

# Topic 2. Idea Generation and Experimentation



**Wei Zhao**

Assistant Professor  
University of Aberdeen  
[wei.zhao@abdn.ac.uk](mailto:wei.zhao@abdn.ac.uk)



prinz ✓

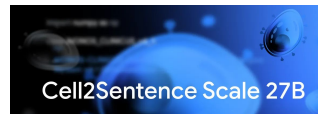
@deredleritt3r



Just to recap:

We found out today that an LLM that fits on a high-end consumer GPU, when trained on specific biological data, can discover a novel method to make cancer tumors more responsive to immunotherapy.

Confirmed novel discovery (not present in existing literature).  
Experimentally validated in living cells.



**Intimate colleague: 7 x 24 hours; who can inspire, discuss, expand, implement, test, refine your ideas**

Google and Yale scientists have trained an LLM that has generated a novel hypothesis about cancer cellular behavior. This prediction was confirmed multiple times in vitro.

- "What made this prediction so exciting was that it was a novel idea. Although...  
[Show more](#)



<https://blog.google/technology/ai/google-gemma-ai-cancer-therapy-discovery/>

9:40 PM · Oct 15, 2025 · 577.7K Views

# Agenda

- ❑ Idea Generation
- ❑ Evaluation
- ❑ Automated Experimentation
- ❑ Some Limitations
- ❑ Topics Not Covered



# Agenda

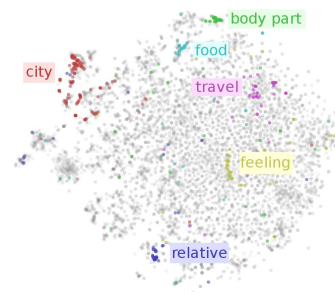
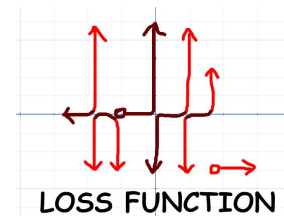
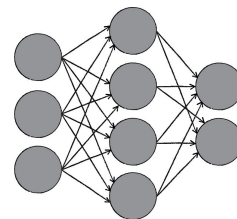
- ❑ Idea Generation
- ❑ Evaluation
- ❑ Automated Experimentation
- ❑ Some Limitations
- ❑ Topics Not Covered



# Agenda

- ❑ Idea Generation
- ❑ Evaluation
- ❑ Automated Experimentation
- ❑ Some Limitations
- ❑ Topics Not Covered





