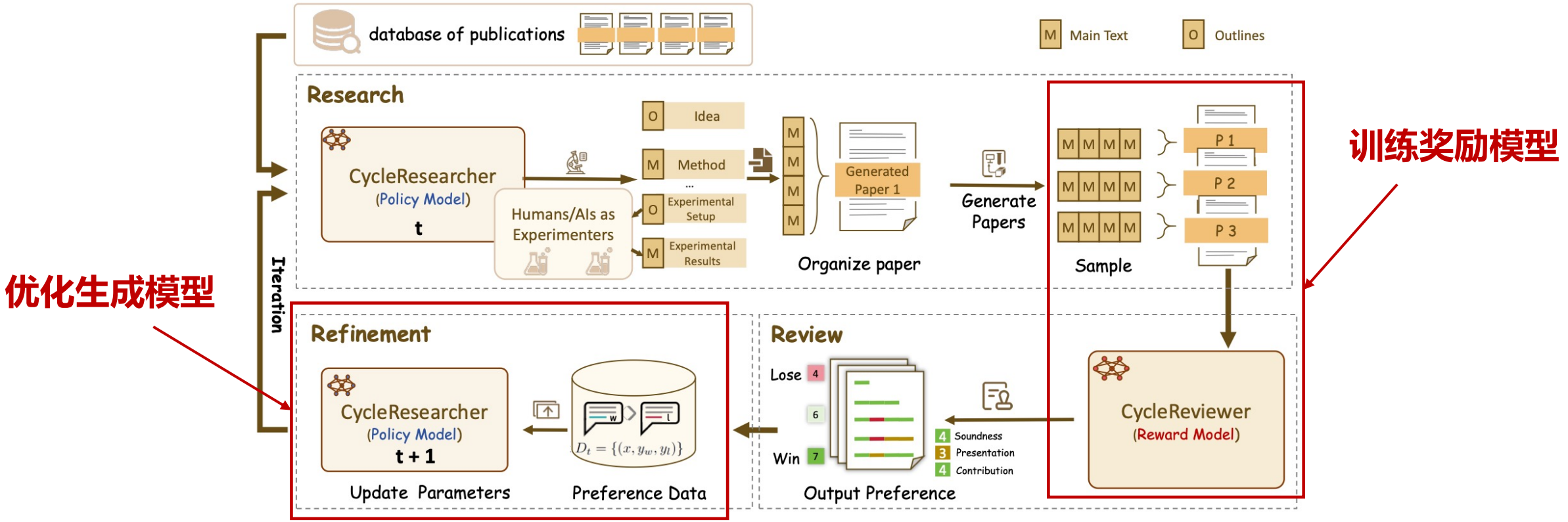
	Consistency
Random	50.0
NeurIPS'21	66.0
ICLR'24	71.9
Ours	56.1
GPT-4o Direct	50.0
GPT-4o Pairwise	45.0
Claude-3.5 Direct	51.7
Claude-3.5 Pairwise	53.3
"AI Scientist" Reviewer	43.3

上：人-人评判一致性
下：人-模型评判一致性

Lu et al., 2024. The AI scientist: Towards fully automated open-ended scientific discovery. arXiv:2408.06292.

优化大语言模型生成研究假设

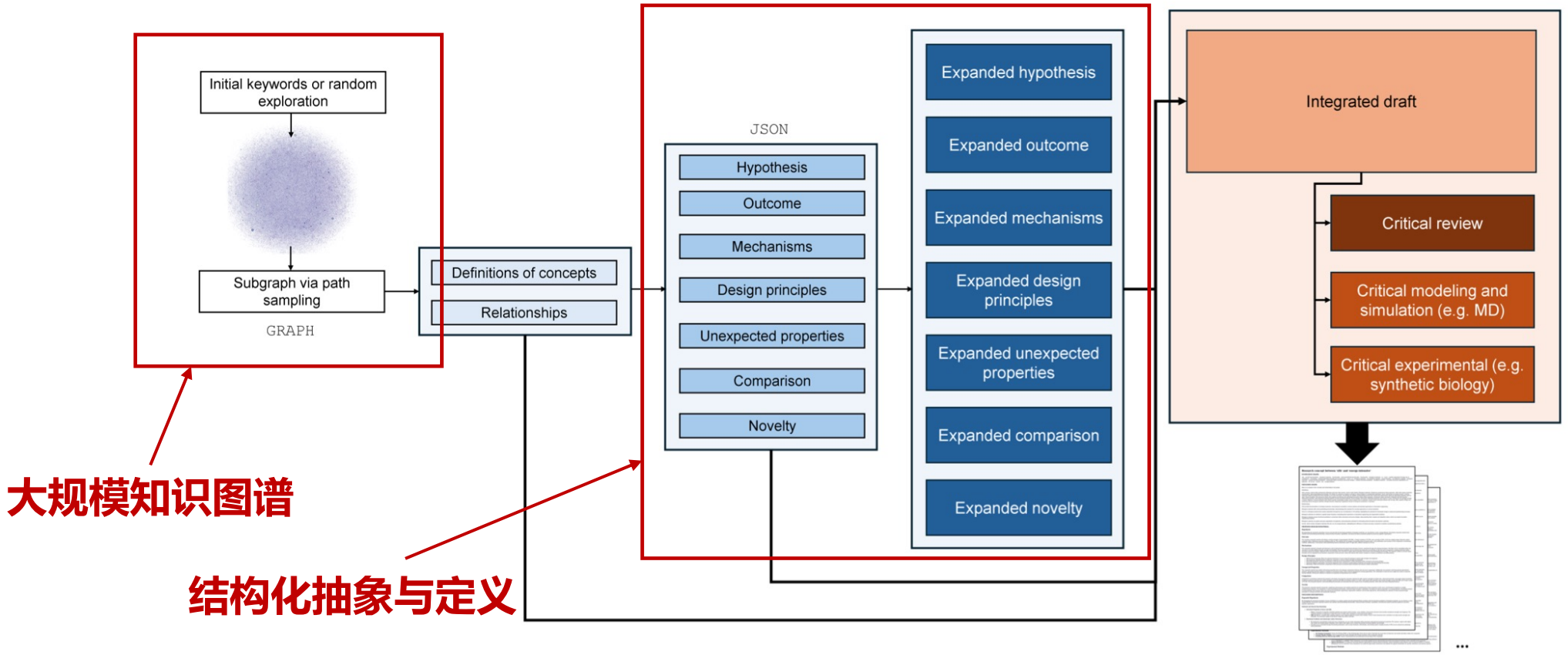
- 基于公开论文数据训练奖励模型，进而指导生成假设模型的迭代优化。



Wang et al., 2024. CycleResearcher: Improving Automated Research via Automated Review. arXiv:2411.00816.

优化大语言模型生成研究假设

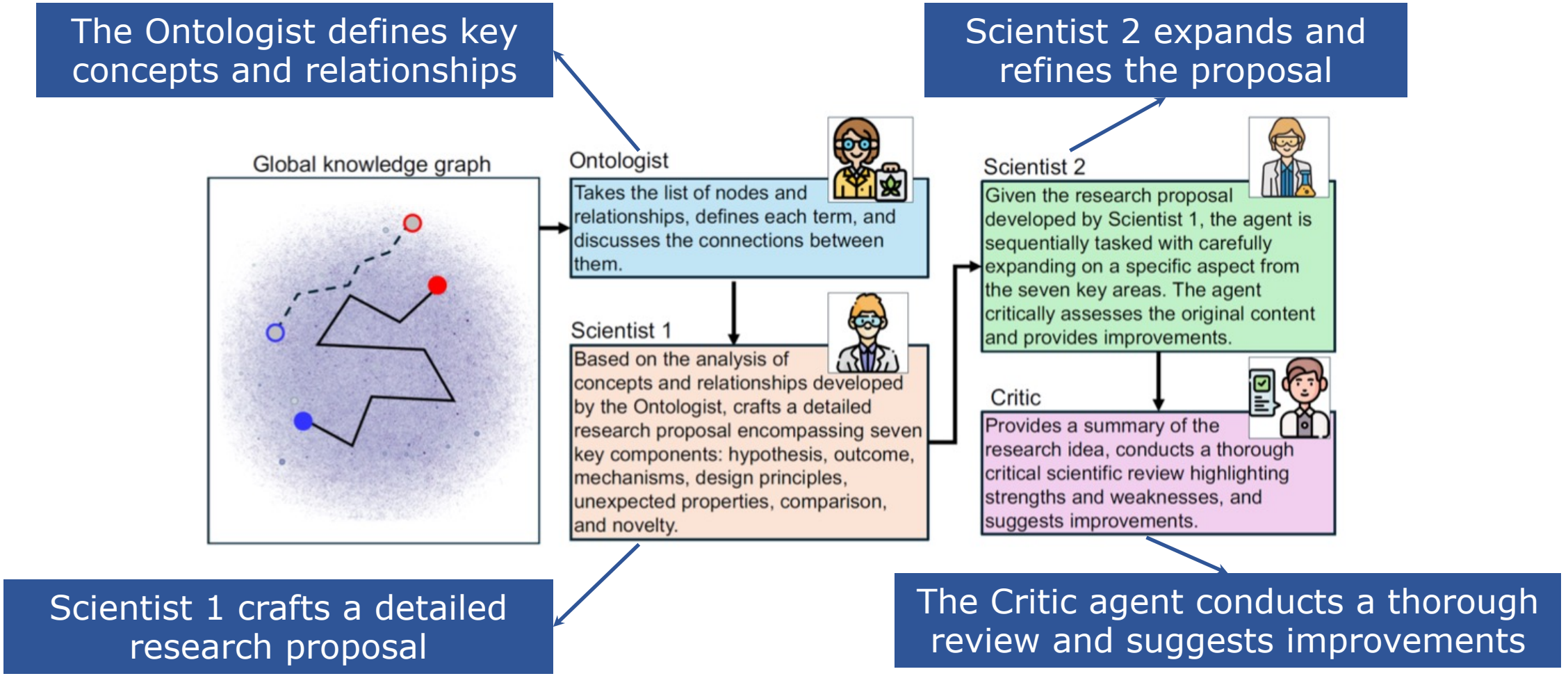
- 使用大规模本体知识图谱来组织和关联多样的科学概念，同时对科学研究过程进行结构化抽象与定义。



Ghafarollahi et al., 2024. SciAgents: Automating Scientific Discovery Through Multi-agent Intelligent Graph Reasoning. arXiv:2409.05556.

优化大语言模型生成研究假设

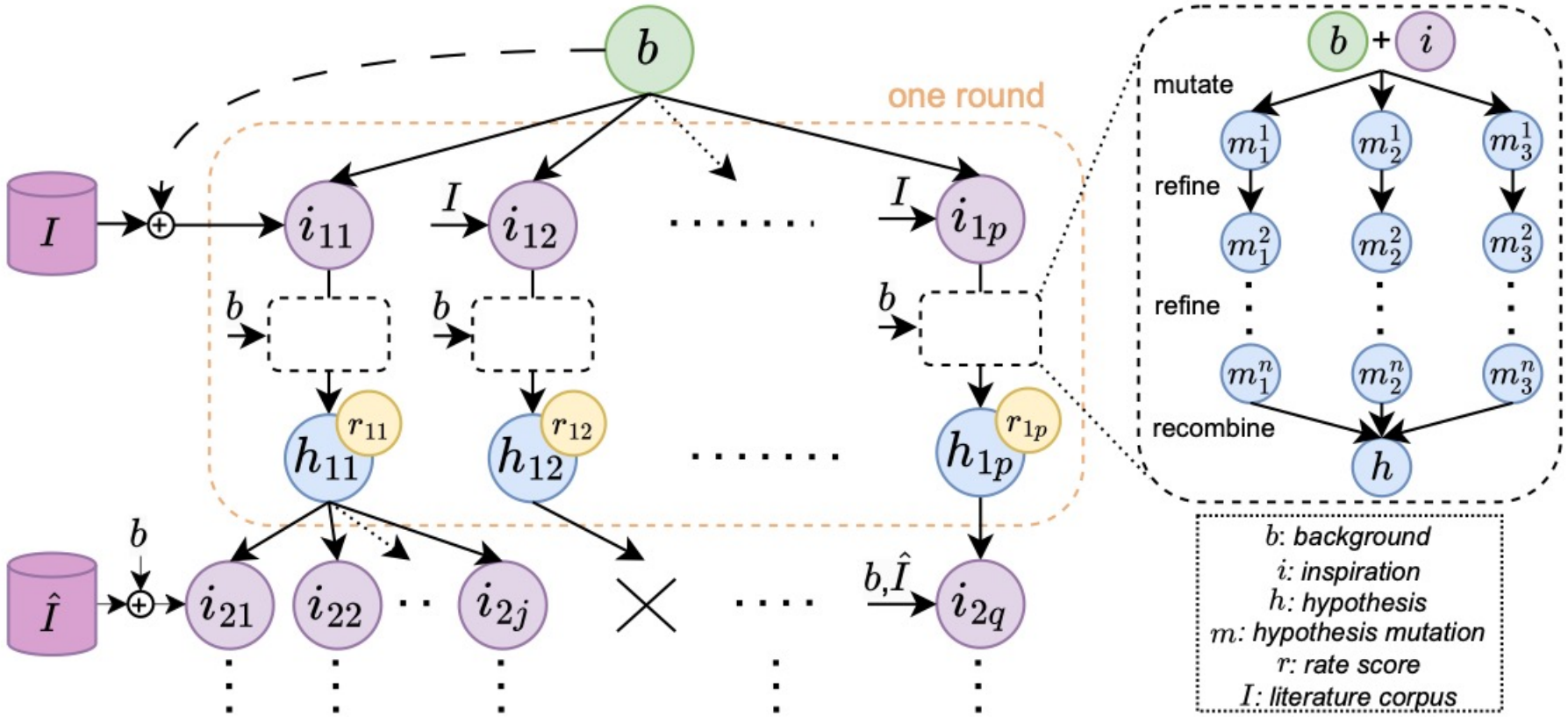
- 使用多智能体系统进一步激发大语言模型的能力。



Ghafarollahi et al., 2024. SciAgents: Automating Scientific Discovery Through Multi-agent Intelligent Graph Reasoning. arXiv:2409.05556.