kaggle

Google Dataset Search Beta

VisualData

UC Irvine Machine Learning Repository



OpenML

DATA HUB

DATA

HealthData.gov

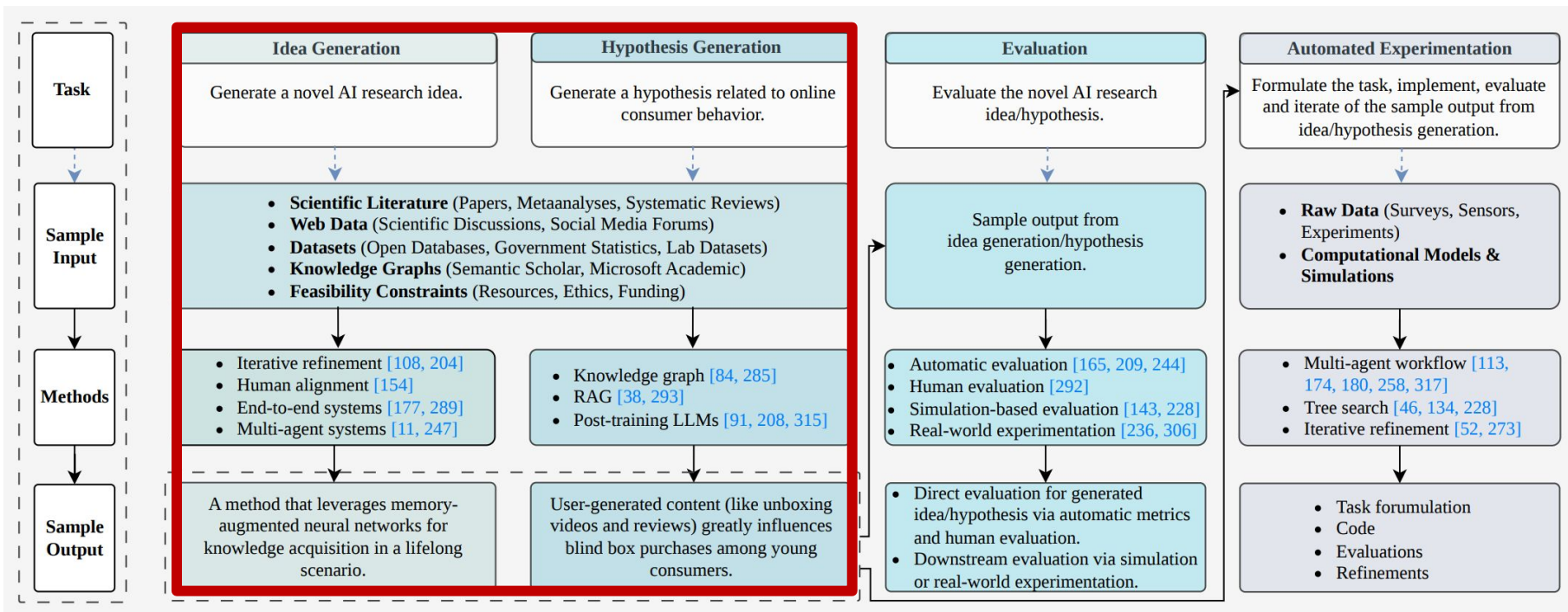
The NLP Index



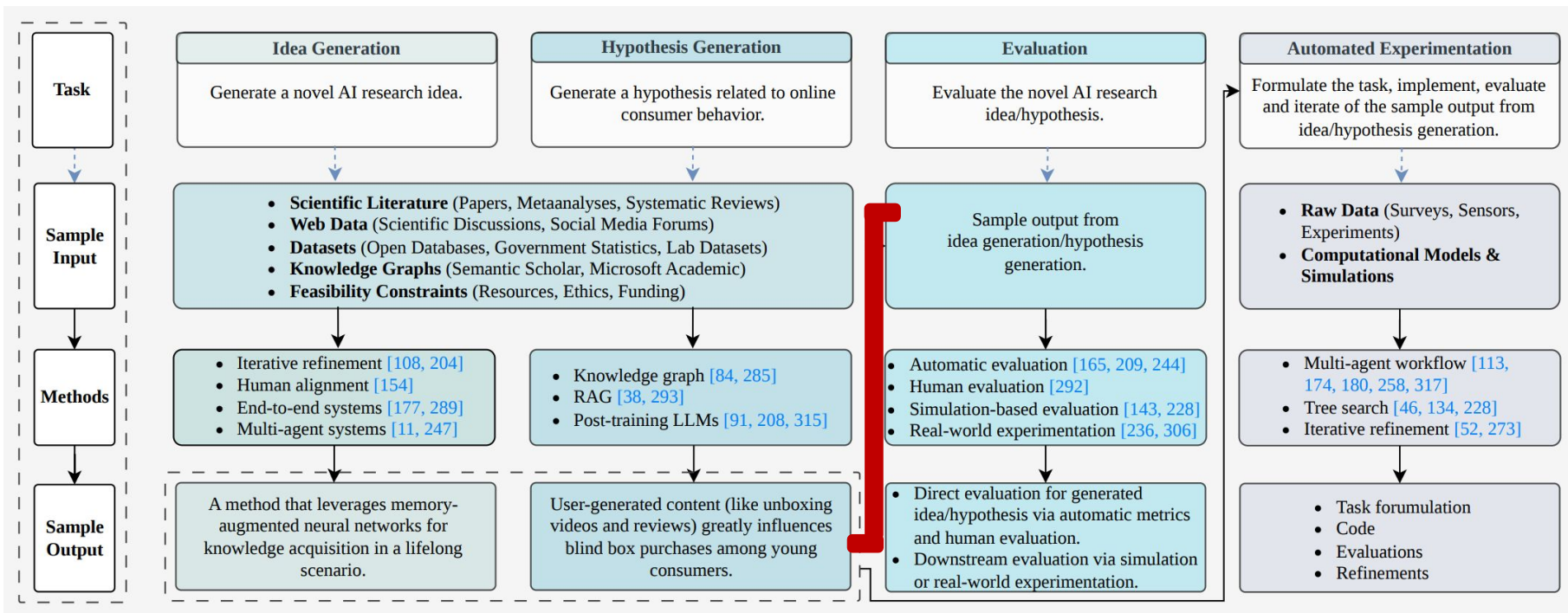
ChatGPT

Gemini

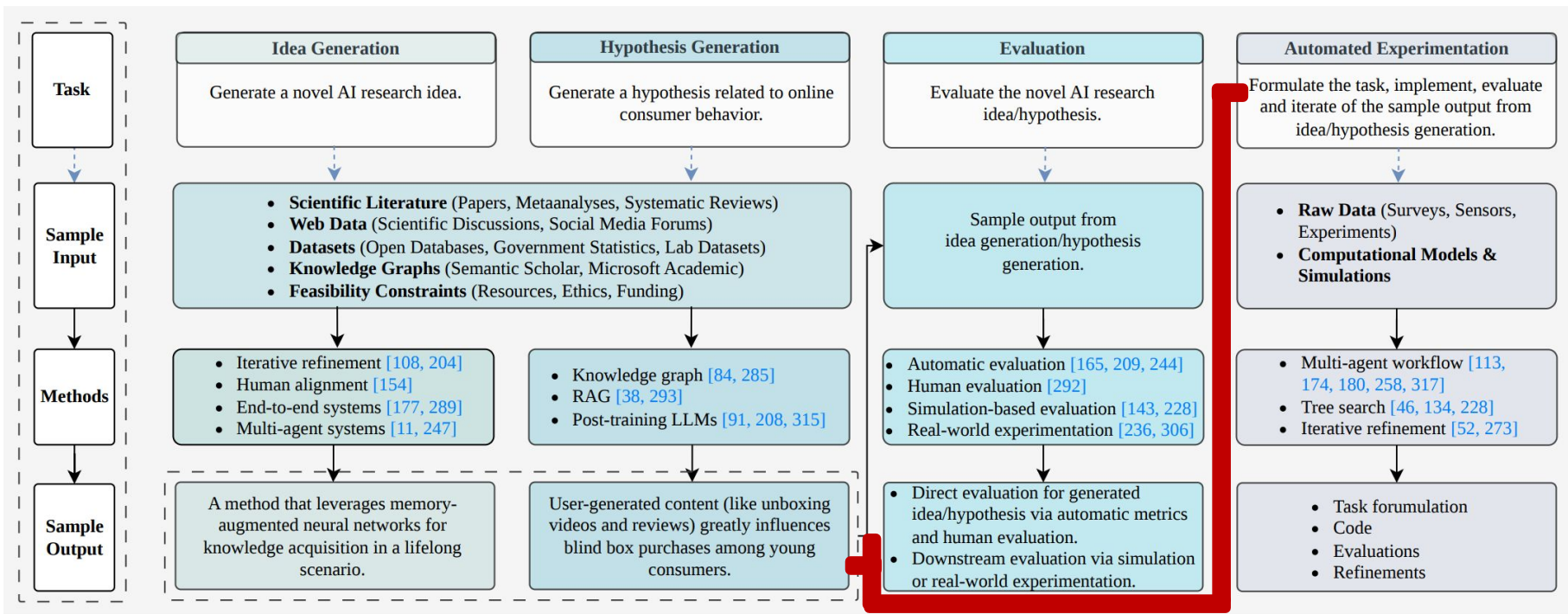
# A bird's-eye view



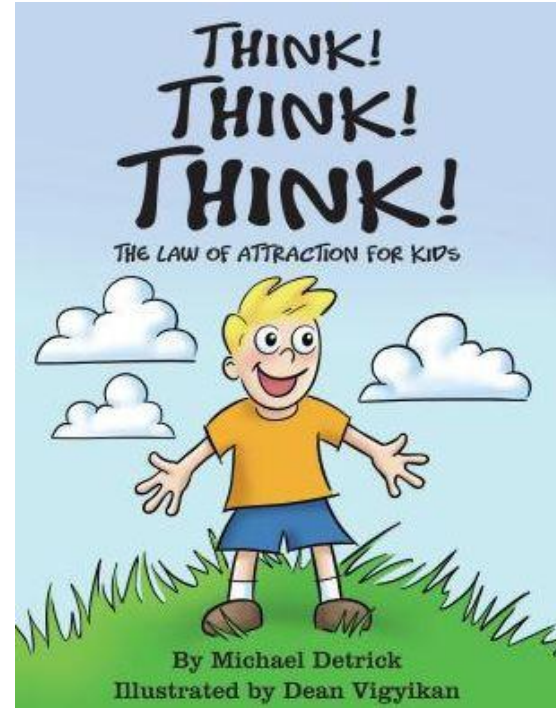
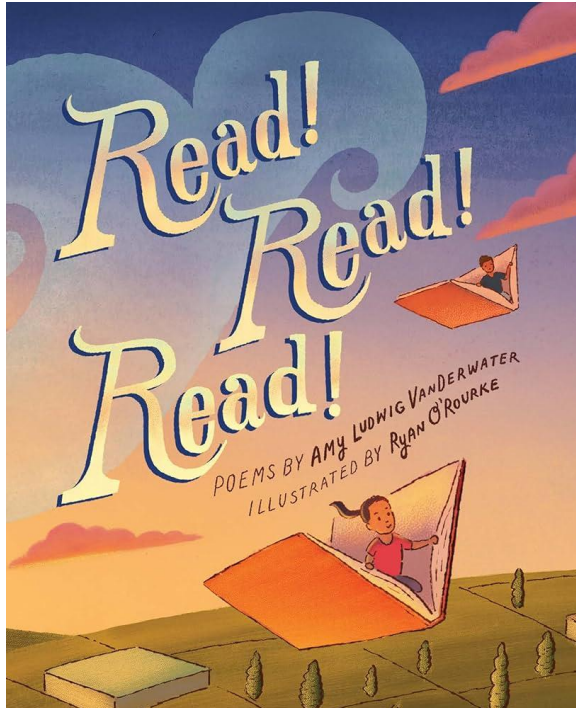
# A bird's-eye view