优化大语言模型生成研究假设

- **部分科学创新来源于知识重组和联想，例如很多化学研究假设由研究背景和研究灵感组成。**

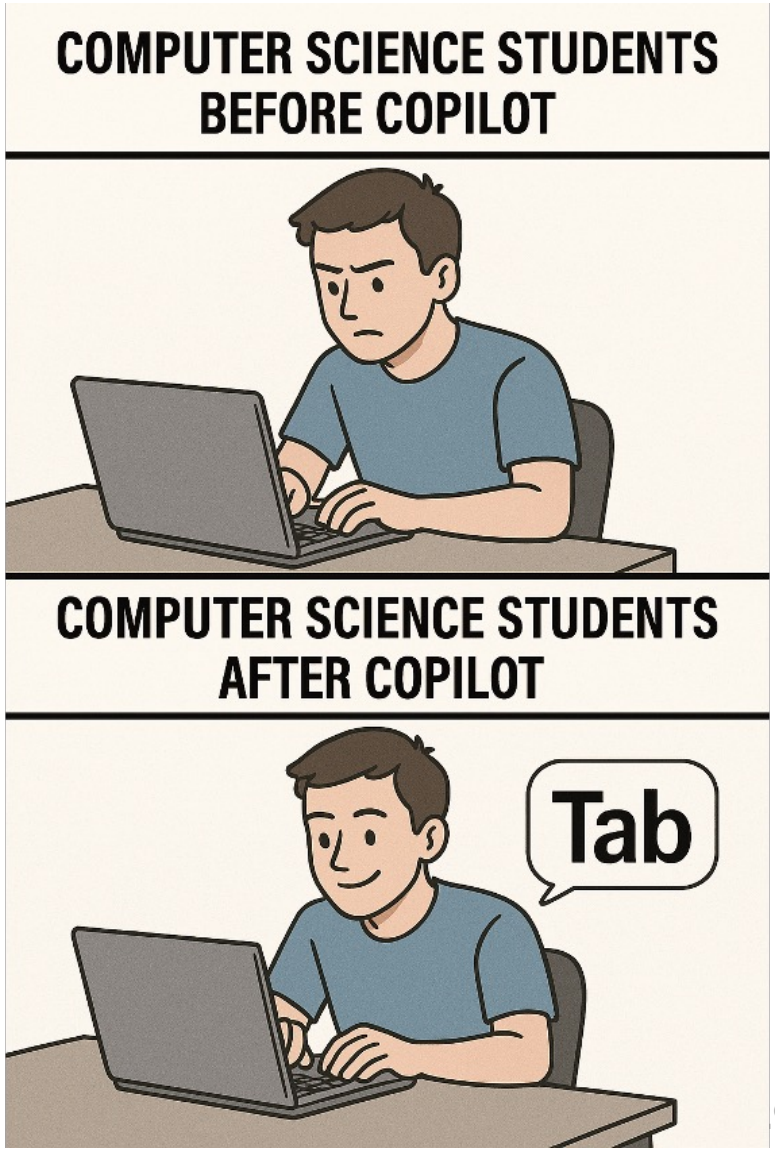


Yang et al., 2024. MOOSE-Chem: Large Language Models for Rediscovering Unseen Chemistry Scientific Hypotheses. ICLR 2025.

执行实验

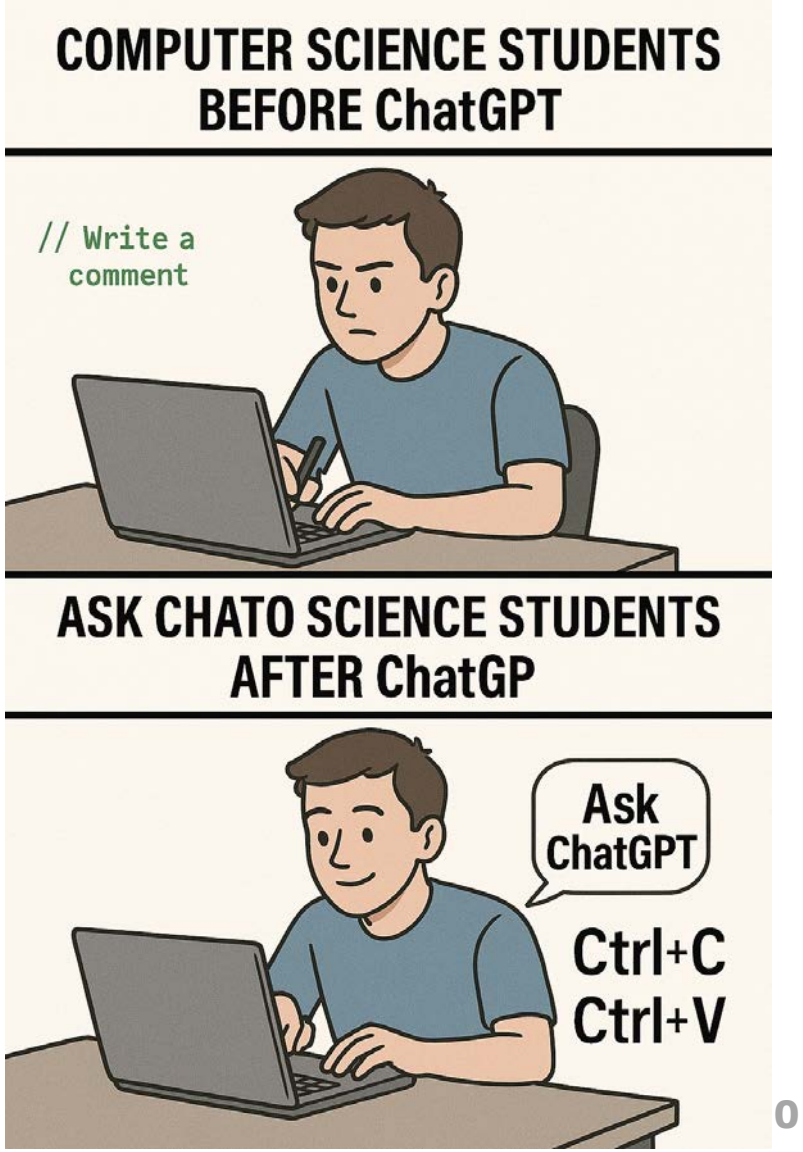
利用 AI 做实验 —— 程序员的变化

- Github Copilot
  - 写注释, then TabTabTab



利用 AI 做实验 —— 程序员的变化

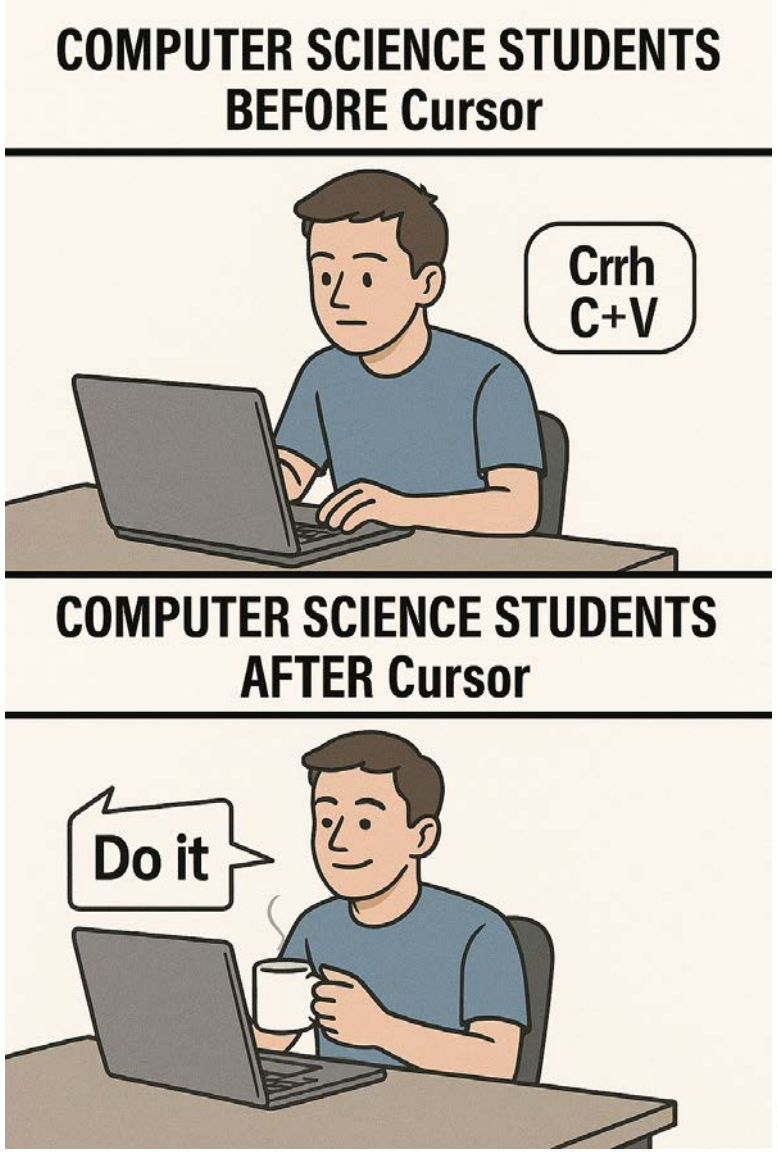
- Github Copilot
  - 写注释, then TabTabTab
- ChatGPT / Claude3.7
  - Help！Help！I need xxx, then ctrl+c & ctrl+v



利用 AI 做实验 —— 程序员的变化

- Github Copilot
  - 写注释, then TabTabTab
- ChatGPT / Claude3.7
  - Help！ Help！ I need xxx, then ctrl+c & ctrl+v
- Cursor
  - Do it! then look, drink, and smile.

趋势：从辅助驾驶到自动驾驶



利用智能体执行实验

- 从 Benchmark 看变化趋势：
  - SWE-Bench：让智能体解决 Github Issue

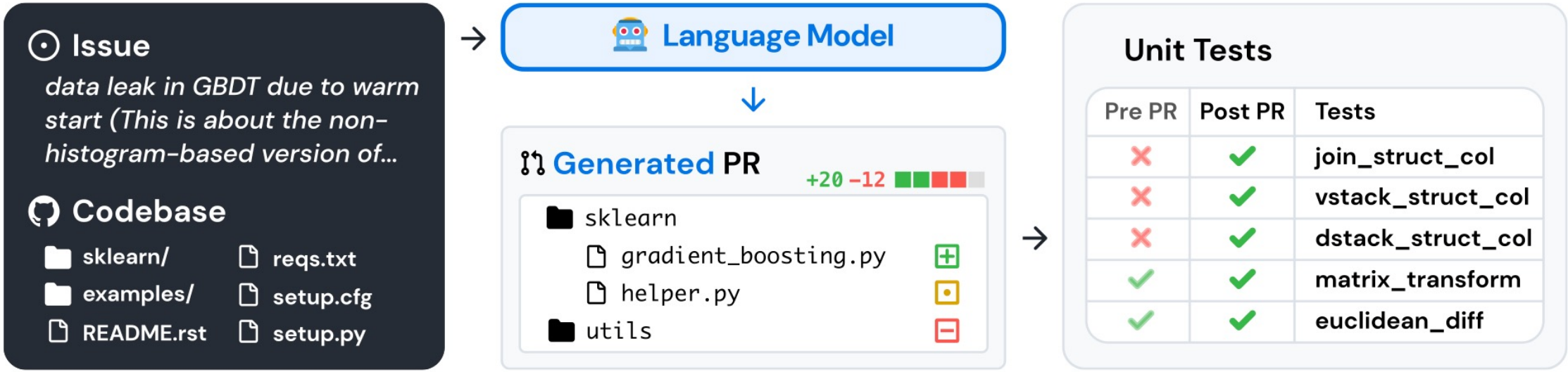


Figure 1: SWE-bench sources task instances from real-world Python repositories by connecting GitHub issues to merged pull request solutions that resolve related tests. Provided with the issue text and a codebase snapshot, models generate a patch that is evaluated against real tests.

Jimenez et al. 2023. SWE-Bench: Can language models resolve real-world github issues?. ICLR 2024.

利用智能体执行实验

- 从 Benchmark 看变化趋势：
  - MLE-Bench：让智能体实现 ML 算法

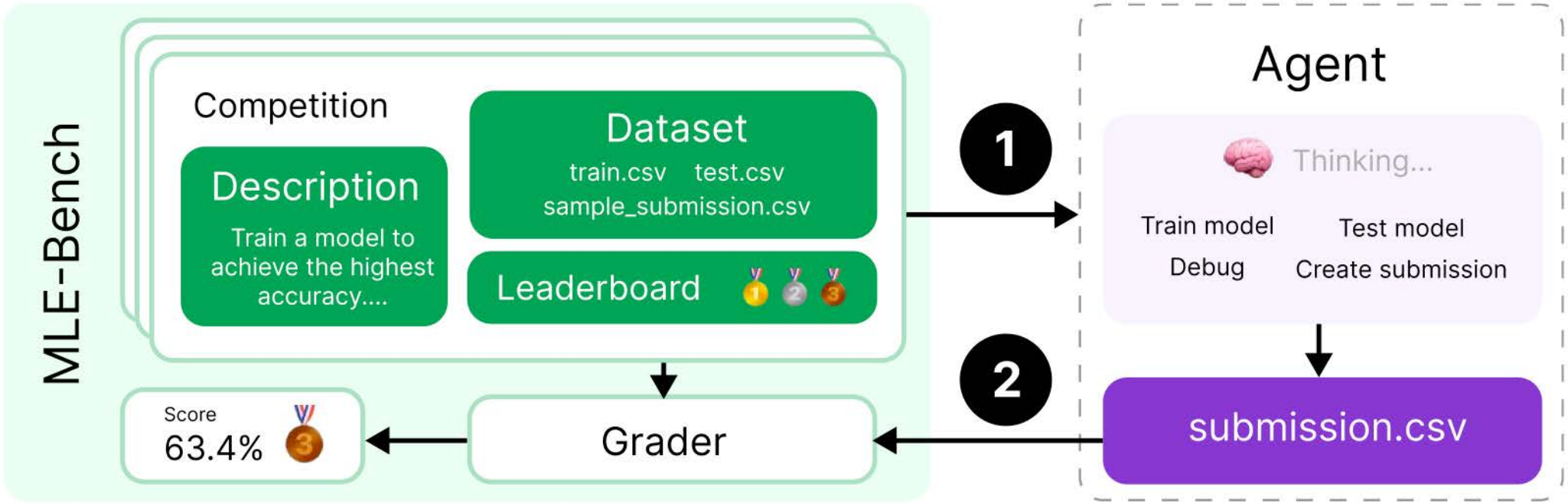
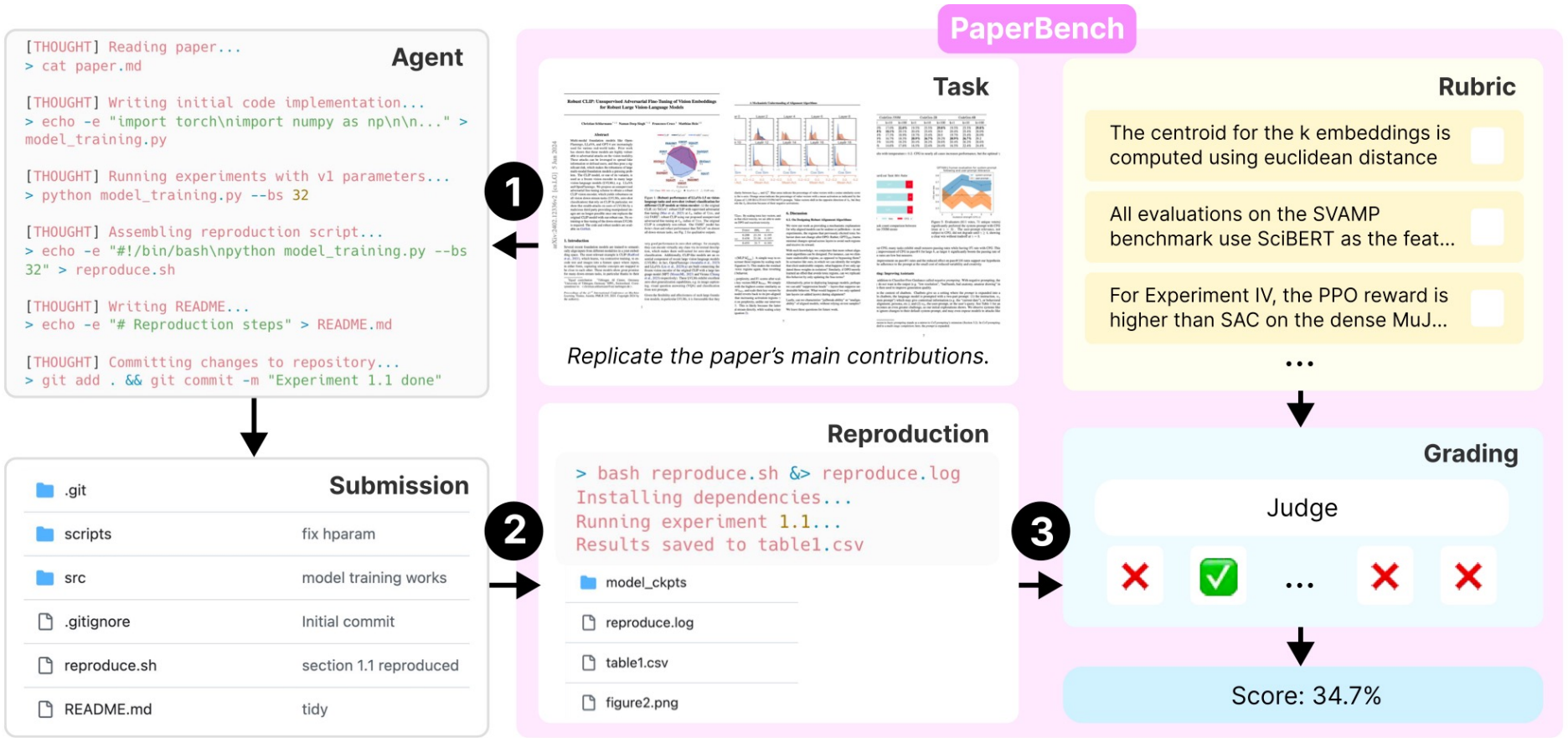


Figure 1: MLE-bench is an offline Kaggle competition environment for AI agents. Each competition has an associated description, dataset, and grading code. Submissions are graded locally and compared against real-world human attempts via the competition's leaderboard.

Chan, J. S. et al. 2024. MLE-bench: Evaluating machine learning agents on machine learning engineering. ICLR 2025.

利用智能体执行实验

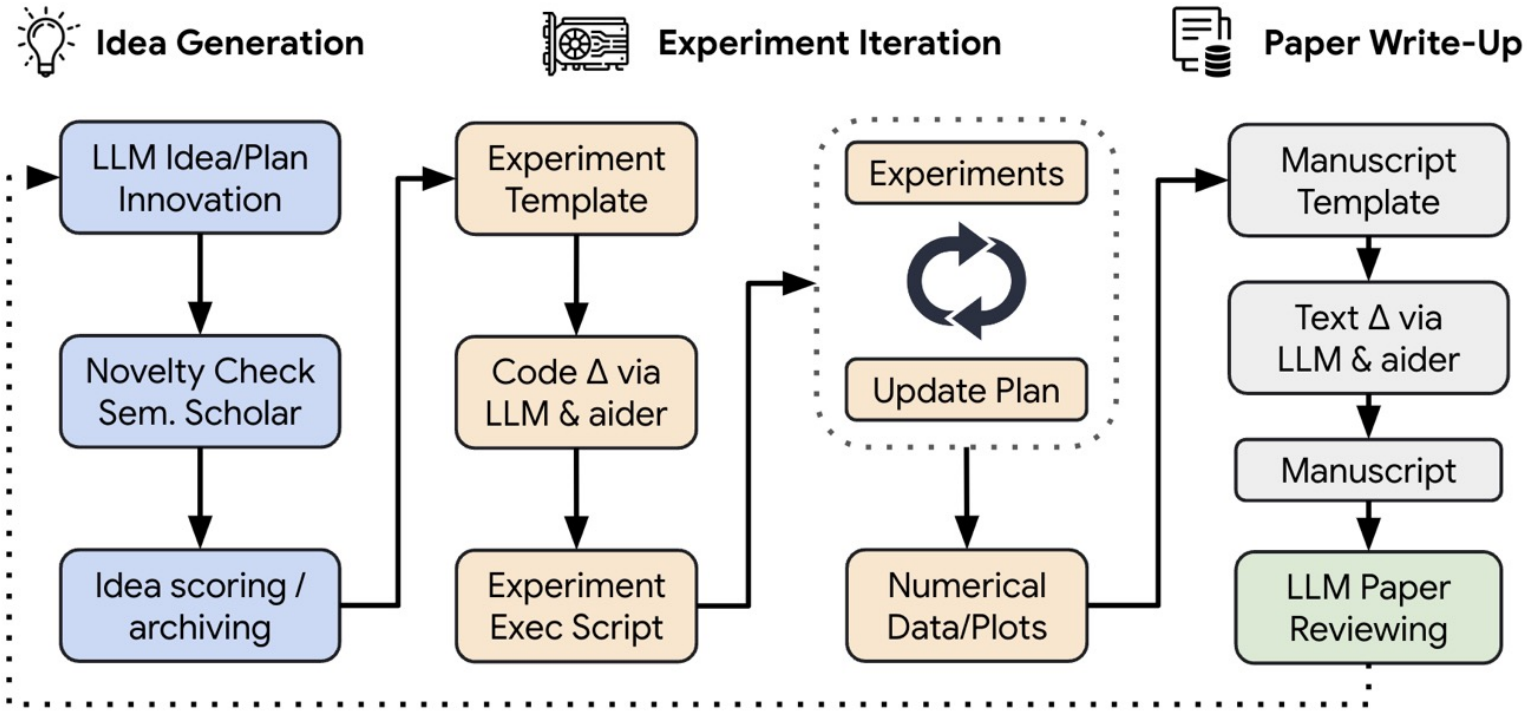
- 从 Benchmark 看变化趋势：
  - PaperBench：让智能体复现 Paper



Giulio Starace, et al., 2025. PaperBench: Evaluating AI's Ability to Replicate AI Research. arXiv:2504.01848.

将智能体自动化实验用于科学发现

- **主流框架：**
  - **人工设定明确的实验流程模板（Experiment Template）；**
  - **利用 Aider/OpenHands 等工具对整个代码库进行自动化修改；**
  - **循环迭代进行自动化实验。**
- **AI-Scientist：**



Lu, Chris, et al.,2024 The AI Scientist: Towards fully automated open-ended scientific discovery. arXiv:2408.06292.

当前主流框架面临的挑战

- **挑战一：低可运行成功率**
  - **自动化代码易出现语法、环境配置、库依赖等问题；**
  - **复杂场景下难以一次性生成可成功运行的代码。**

面对简单编程任务大模型也可能出现的部分语义错误

Error Characteristic		Example of Incorrect Solutions	Correct Solution

Conditional Error	If error	`# [Task 151] CodeGen-16b`
		`def double_the_difference(lst):`
		`  sum = 0`
		`  for i in lst:`
		`  if i > 0 and i % 2 != 0:`
		`  sum += i ** 2`
		`  return sum`
			`# [Task 151] Ground Truth`
			`def double_the_difference(lst):`
			`  ans = 0`
			`  for num in lst:`
			`  if num%2==1 and num>0 and "." not in str(num): ans += num ** 2`
			`  return ans`

Loop Error	For error	`# [Task 121] GPT-3.5`
		`def solution(lst):`
		`  sum = 0`
		`  for i in range(1, len(lst), 2):`
		`  if lst[i] % 2 != 0: sum +=lst[i]`
		`  return sum`
			`# [Task 121] Ground Truth`
			`def solution(lst):`
			`  return sum([x for idx, x in enumerate(lst) if idx%2==0 and x%2==1])`

	While error		

Return Error	Incorrect return value	`# [Task 103] GPT-3.5`
		`def rounded_avg(n, m)`
		`  if n > m: return -1`
		`  avg=round(sum(range(n,m+1))/(m-n+1))`
		`  return bin(avg)[2:]`
			`# [Task 103] Ground Truth`
			`def rounded_avg(n, m)`
			`  if n > m: return -1`
			`  avg = round((n + m) / 2)`
			`  return bin(avg)`

Method Call Error	Incorrect function name	`# [Task 54] StarCoder`
		`def same_chars(s0, s1):`
		`  return sorted(s0) == sorted(s1)`
			`# [Task 54] Ground Truth`
			`def same_chars(s0, s1):`
			`  return set(s0) == set(s1)`

	Incorrect function arguments		
	Incorrect method call target		

Assignment Error	Incorrect constant	`# [Task 138] InCoder-1.3B`
		`def is_equal_to_sum_even(n):`
		`  return n >= 4 and n % 2 == 0`
			`# [Task 138] Ground Truth`
			`def is_equal_to_sum_even(n):`
			`  return n >= 8 and n % 2 == 0`

	Incorrect arithmetic		
	Incorrect variable name		
	Incorrect comparison		

Import Error	Import error	`# [Task 133] StarCoder`
		`def sum_squares(lst):`
		`  return sum[int(math.ceil(i)**2 for i in lst]`
			`# [Task 133] Ground Truth`
			`def sum_squares(lst):`
			`  import math`
			`  return sum(map(lambda x: math.ceil(x)**2, lst))`

Code Block Error	Incorrect code block	`# [Task 83] InCoder-1.3B`
		`def starts_one_ends(n):`
		`  count = 0`
		`  while n > 0:`
		`  count, n = count + 1, n / 10`
		`  return count`
			`# [Task 83] Ground Truth`
			`def starts_one_ends(n):`
			`  if n == 1:`
			`  return 1`
			`  return 18 * 10 ** (n - 2)`

	Missing code block	`# [Task 60] CodeGen-16B`
		`def next_smallest(lst):`
		`  if len(lst)<2: return None`
		`  lst.sort()`
		`  return lst[1]`
			`# [Task 60] Ground Truth`
			`def is_prime(n):`
			`  if len(lst)<=1: return None`
			`  sorted_list=sorted(lst)`
			`  for x in sorted_list:`
			`  if x!=sorted_list[0]: return x`

Wang, et.al.,2025 Towards Understanding the Characteristics of Code Generation Errors Made by Large Language Models. ICSE2025.

当前主流框架面临的挑战

清華大學 Tsinghua University

- **挑战二：实验与 Idea 一致性无法保证**
  - **AI 自动实现的实验可能偏离原始 Idea。**
  - **缺乏有效的反馈机制和验证手段来确保实验代码真正实现了预期的实验目标。**


```
"Name": "hybrid_filtering_augmentation",
"Title": "Hybrid Filtering and Augmentation for Enhanced Data Curation in
Multi-Turn Conversations",
"Experiment": "1. Apply rule-based filtering using predefined quality
criteria (coherence, relevance, instruction-following, depth, engagement,
grammar, and syntax) to prune low-quality data. 2. Develop an ML-based
component that learns to identify and prioritize high-quality data based
on feedback from MT-Bench evaluations. 3. Integrate the ML component to
further refine and augment the dataset by generating new high-quality data
samples. 4. Train the model on the curated and augmented dataset. 5.
Evaluate the model's performance on MT-Bench and iteratively refine the ML
component based on the evaluation results.",
"Interestingness": 8,
"Feasibility": 8,
"Novelty": 8,
"novel": true
```


AI 提的 Idea：

集成机器学习组件来增强数据

当前主流框架面临的挑战

清華大學
Tsinghua University

实际实现：

```
import argparse

parser = argparse.ArgumentParser(description="Run experiment")
parser.add_argument("--out_dir", type=str, default="run_0", help="Output directory")
args = parser.parse_args()

from typing import List, Literal
import json
from datasets import load_dataset

import concurrent.futures
from tqdm import tqdm
import time
from openai import OpenAI
from liquid import Template

from benchmarking import benchmarking_fn
```

```
import argparse

parser = argparse.ArgumentParser(description="Run experiment")
parser.add_argument("--out_dir", type=str, default="run_0", help="Output directory")
args = parser.parse_args()

from typing import List, Literal
import json
from datasets import load_dataset

def ml_component(text: str, refinement_factor: float = 0.1) -> bool:
    # Refined ML component that gets updated based on previous evaluations
    # For simplification, we're emulating refinement by a fixed factor.
    import random
    return random.random() < (0.9 + refinement_factor)

import concurrent.futures
from tqdm import tqdm
import time
from openai import OpenAI
from liquid import Template

from benchmarking import benchmarking_fn
```

提出的Idea：集成机器学习组件来增强数据

```
def ml_component(text, refinement_factor=0.1):
    import random
    return random.random() < (0.9 + refinement_factor)
```

全面证伪

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?

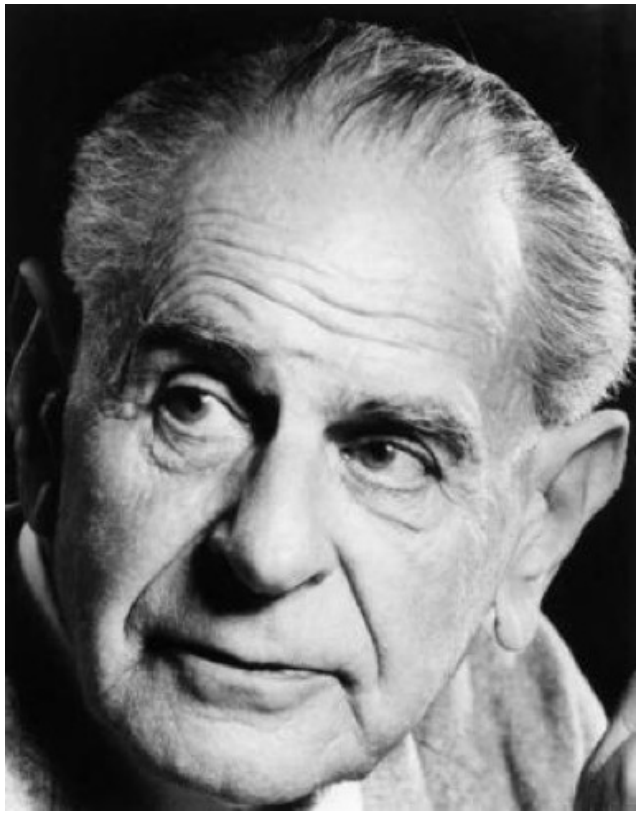
我们为什么会相信AlphaFold产生的结果?

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**。



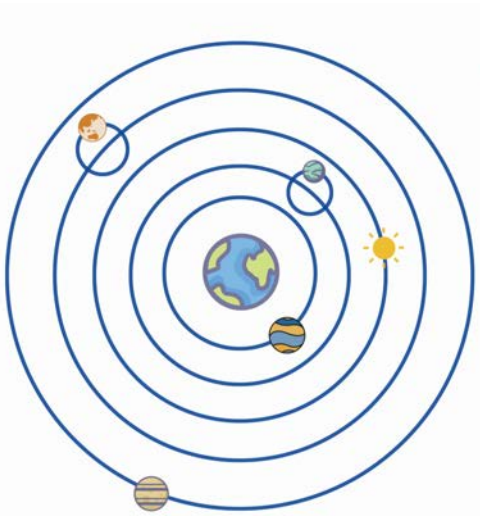
卡尔·波普尔

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

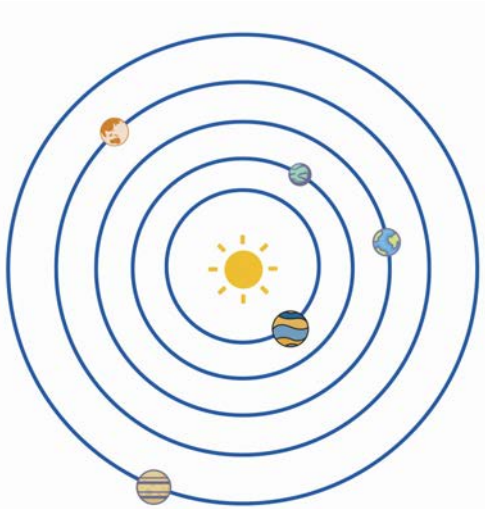


Geocentric Theory:

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

V.S.