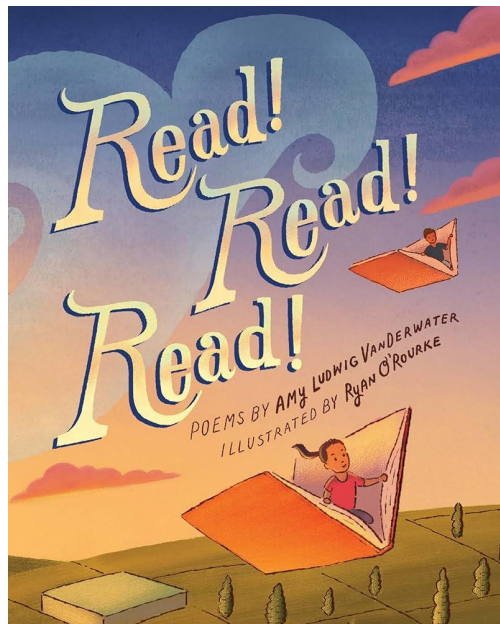


# A bird's-eye view



# Idea Generation





kaggle

Google  
Dataset Search Beta

VisualData

UC Irvine  
Machine Learning  
Repository



OpenML

DATA  
HUB

USA DATA

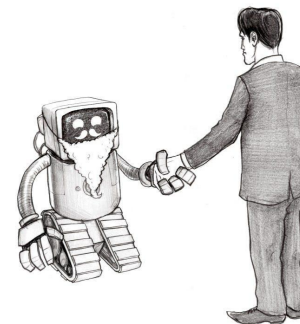
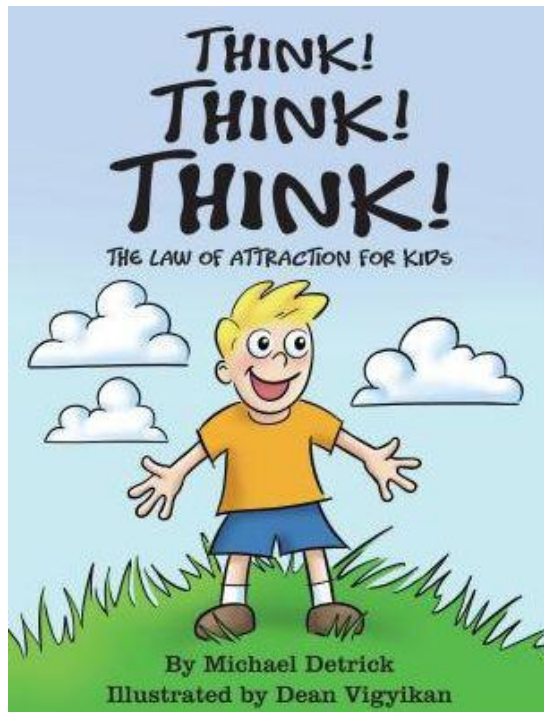
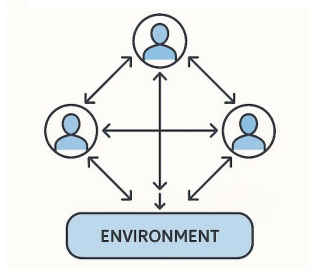
HealthData.gov

The NLP Index



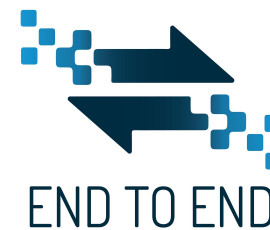
Iterative thinking  
(Hu et al. 2024)

Multi-agent thinking  
(Su et al. 2024)



Human-like thinking  
(Li et al. 2024)

End-to-end thinking  
(Lu et al. 2024)



# Idea Generation

**THE AI SCIENTIST-v2: Workshop-Level  
Automated Scientific Discovery via Agentic  
Tree Search**

**AI Can Learn Scientific Taste**

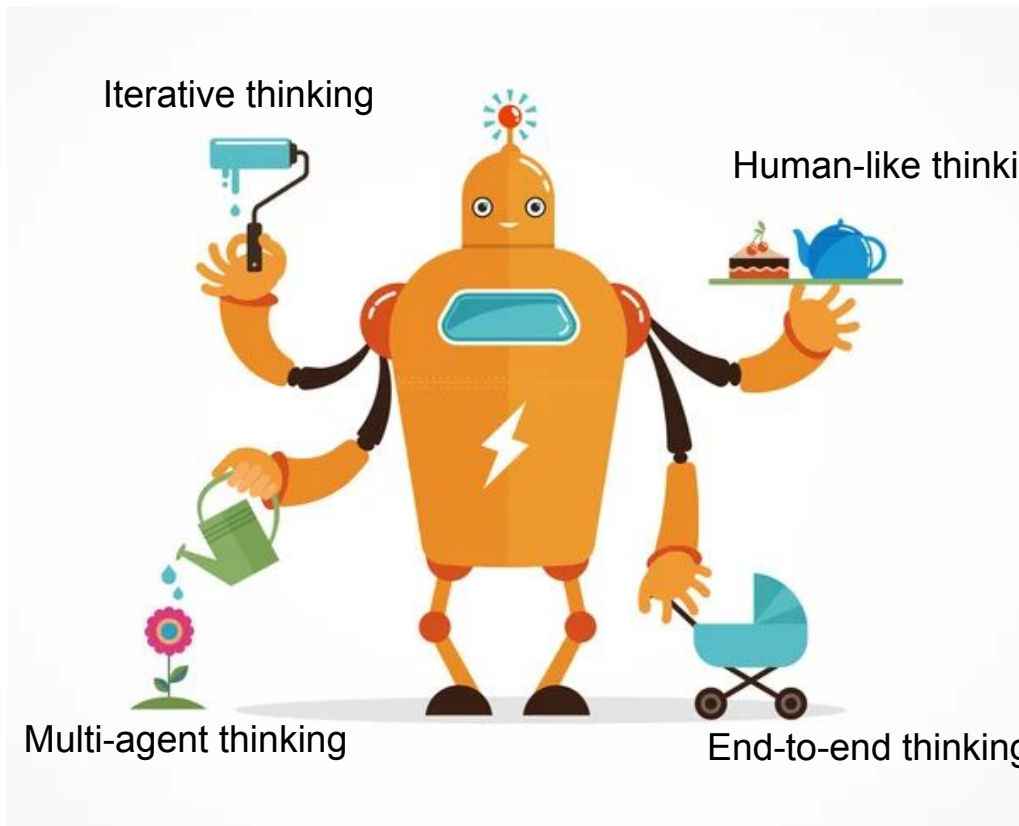
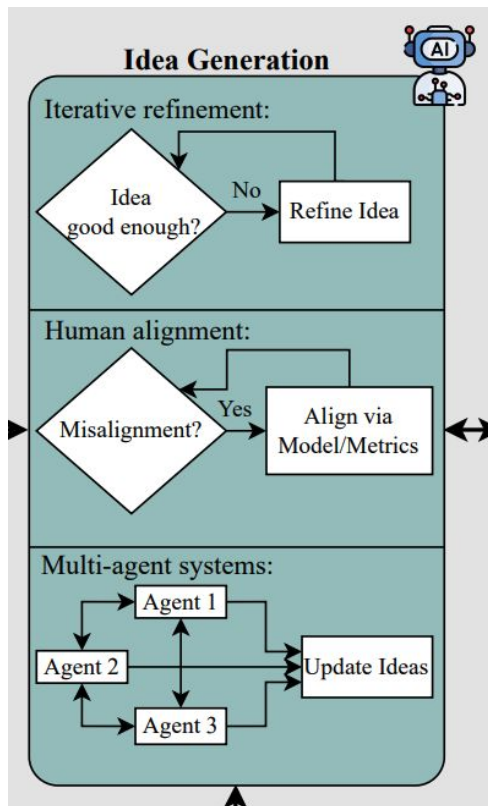
**Agent Laboratory: Using LLM Agents as  
Research Assistants**