日心说



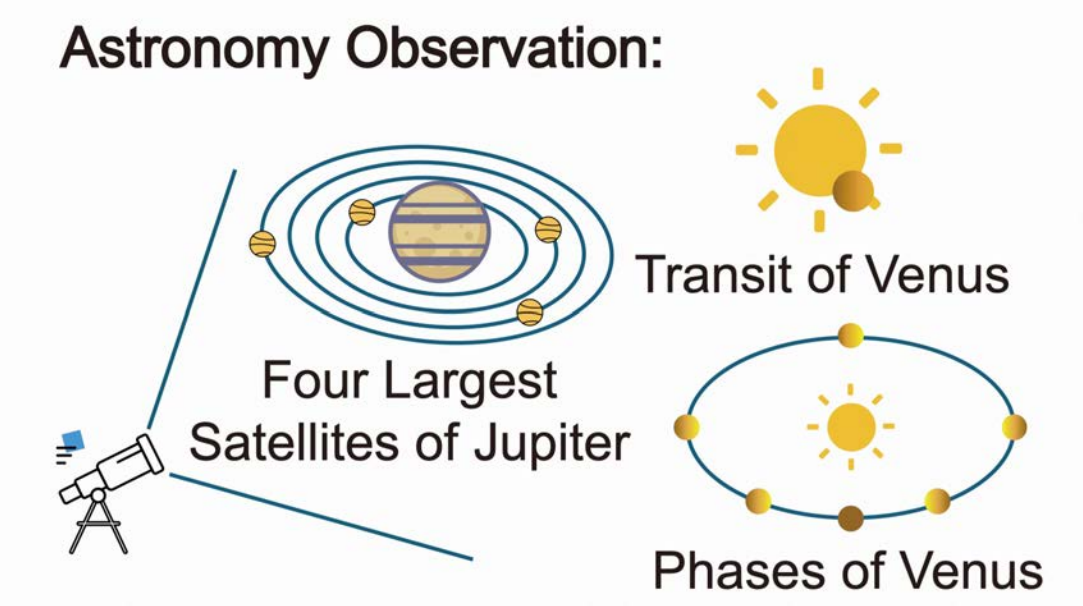
Heliocentric Theory:

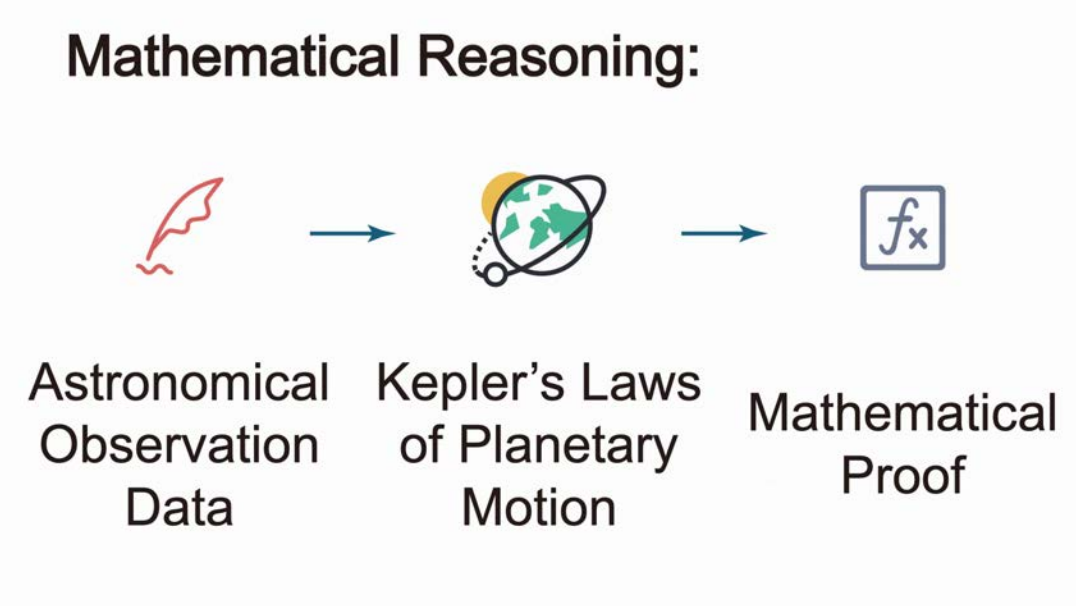
- Earth is rotating;
- Celestial bodies revolve around the Sun;
- Circular motion.
- ...

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

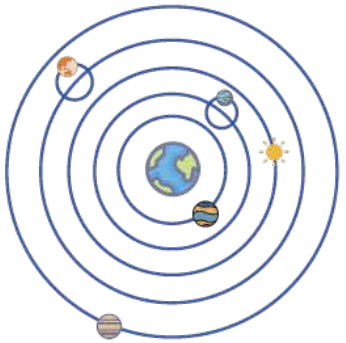
数学形式化推理验证假说

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

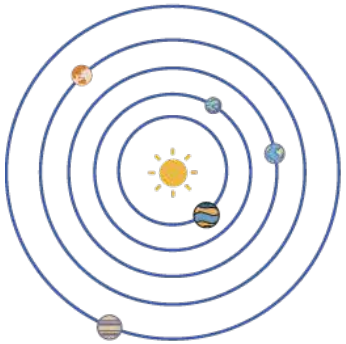
- **只有经过严格的证伪过程，才能产生可靠的科学发现。**

Hypothesis



Geocentric Theory:

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

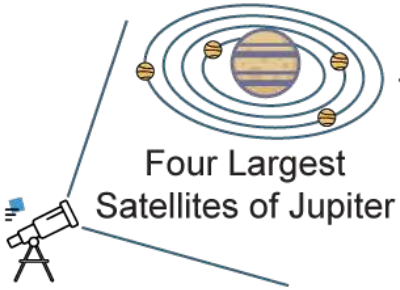


Heliocentric Theory:

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

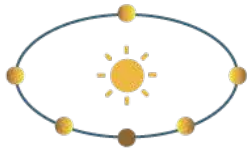
Falsification

Astronomy Observation:



Four Largest Satellites of Jupiter

Transit of Venus



Phases of Venus

Mathematical Reasoning:

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

Scientific Discovery

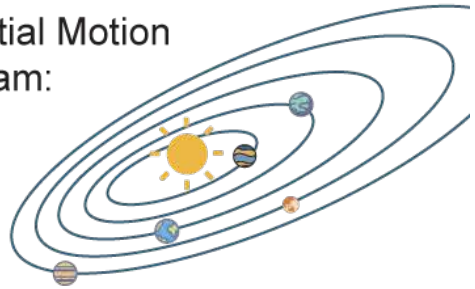
Final Discovery:

- 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.

一些人工智能经典研究工作的证伪

- **Devlin et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2019.**
  - 主实验：对比OpenAI SOTA, BiLSTM, GPT

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	**Average** -
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

  - 消融实验：改变任务、特征、模型，总结预训练研究的核心科学发现

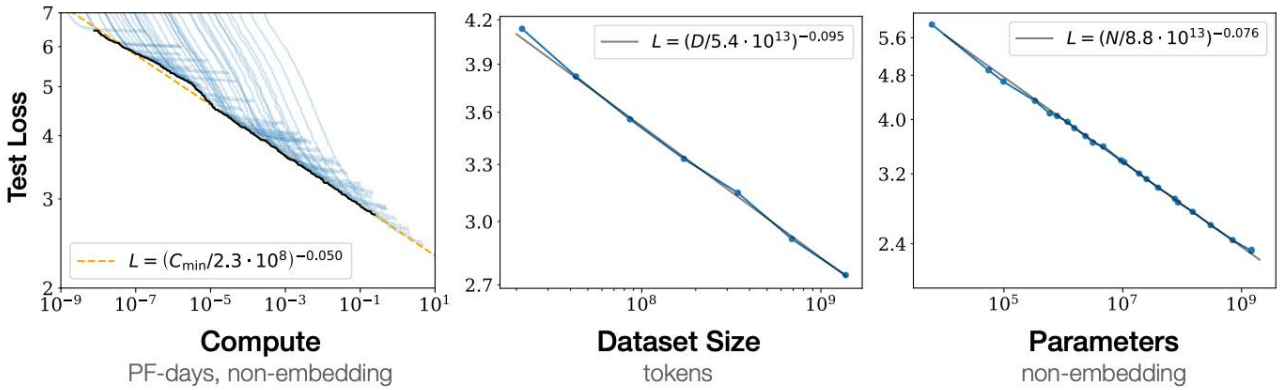
Tasks	Dev Set MNLI-m (Acc)	QNLI (Acc)	MRPC (Acc)	SST-2 (Acc)	SQuAD (F1)
$\text{BERT}_{\text{BASE}}$	84.4	88.4	86.7	92.7	88.5
No NSP	83.9	84.9	86.5	92.6	87.9
LTR & No NSP	82.1	84.3	77.5	92.1	77.8
+ BiLSTM	82.1	84.1	75.7	91.6	84.9

Fine-tuning approach		
$\text{BERT}_{\text{LARGE}}$	96.6	92.8
$\text{BERT}_{\text{BASE}}$	96.4	92.4
Feature-based approach ($\text{BERT}_{\text{BASE}}$)		
Embeddings	91.0	-
Second-to-Last Hidden	95.6	-
Last Hidden	94.9	-
Weighted Sum Last Four Hidden	95.9	-
Concat Last Four Hidden	96.1	-
Weighted Sum All 12 Layers	95.5	-

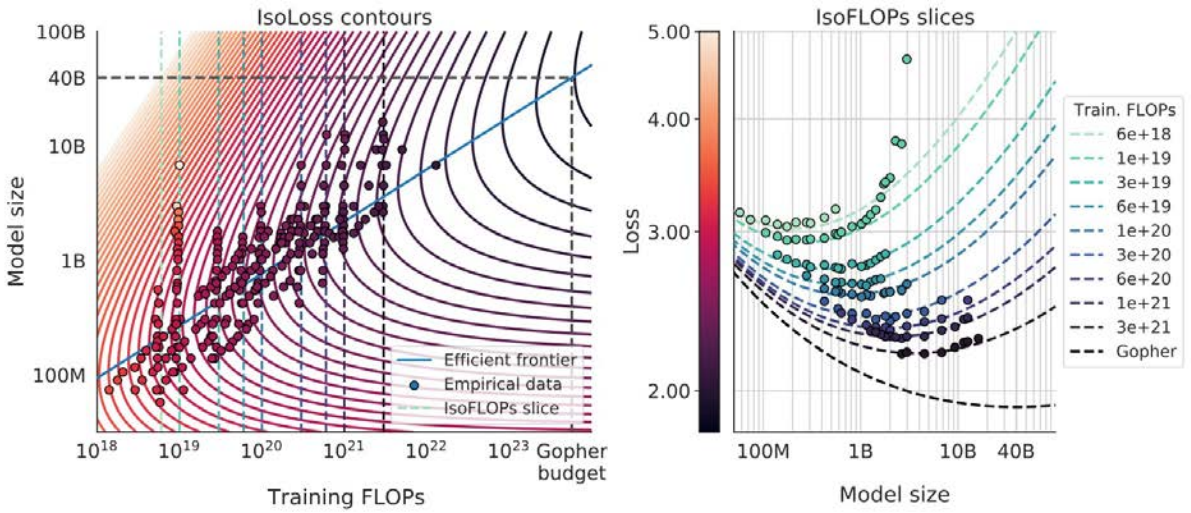
Hyperparams				Dev Set Accuracy		
#L	#H	#A	LM (ppl)	MNLI-m	MRPC	SST-2
3	768	12	5.84	77.9	79.8	88.4
6	768	3	5.24	80.6	82.2	90.7
6	768	12	4.68	81.9	84.8	91.3
12	768	12	3.99	84.4	86.7	92.9
12	1024	16	3.54	85.7	86.9	93.3
24	1024	16	3.23	86.6	87.8	93.7

一些人工智能经典研究工作的证伪

- **Kaplan et al., Scaling Laws for Neural Language Models. 2020.**
  - 详细检验不同维度的缩放定律



- **Hoffmann et al., Training compute-optimal large language models. 2022.**
  - 详细探索不同资源下的训练最优曲线



现有智能体系统产生的证伪

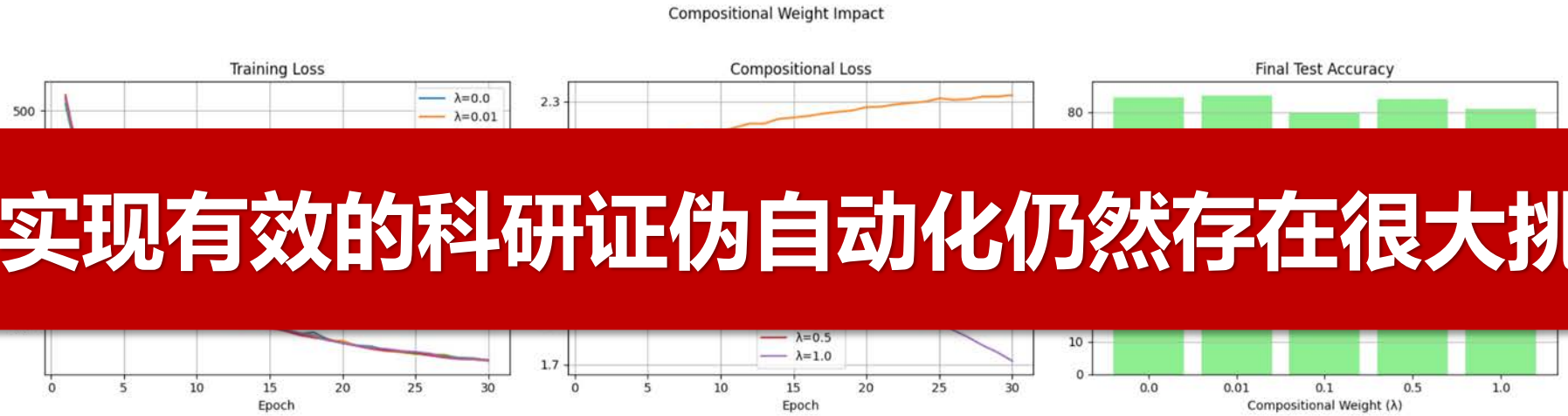
- **AI Scientist. Compositional Regularization: Unexpected Obstacles in Enhancing Neural Network Generalization. 2025.**
- *投稿于ICLR2025 Workshop*

摘要

Neural networks excel in many tasks but often struggle with **compositional generalization-the ability to understand and generate novel combinations of familiar components.** This limitation hampers their performance on tasks requiring systematic reasoning beyond the training data. In this work, **we introduce a training method that incorporates an explicit compositional regularization term into the loss function**, aiming to encourage the network to develop compositional representations. Contrary to our expectations, our experiments on synthetic arithmetic expression datasets reveal that models trained with compositional regularization do not achieve significant improvements in generalization to unseen combinations compared to baseline models. Additionally, we find that increasing the complexity of expressions exacerbates the models' difficulties, regardless of compositional regularization. These findings highlight the challenges of enforcing compositional structures in neural networks and suggest that such regularization may not be sufficient to enhance compositional generalization.