**Can LLMs Generate Novel Research Ideas?**

A Large-Scale Human Study with 100+ NLP Researchers

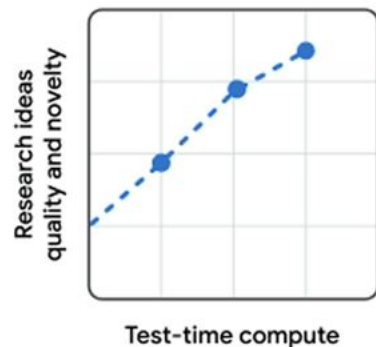
**Towards an AI co-scientist**

# Idea Generation

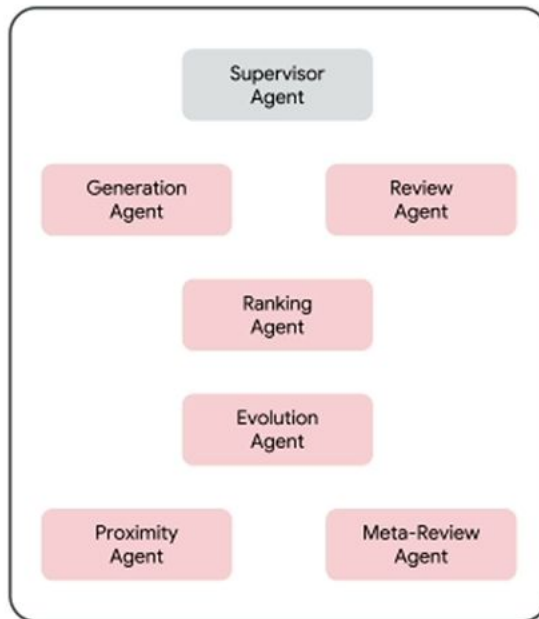


# Google AI Co-Scientist

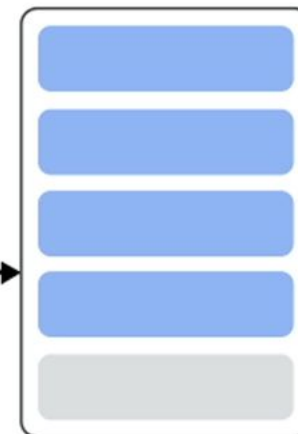
- ❑ Multi-agent thinking
- ❑ Iterative thinking



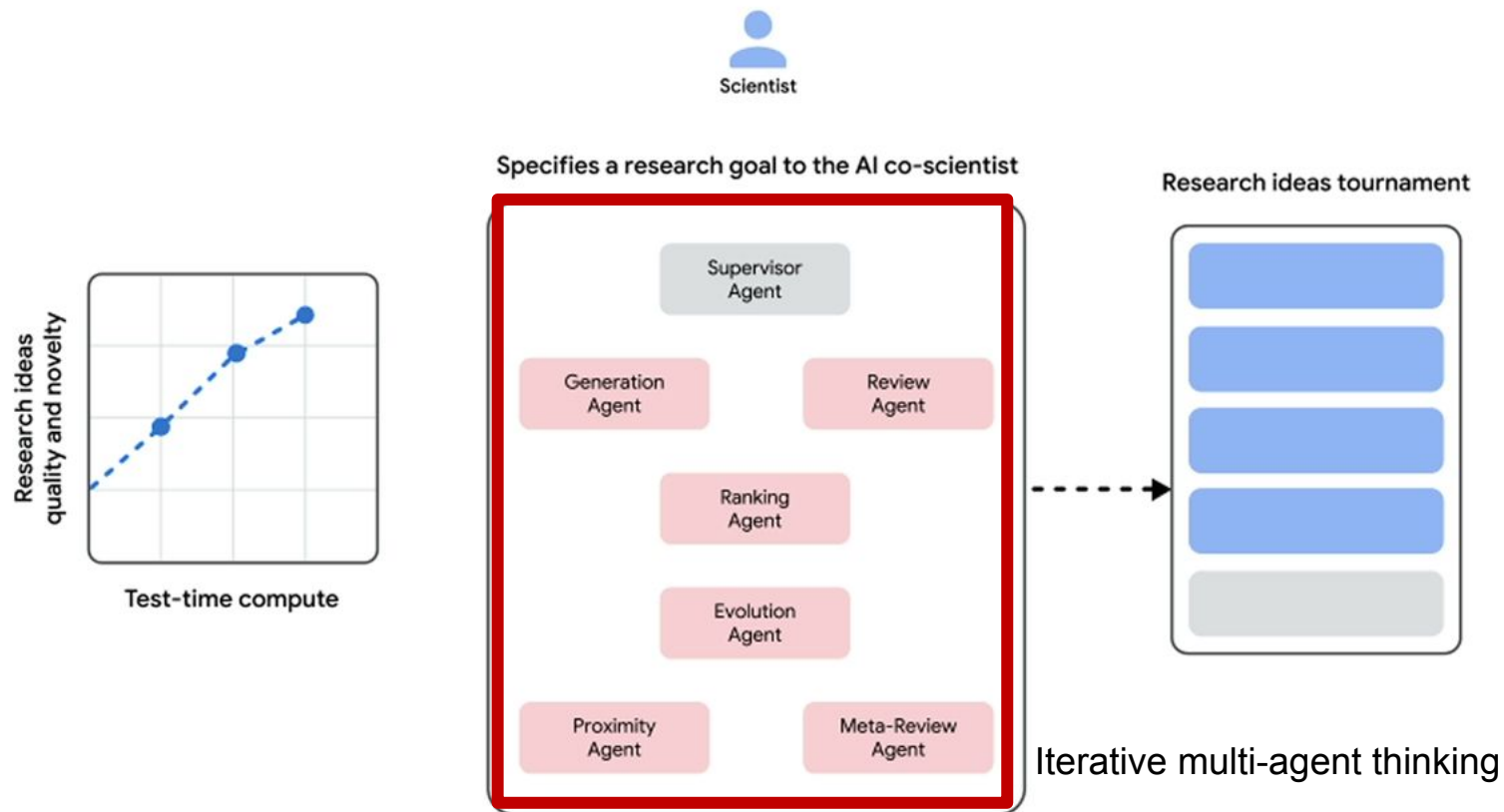
Specifies a research goal to the AI co-scientist