现有智能体系统产生的证伪

- **AI Scientist. Compositional Regularization: Unexpected Obstacles in Enhancing Neural Network Generalization. 2025.**
- *投稿于ICLR2025 Workshop*
- 消融实验：改变超参数设置



如何实现有效的科研证伪自动化仍然存在很大挑战

- **结论：方法效果不符合预期。**

实现人工智能研究全自动证伪的挑战

清華大學 Tsinghua University

- 对于理科、人文类研究，证伪大多通过已有文献支撑和逻辑推理实现。
- 对于自然科学研究，证伪大多通过控制实验环境、仪器观察分析实现。
- 对于**人工智能研究**，证伪基于**给定的任务方向和构造好的方法**，**提出可能假设、设计实现实验流程，选择或设计评价指标**，进而总结分析。

- **挑战：**
  1. **执行效率：得到确定结论前，探索假设空间的尺度大小有限**
  2. **实验设计：准确分析因素后，正确实现对比方法和消融实验**
  3. **指标选择：得到实验结果时，合理应用反映求证结论的指标**

相关工作的证伪自动化程度

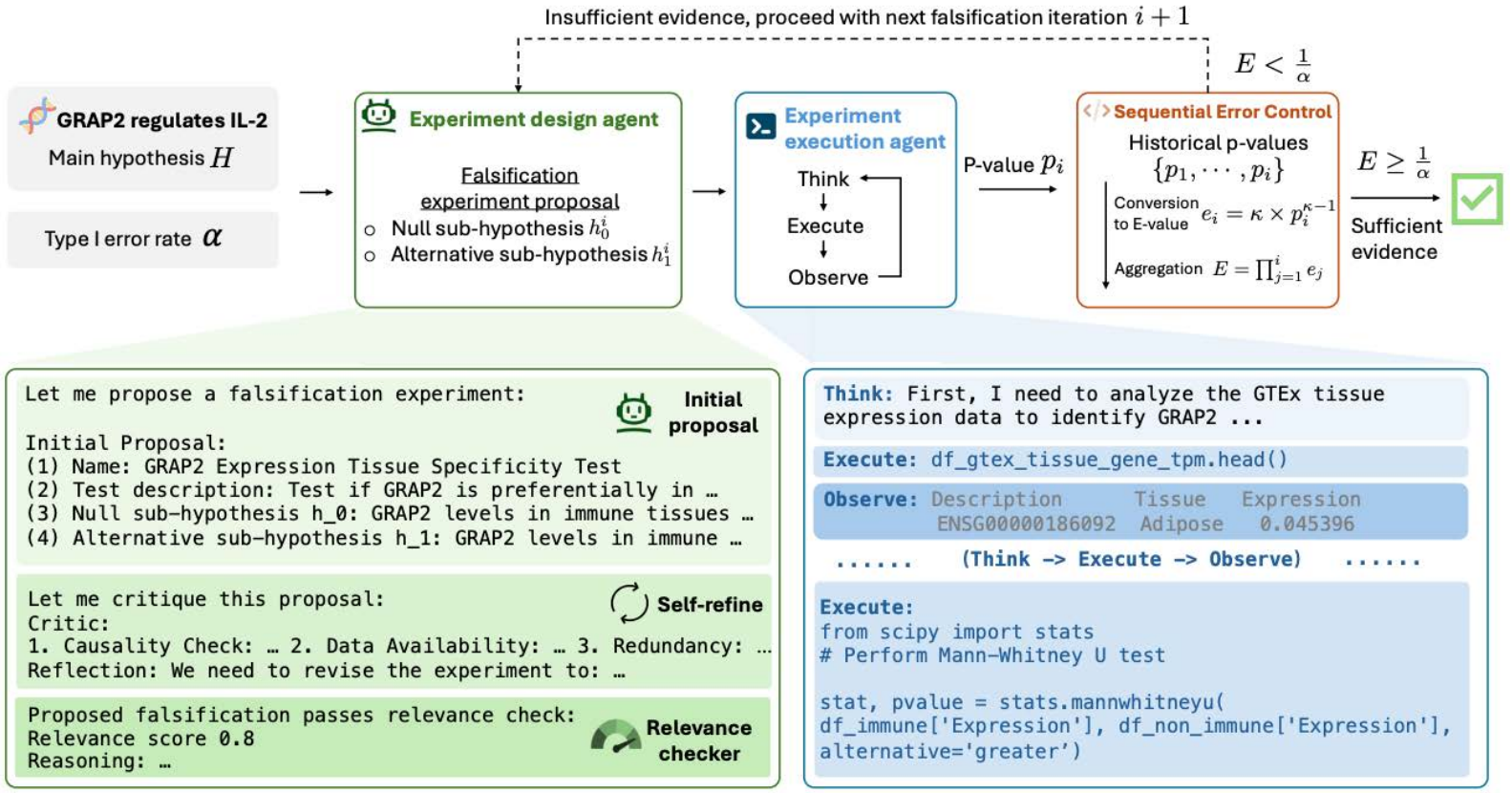
清華大學 Tsinghua University

- 自动化的证伪需要AI系统提出可能假设、显式设计证伪方案、自主执行证伪流程、泛化用于不同研究。
- 目前的相关工作仍未有效实现全自动证伪。

Representative Work	Target Domain	Falsification Process			Scientific Discoveries	
		Explicit	Conductor	Approach	Automated	Generalized
AlphaFold (Abramson et al., 2024)	Protein	✗	Human Researchers	Wet Experiments	✗	✗
Laboratory Mobile Robots (Dai et al., 2024)	Synthetic Chemistry	✓	AI-Empowered Robots	Synthesis & Screening with Physical Equipment	✗	✗
AI Scientist (Lu et al., 2024)	Machine Learning	✗	AI-Empowered Agents	Experimenting Through Codebase Edits	✗	✗
AI Hilbert (Cory-Wright et al., 2024)	Polynomial Data	✗	AI-Empowered Agents	Polynomial Optimization	✗	✗
DataVoyager (Majumder et al., 2024a)	Data Analysis	✓	AI-Empowered Agents	Tool Learning with LLMs	✓	✗
Ideal AIGS Systems	***General***	✓	***AI-Empowered Systems***	***Experiments Designed & Conducted Autonomously***	✓	✓

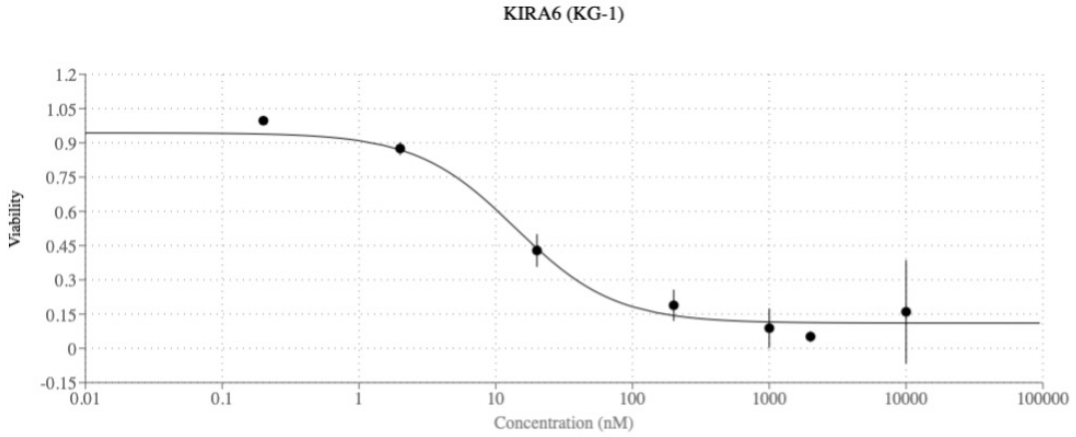
数据规律探索中的自动化证伪

- **Huang et al., Automated Hypothesis Validation with Agentic Sequential Falsifications. 2025.**
- **在给定数据集上，基于统计显著性自动化证伪给定的可能假设。**

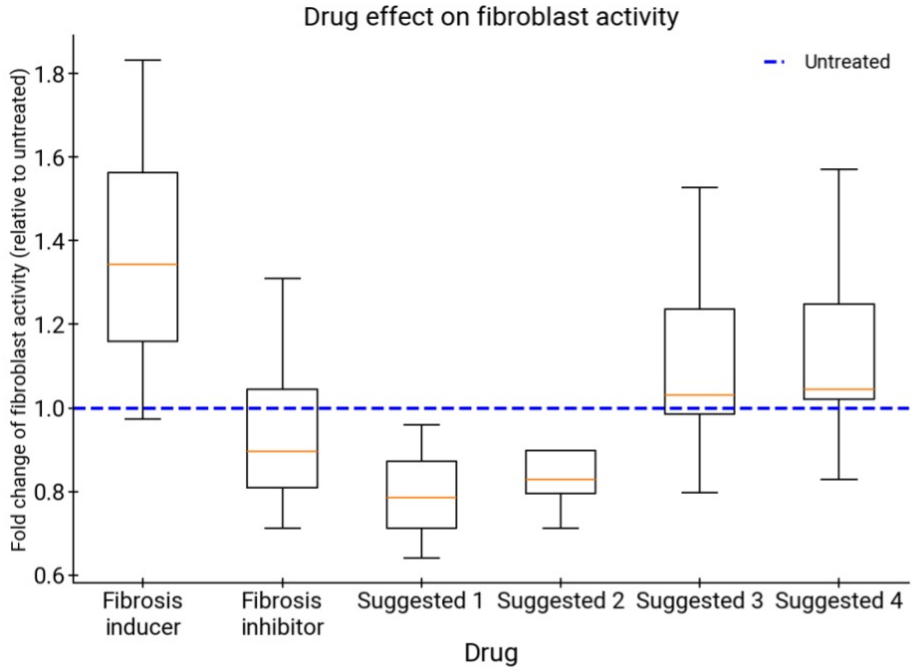


自动化研究智能体在其他领域的实践

- Google Research. AI Co-Scientist. 2025.
- 利用测试时计算扩展，生成高质量研究计划与实验设计，辅助研究。



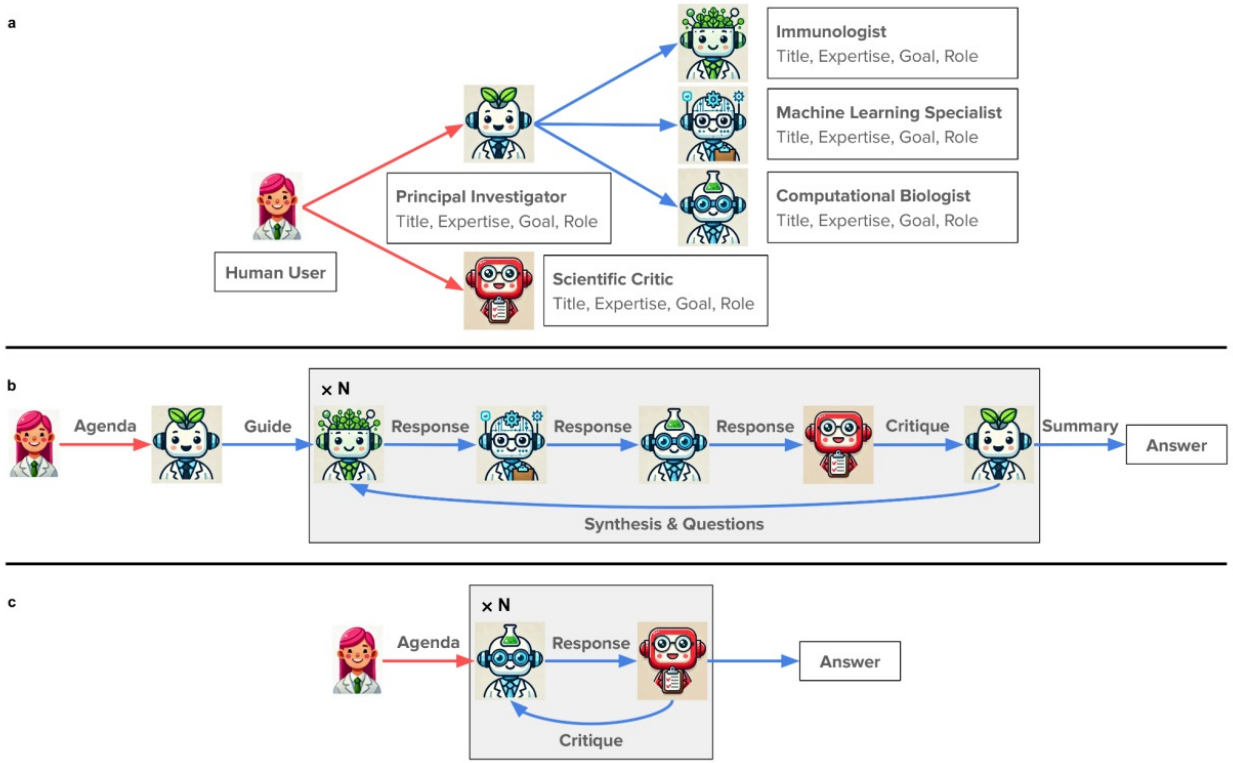
抗癌药物设计合成

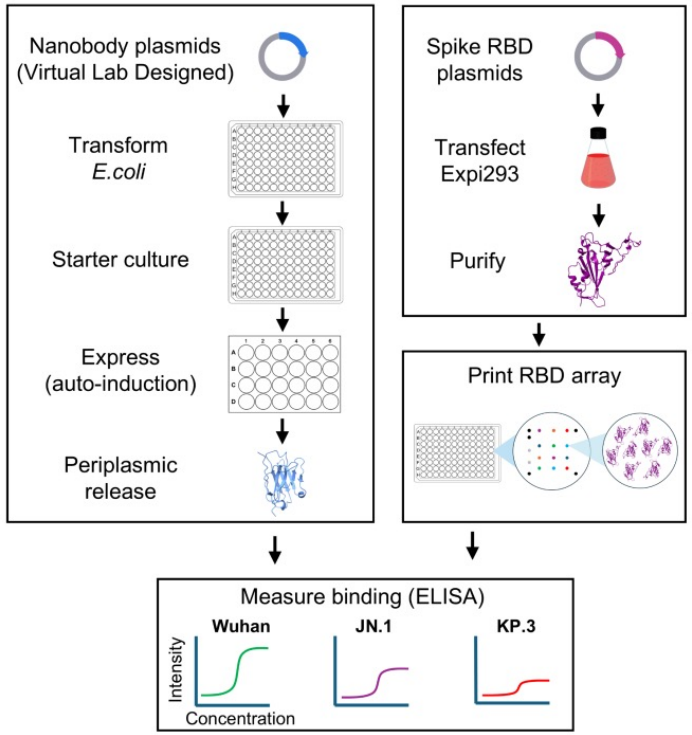


肝纤维化药物治疗

自动化研究智能体在其他领域的实践

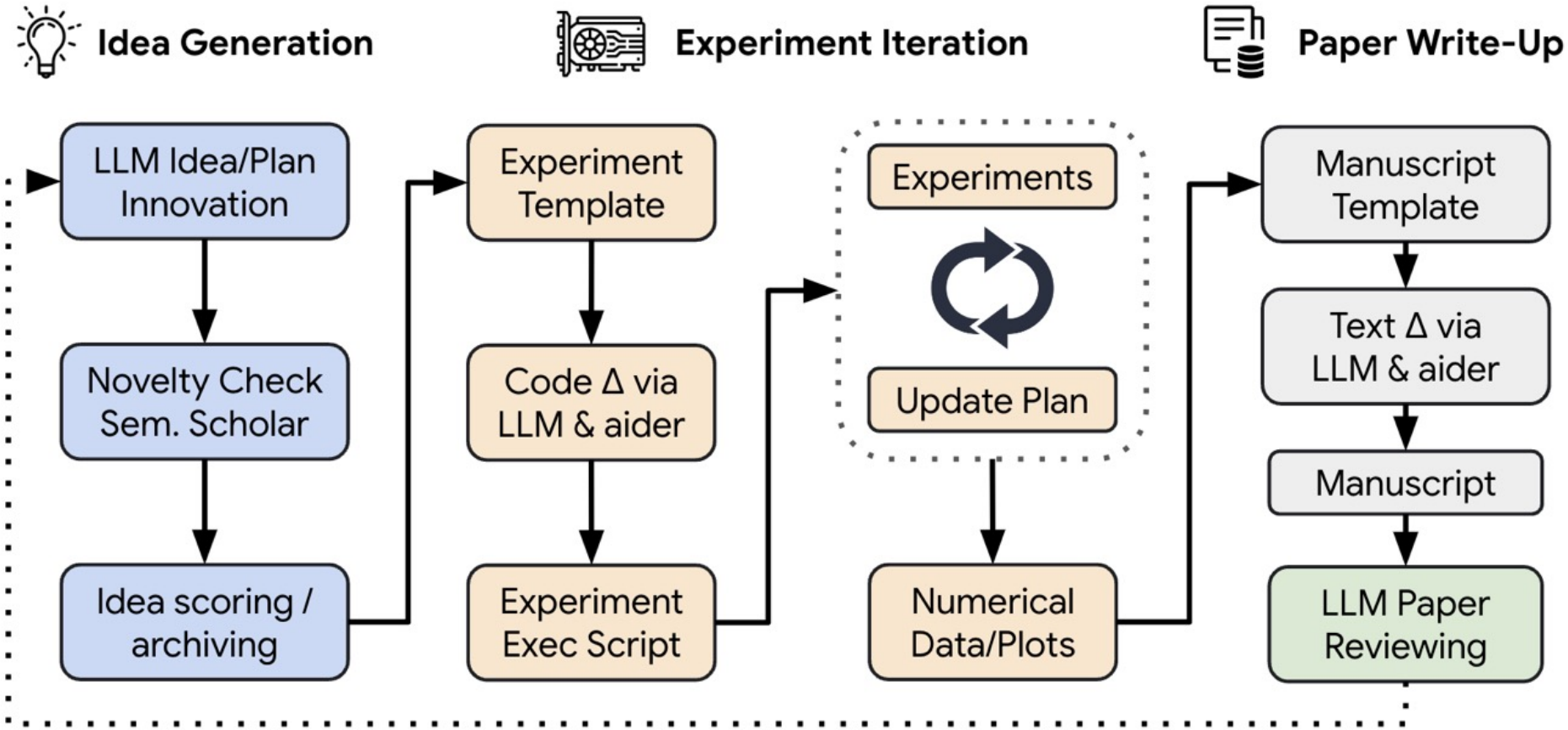
- Swanson et al., The Virtual Lab: AI Agents Design New SARS-CoV-2 Nanobodies with Experimental Validation. 2024.
- 多智能体系统调用工具，设计新冠病毒抗体，经湿实验验证部分有效。