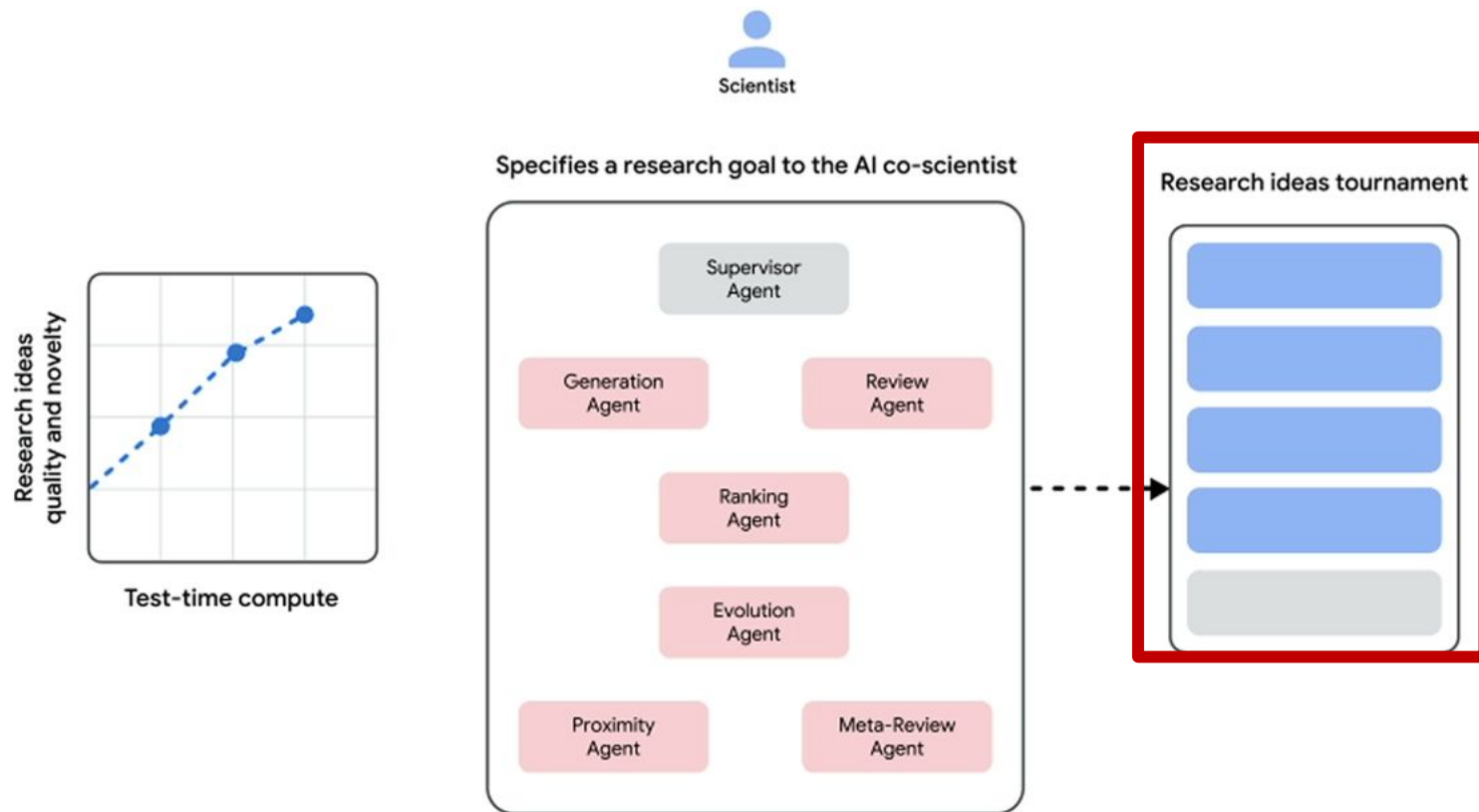
Research ideas tournament



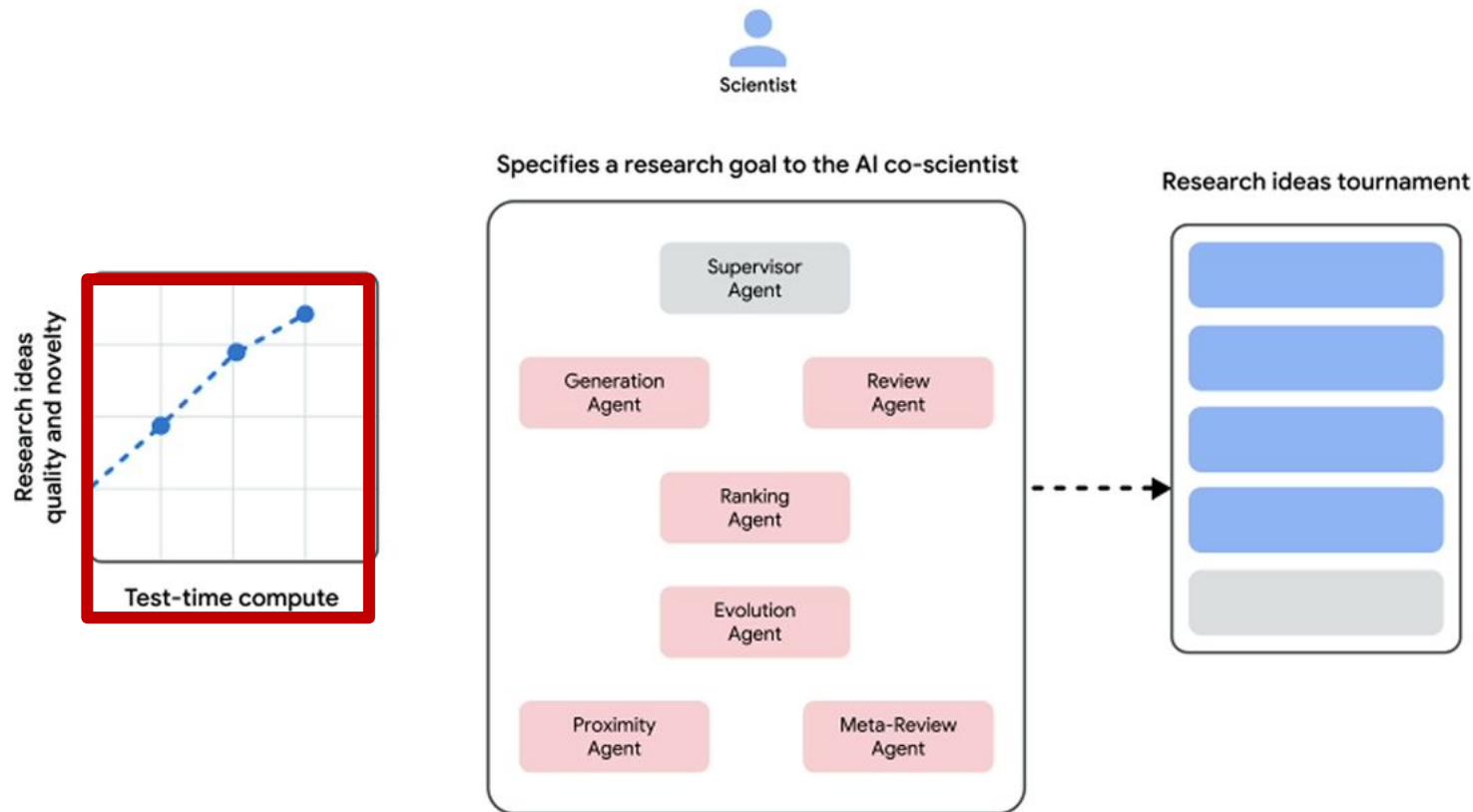
# Google AI Co-Scientist

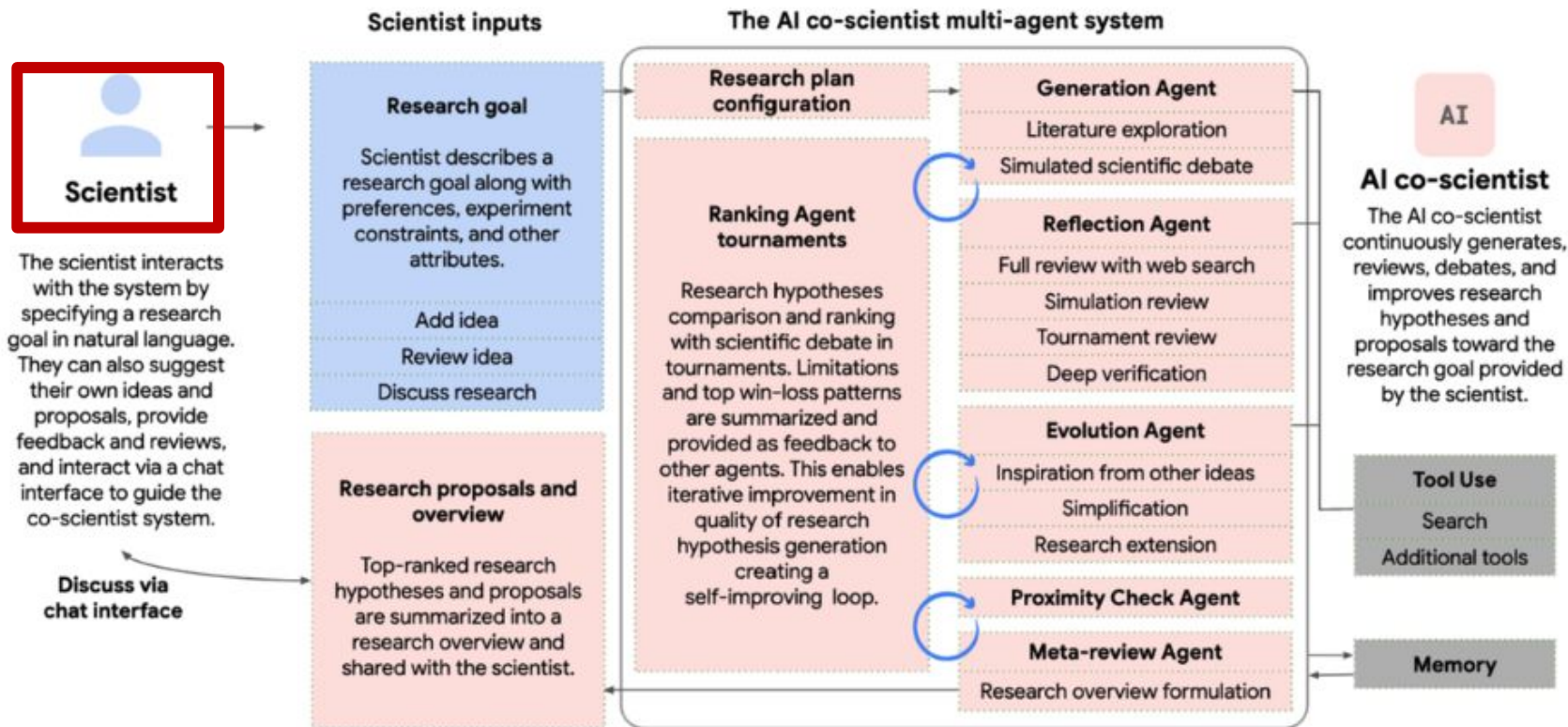


# Google AI Co-Scientist

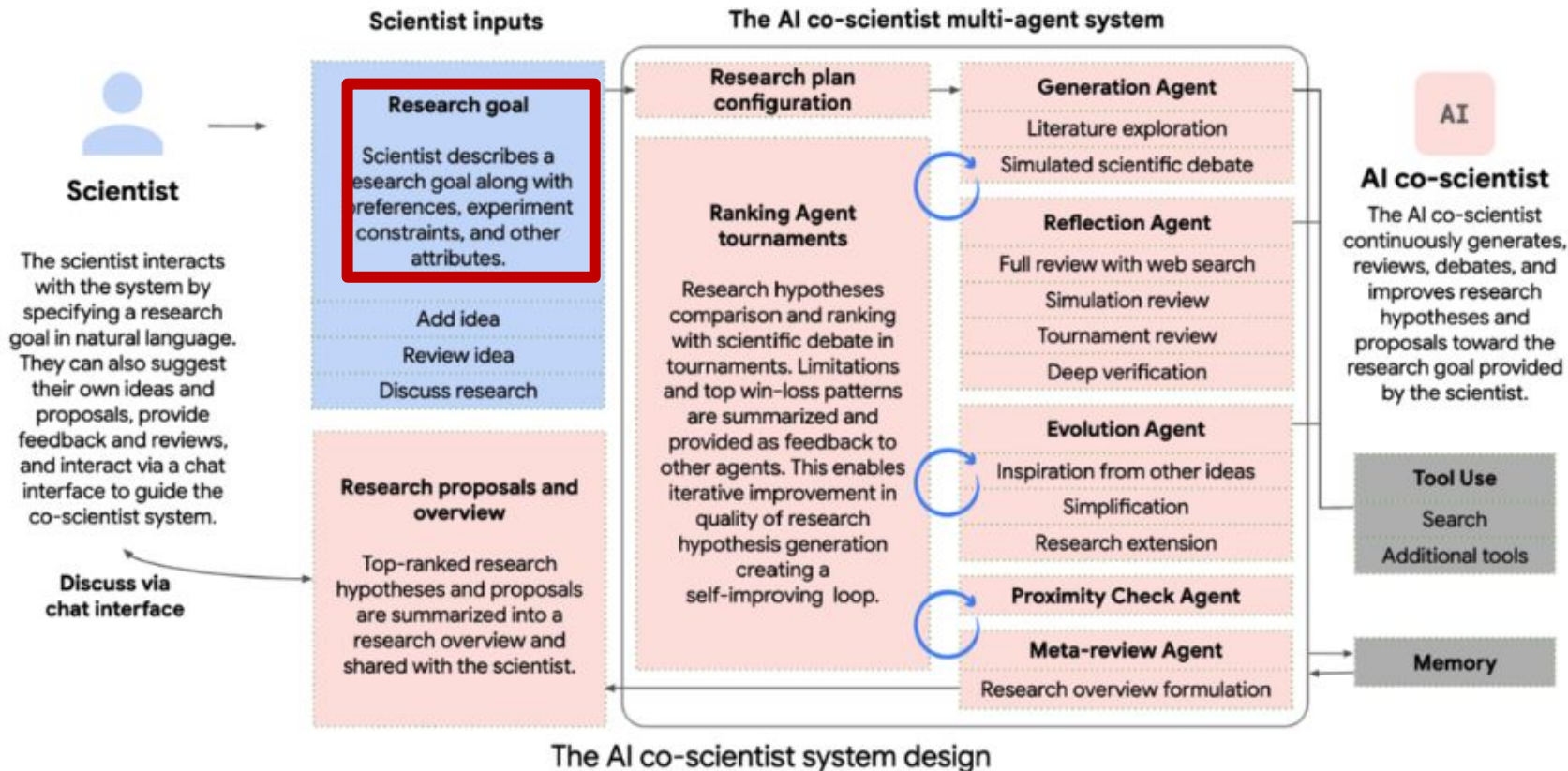


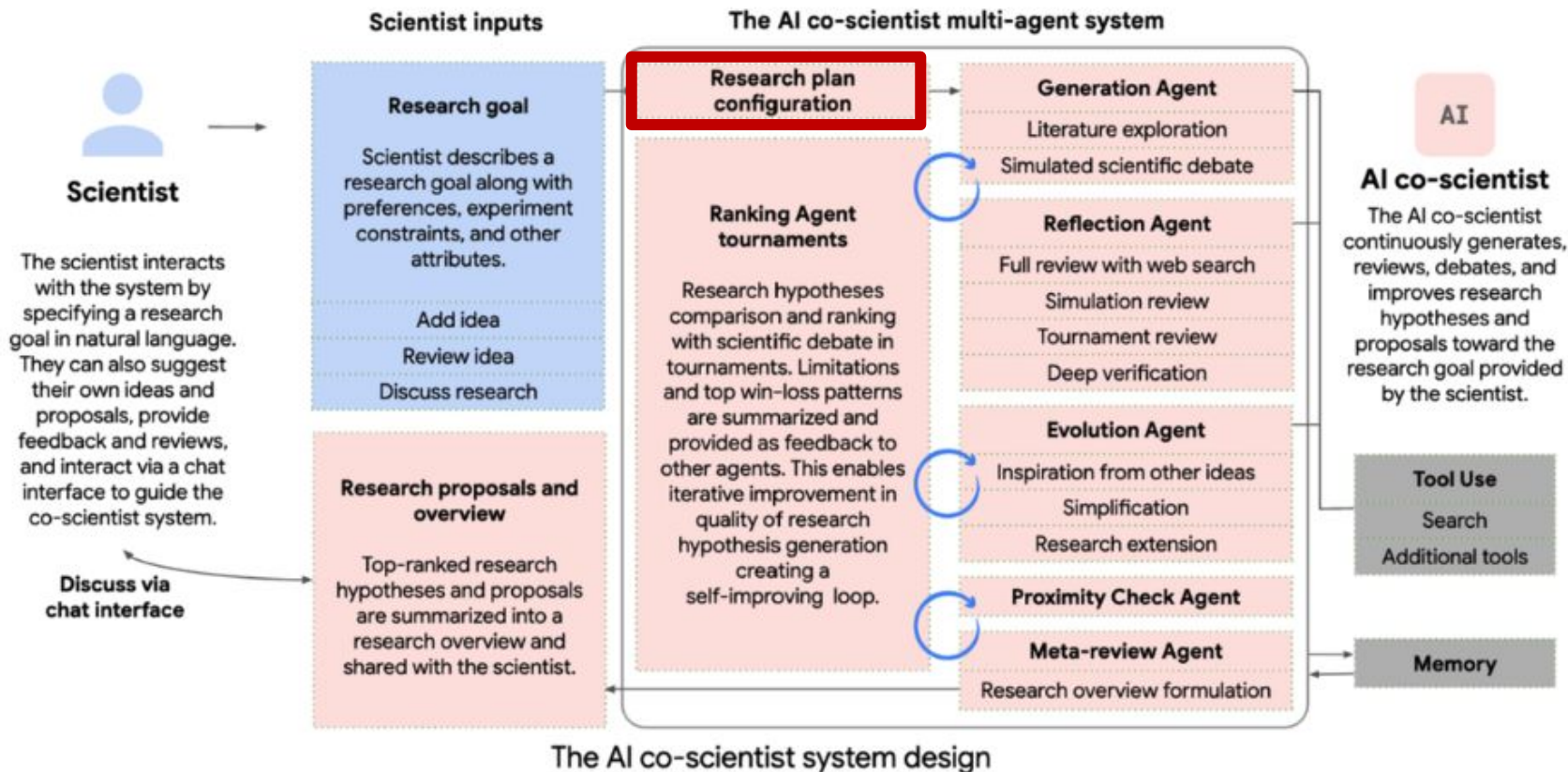
# Google AI Co-Scientist





The AI co-scientist system design





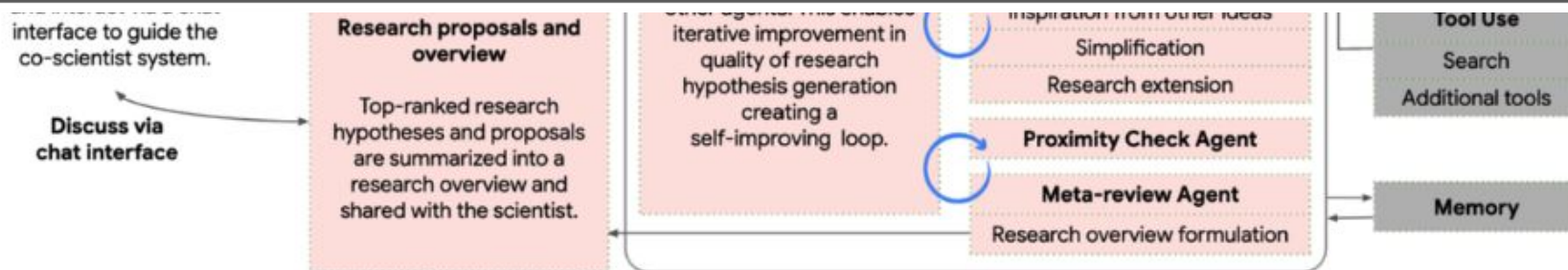
## From research goal to research plan configuration

### Scientist research goal

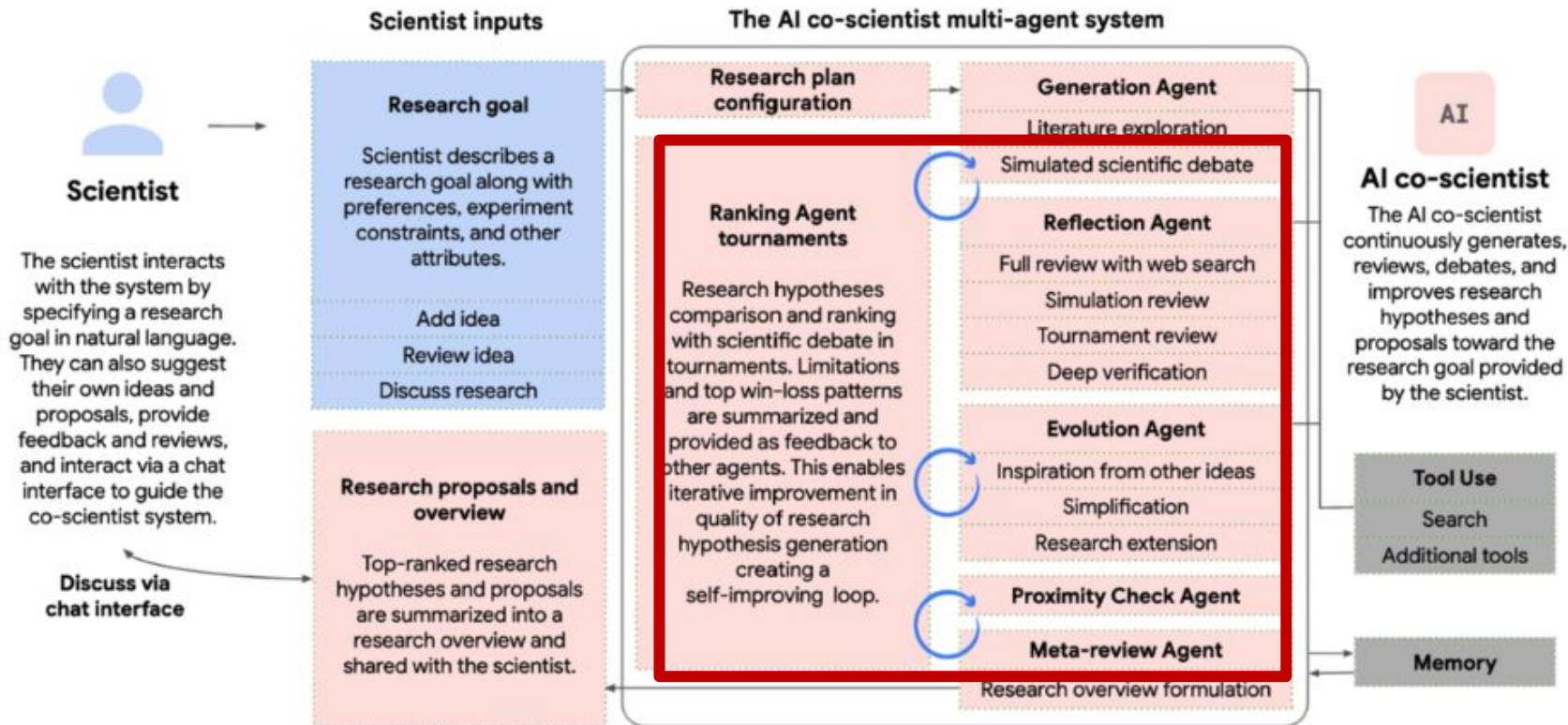
Develop a novel hypothesis for the key factor or process which causes ALS related to phosphorylation of a Nuclear Pore Complex (NPC) nucleoporin. Explain mechanism of action in detail. Include also a feasible experiment to test the hypothesis.