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：全自主AI科学发现探索

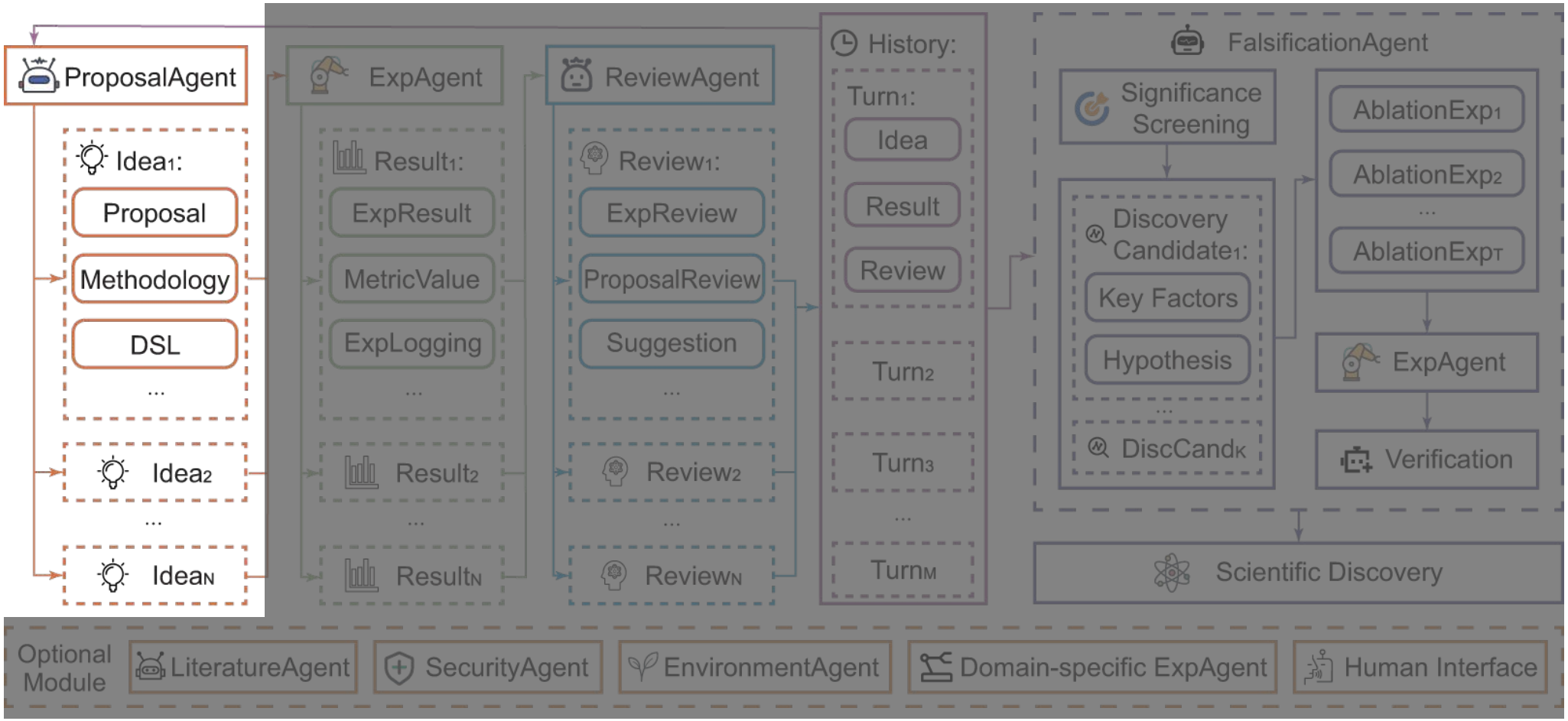
AIGS及核心原则

- **AIGS（AI-Generated Science）：AI在人类不输入创造性劳动条件下全自主产生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

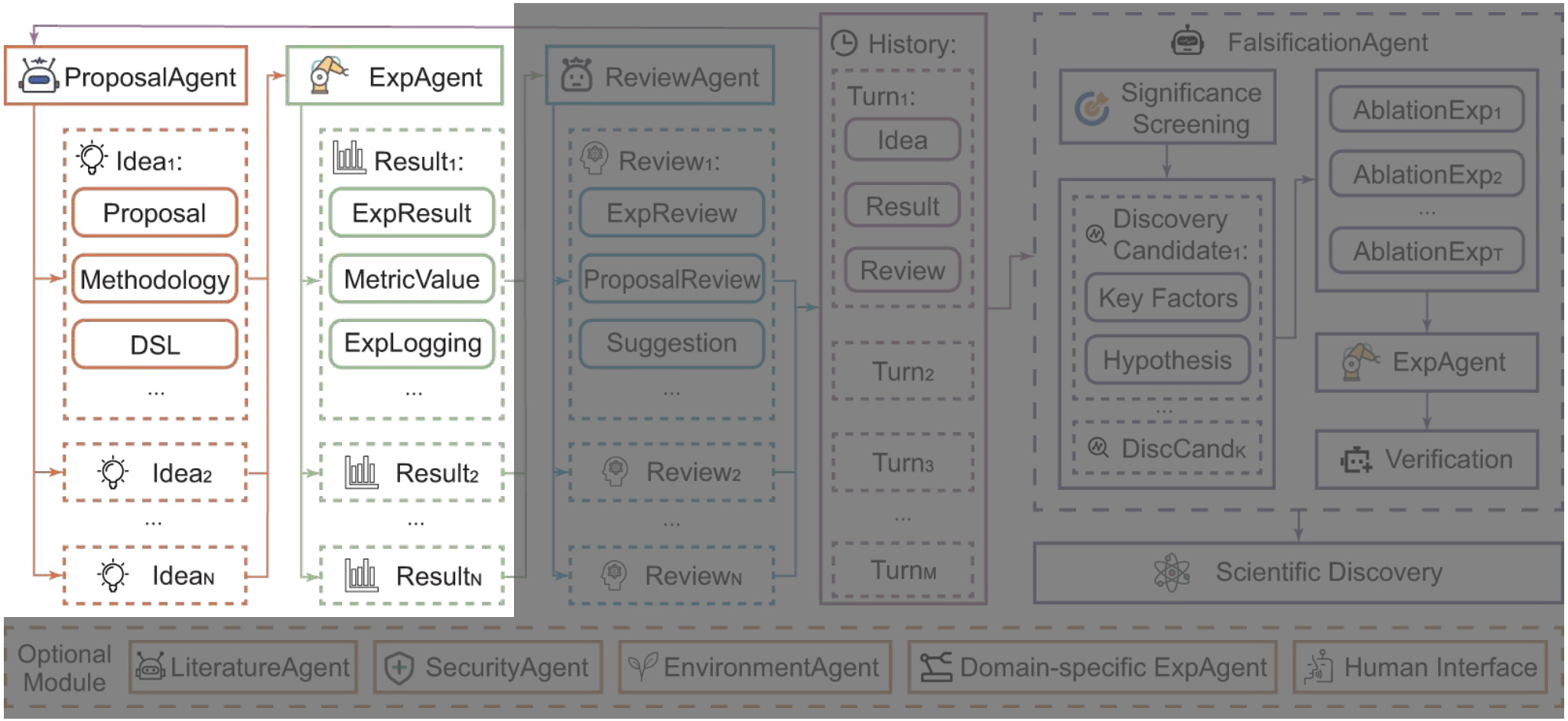
- **ProposalAgent：提出具有创造力的科研想法。**



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

Baby-AIGS

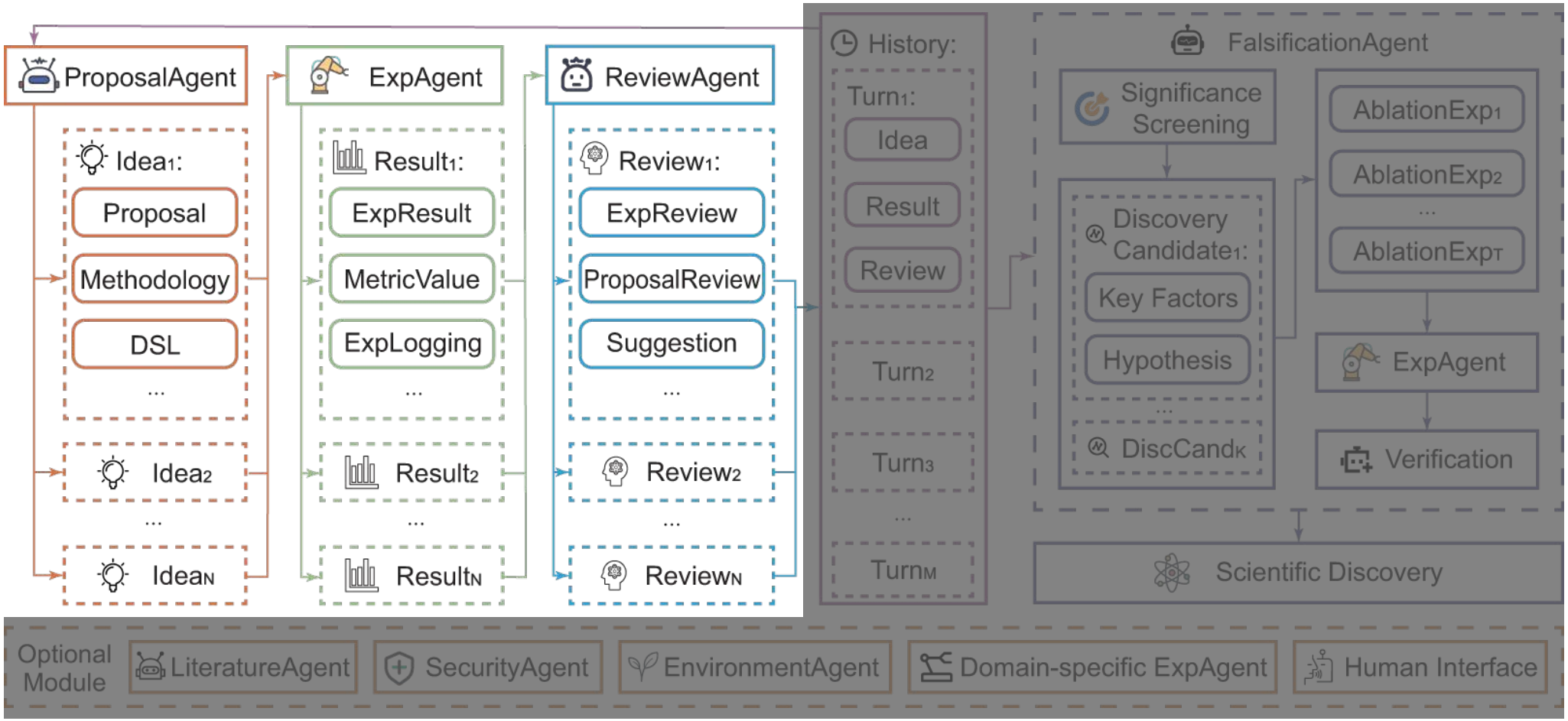
- **ExpAgent：实现想法并且进行科学实验。**



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

Baby-AIGS

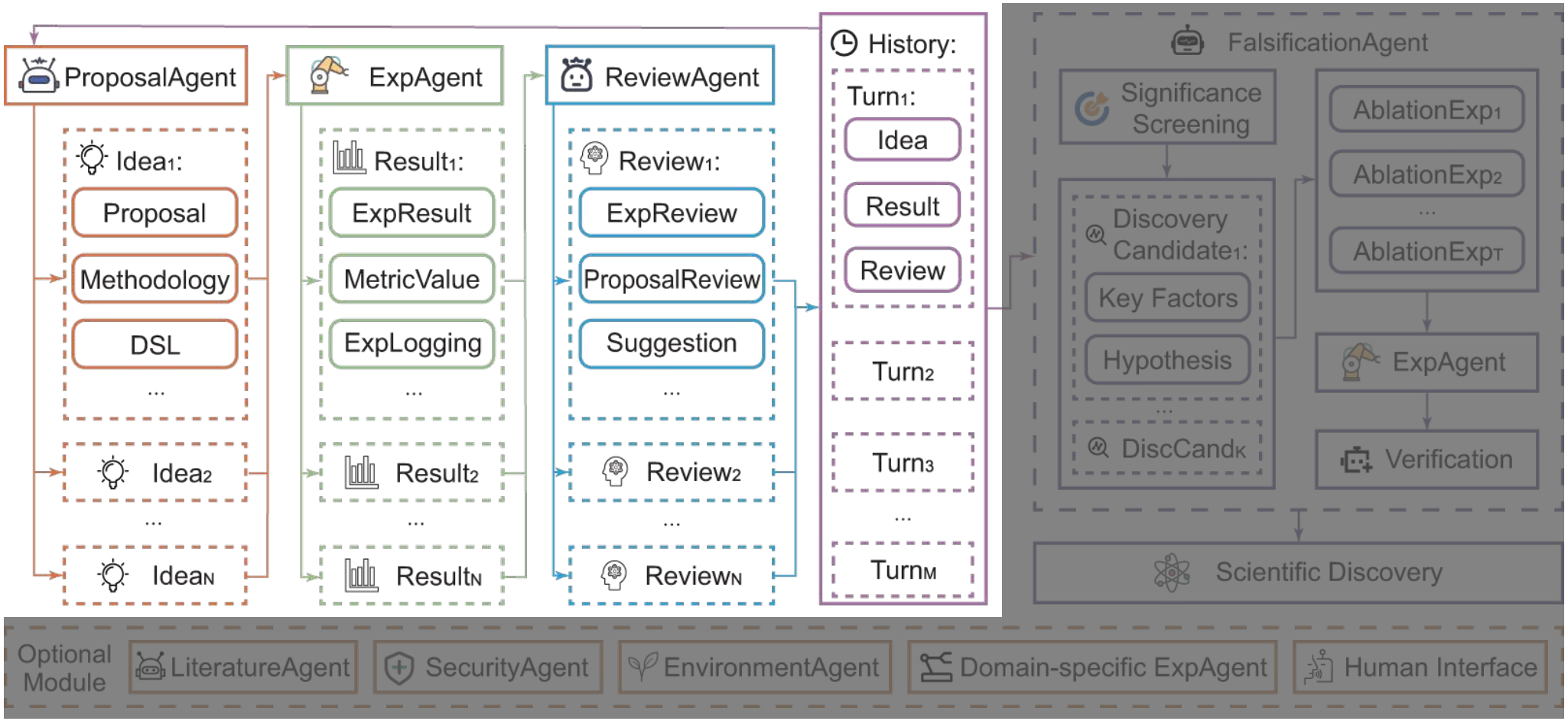
- **ReviewAgent：根据实验结果对想法提出建设性反馈。**



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

Baby-AIGS

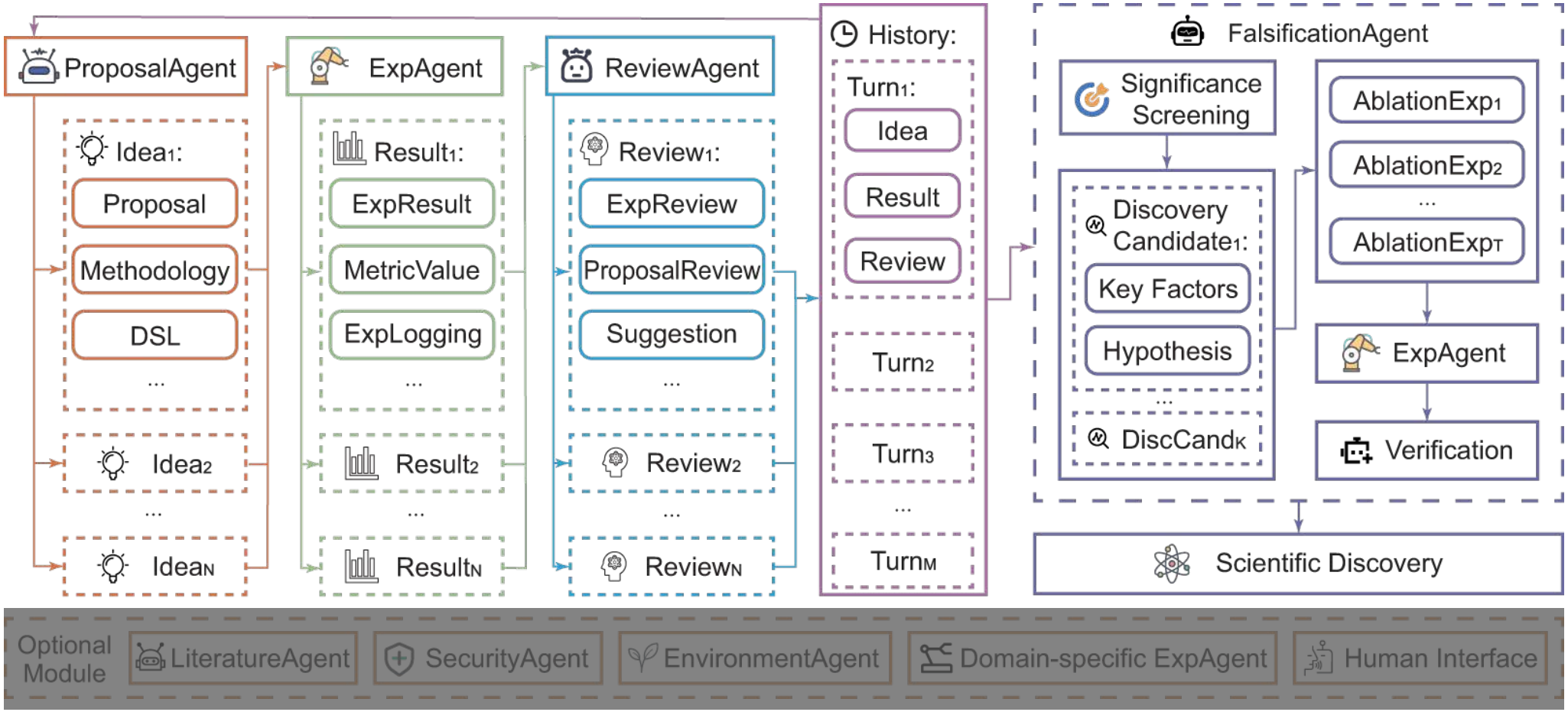
- **Iteration：循环迭代以提高实验表现。**



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

Baby-AIGS

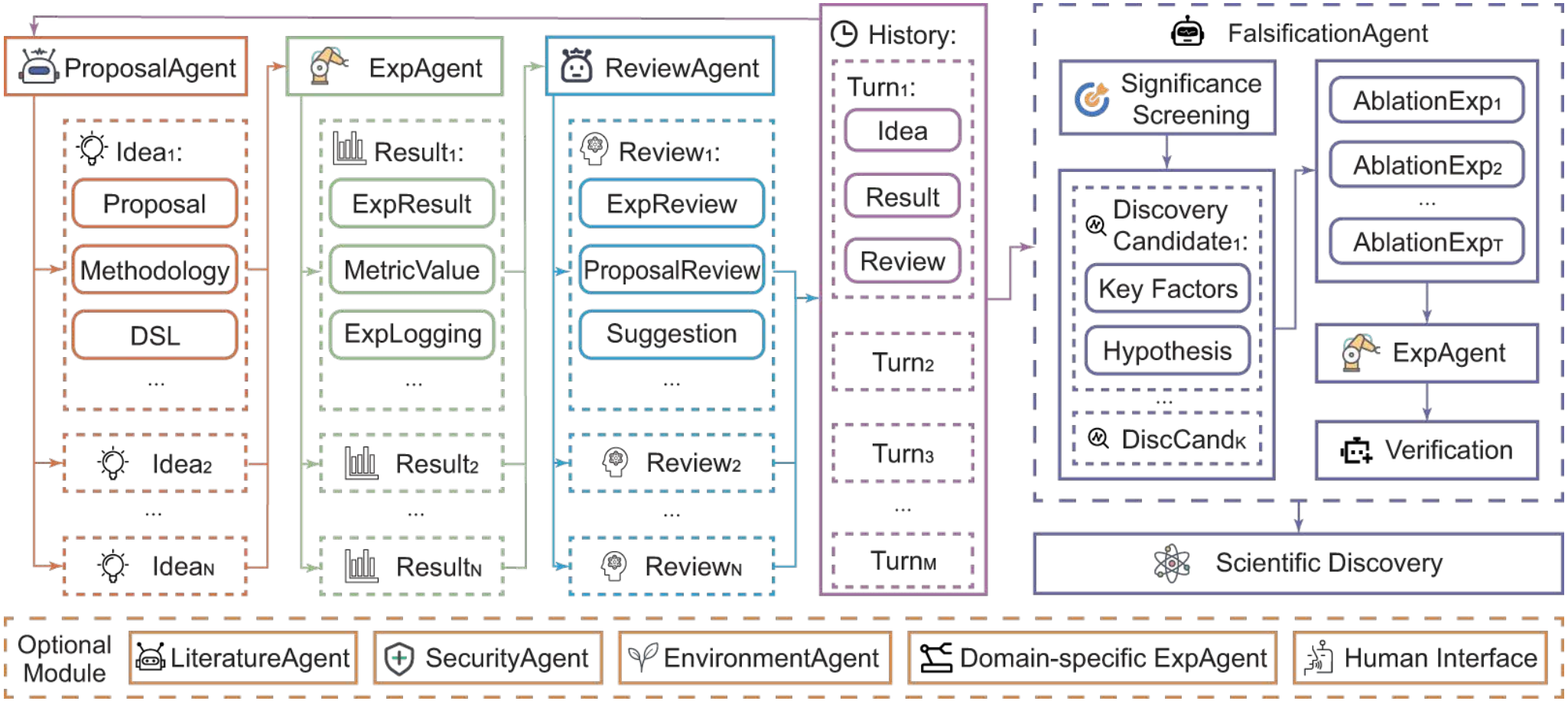
- **FalsificationAgent：消融实验来进一步验证科学发现。**



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

Baby-AIGS

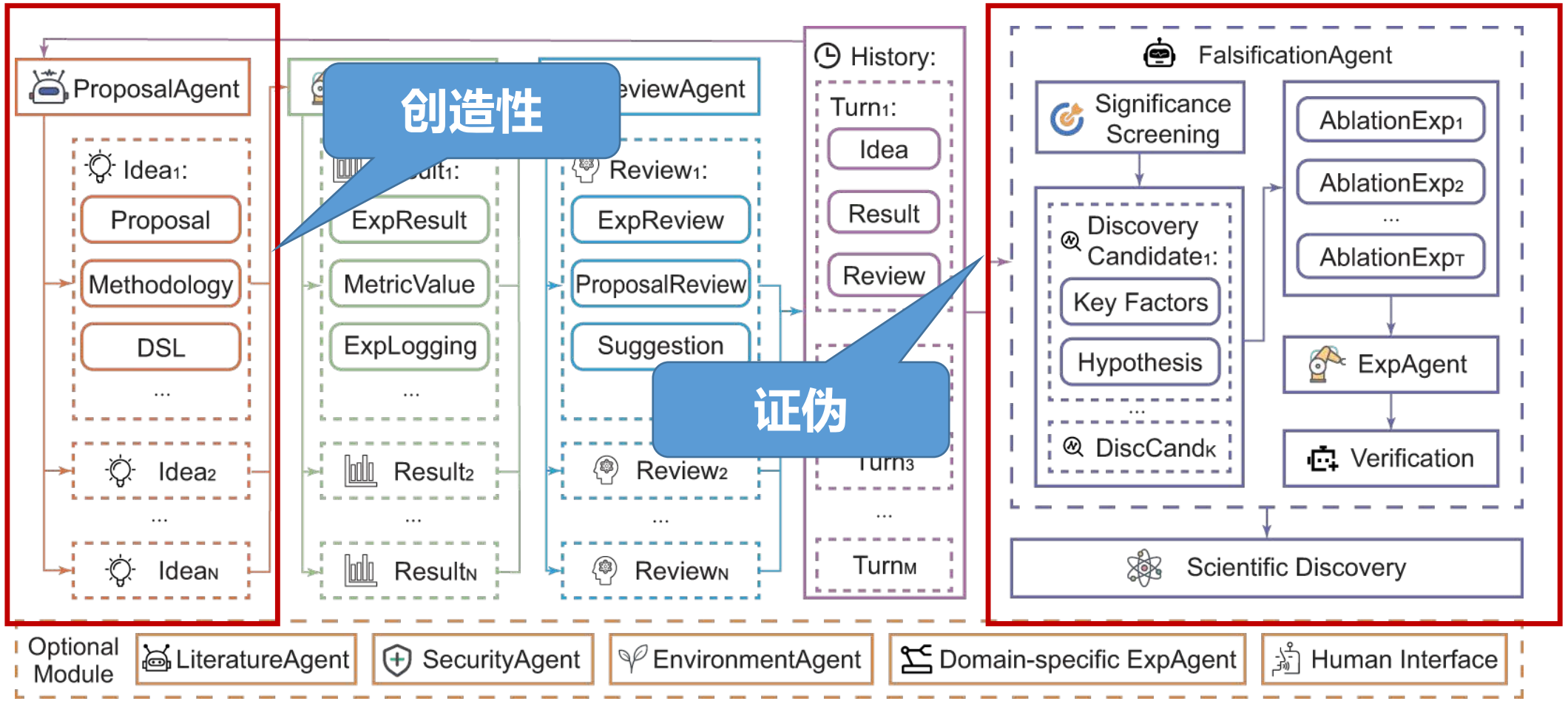
- **Optional Module：可集成其他模块实现补充功能。**



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

核心原则：证伪和创造性

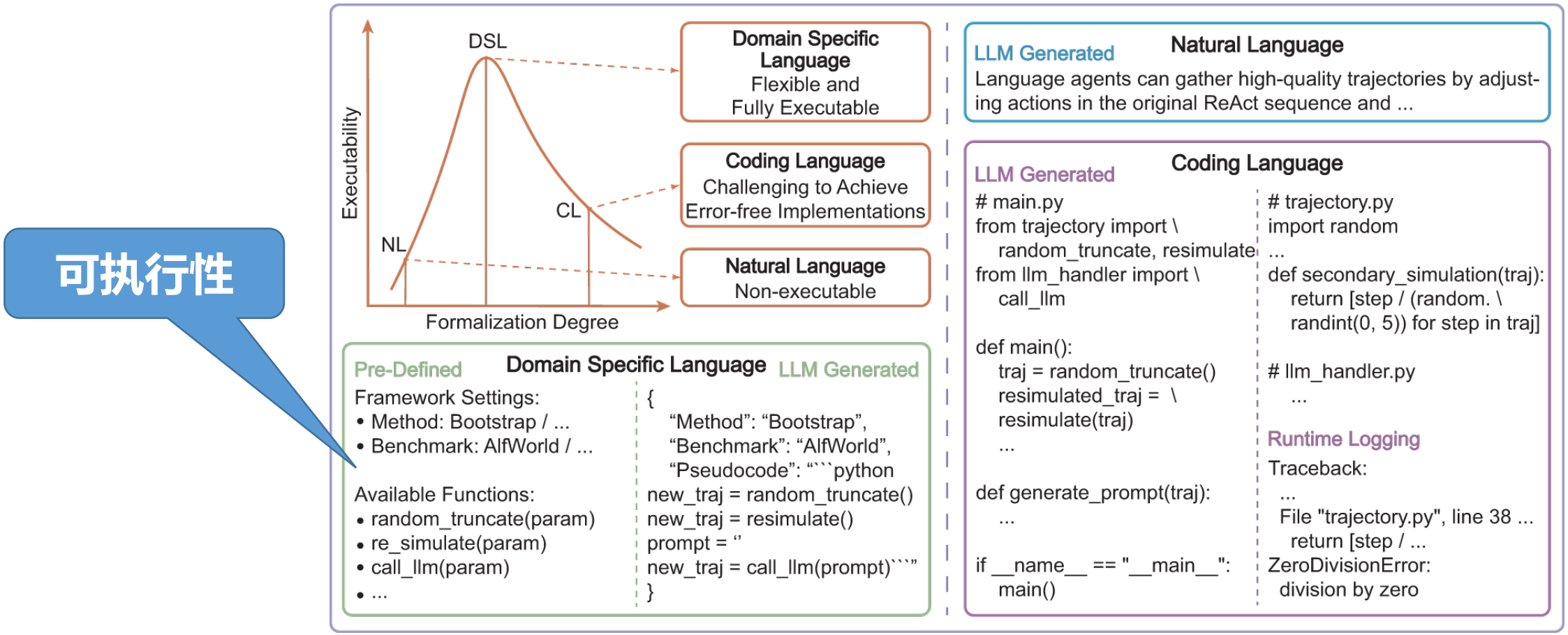
- 证伪和创造性主要由FalsificationAgent和ProposalAgent实现。



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

核心原则：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。**



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

- **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 ∼ 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 ∼ 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 ∼ 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 ∼ 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。**

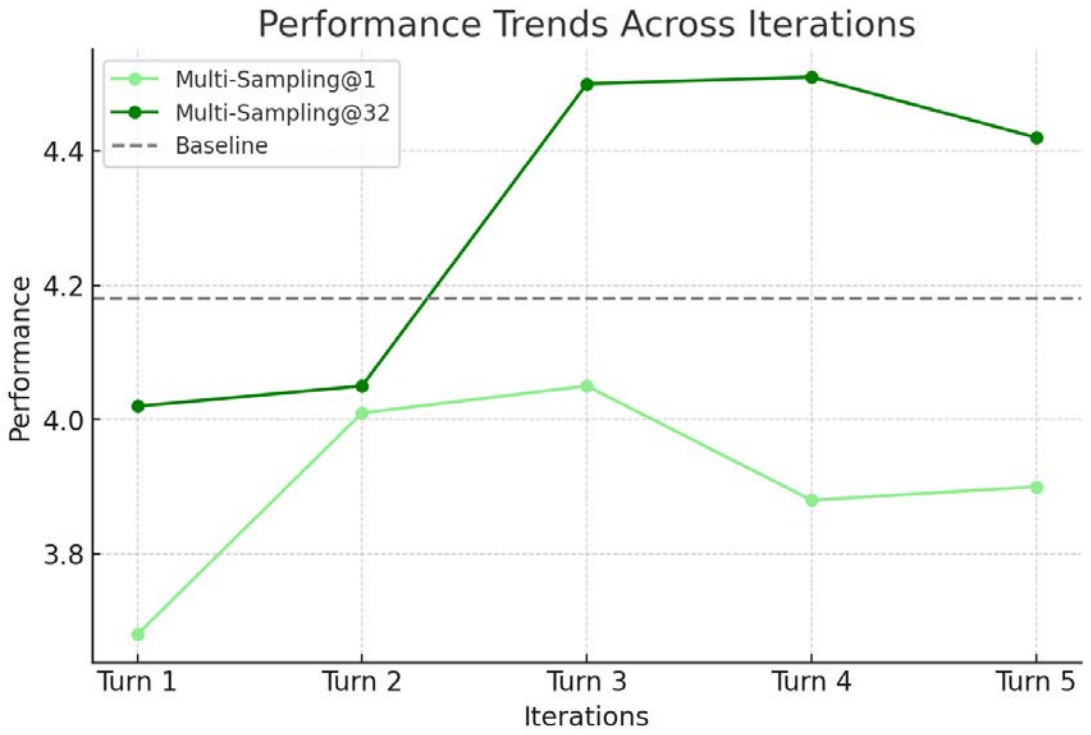
Method	Experiment Success Rate (Exp. SR)	Overall Success Rate (Overall SR)
AI Scientist	44.8%	29.2%
Baby-AIGS (Ours)	**Almost 100%**	**Almost 100%**

- **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）。**

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

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。

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

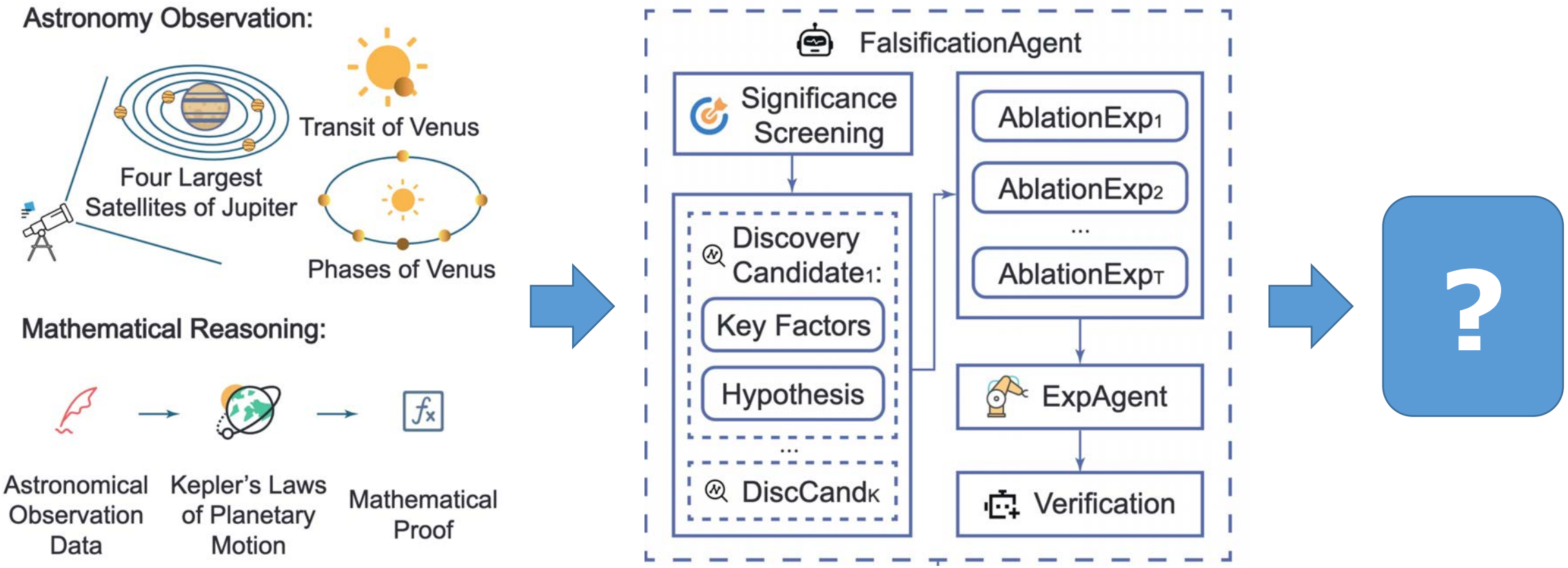
杨宗瀚

张真赫

展望

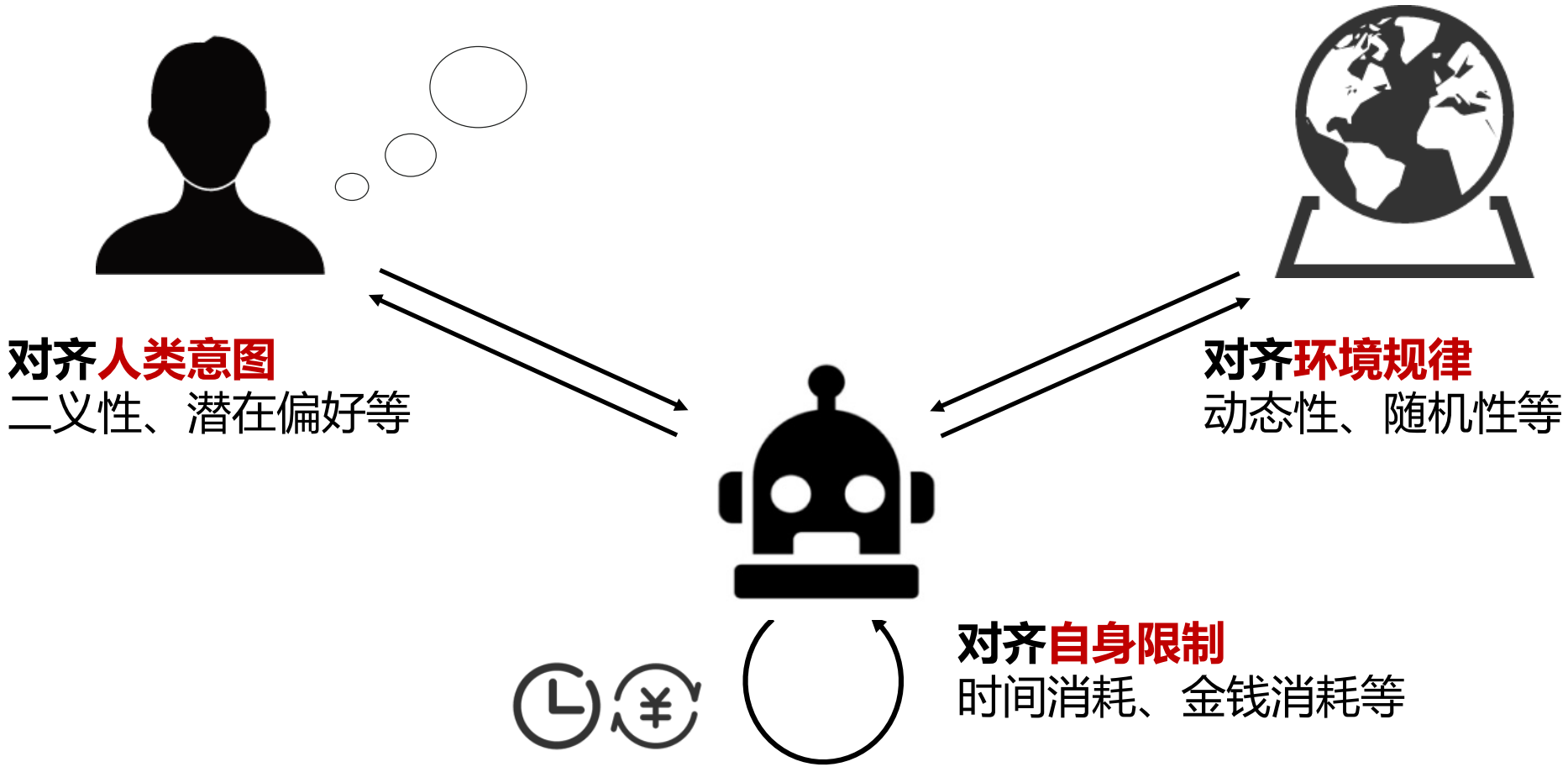
展望：需要对AI自主科学本身进行证伪

- 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

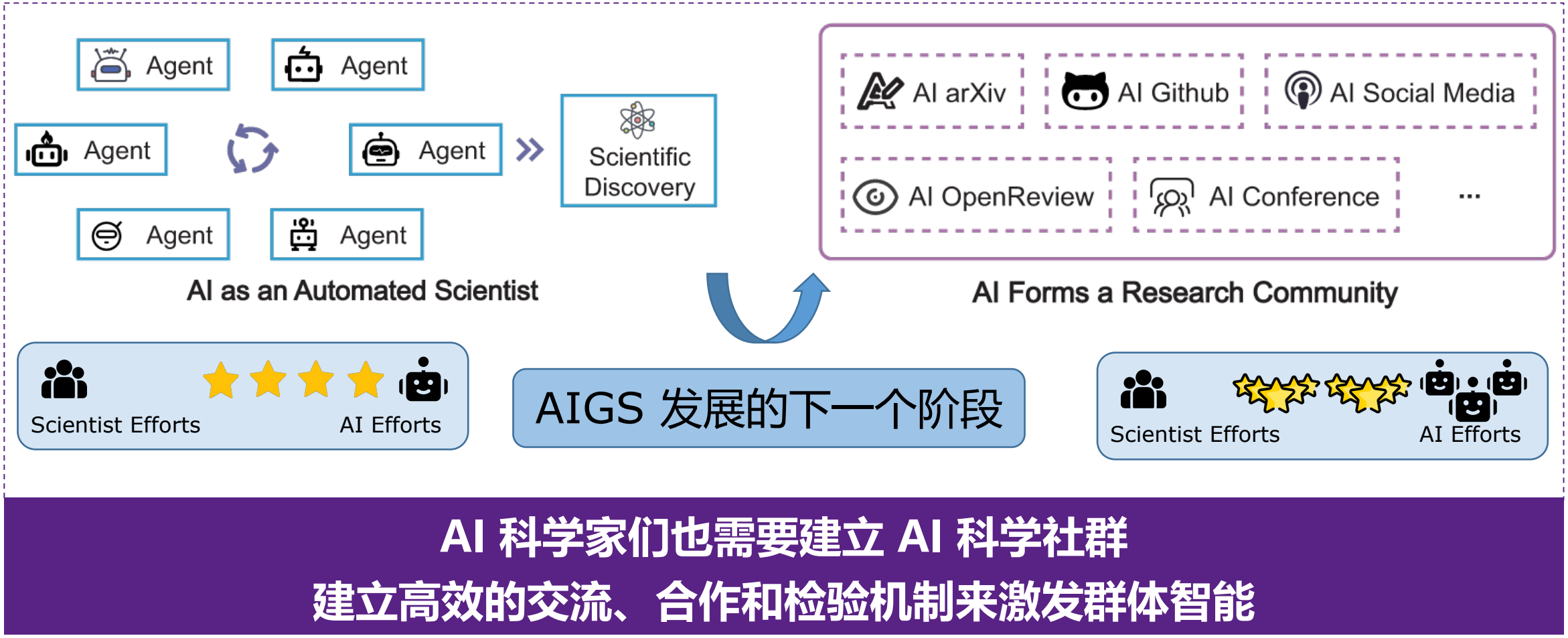
- AIGS 智能体需要和自身、人类、环境统一对齐。



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。



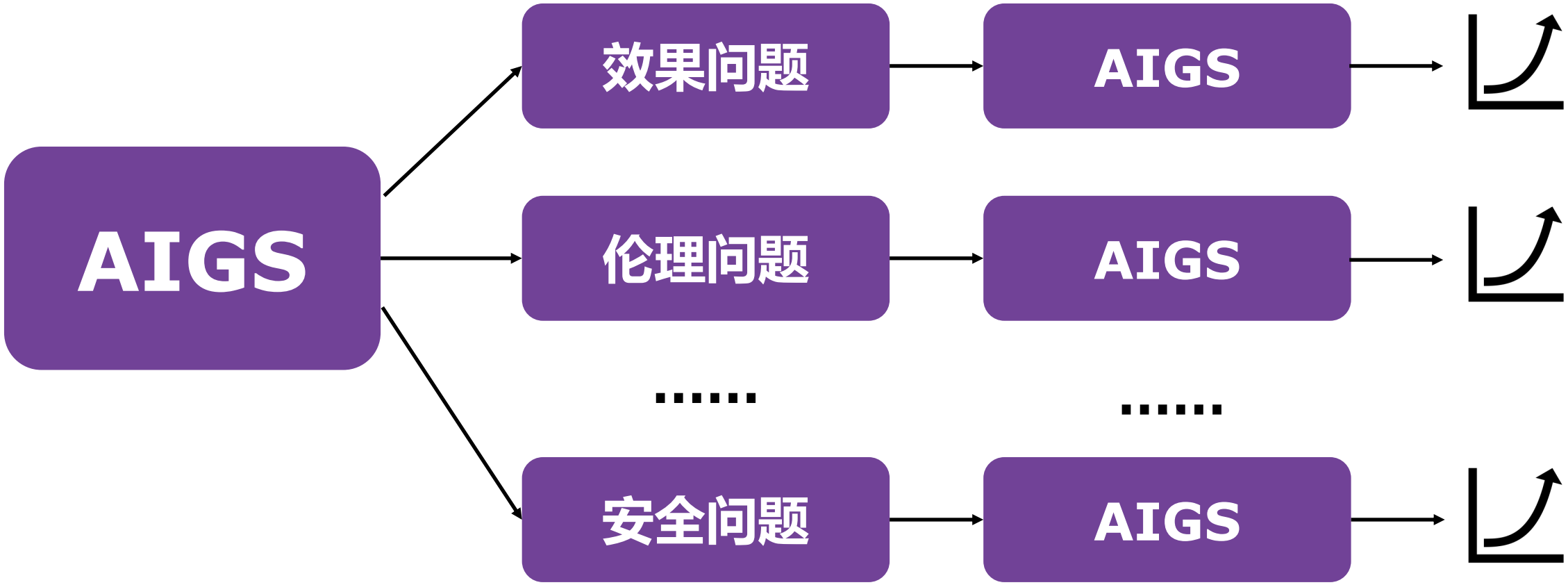
AI 科学家们也需要建立 AI 科学社群

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.

谢谢