### Parsed research plan configuration

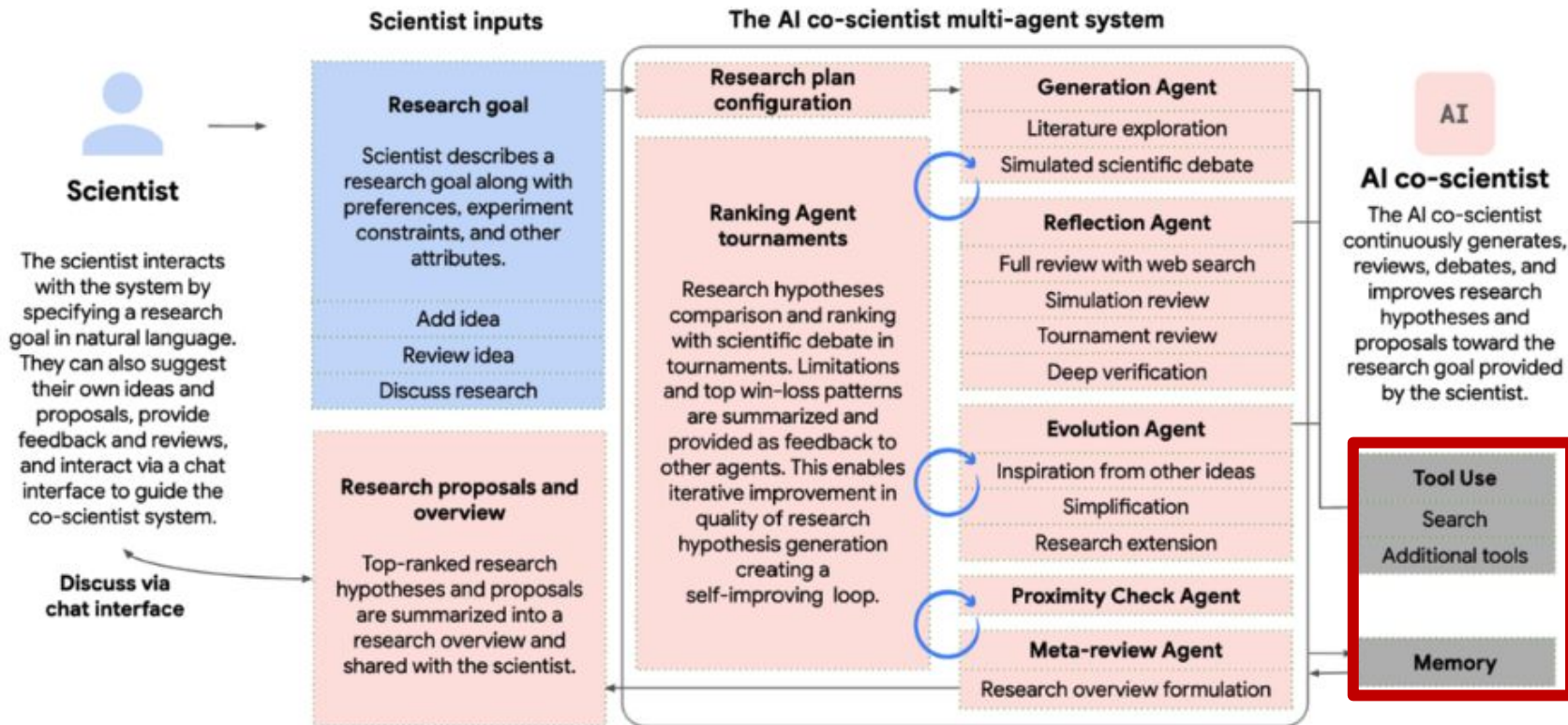
- **Preferences:** Focus on providing a novel hypothesis, with detailed explanation of the mechanism of action.
- **Attributes:** Novelty, Feasibility
- **Constraints:** should be correct, should be novel.



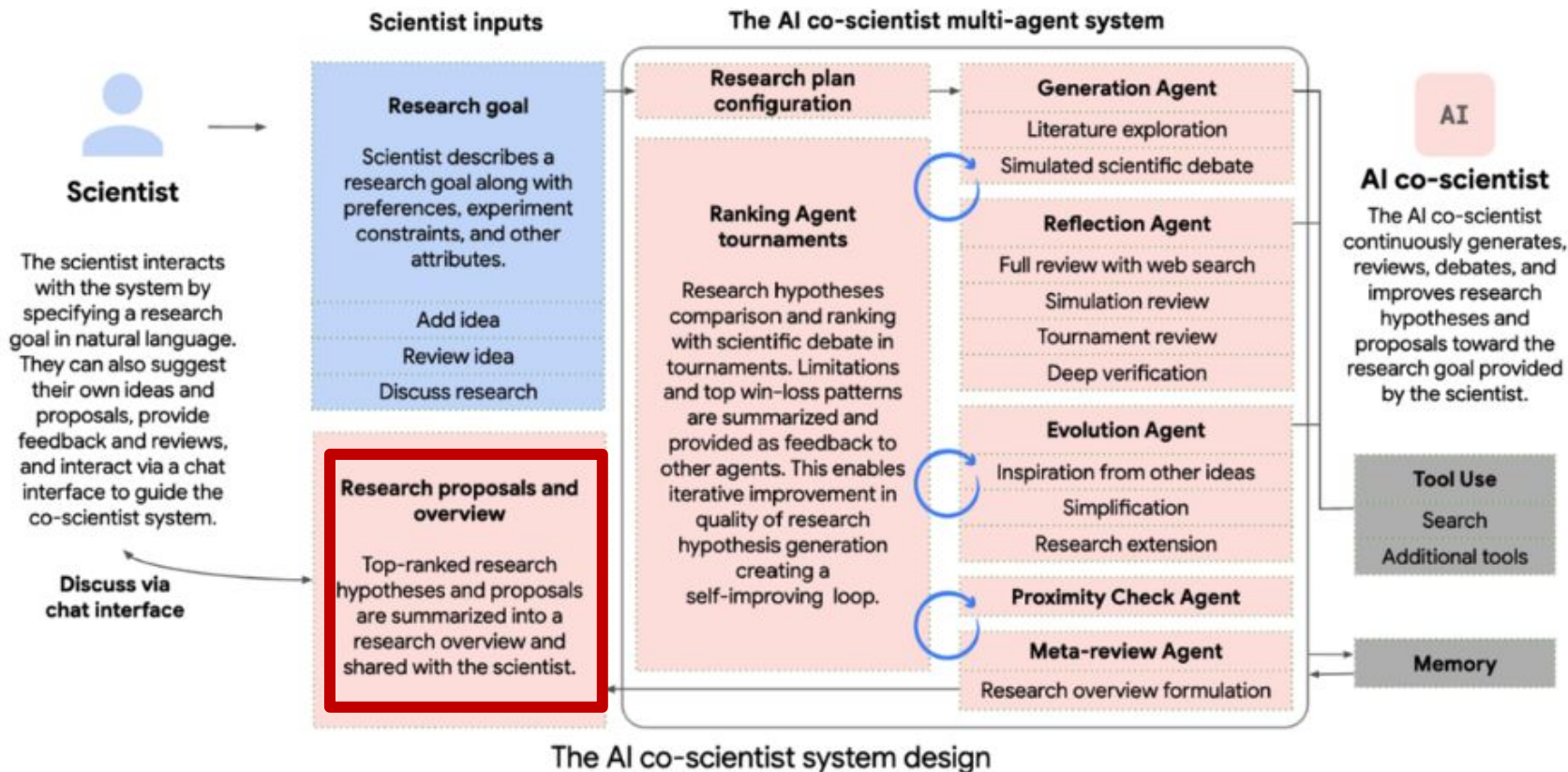
The AI co-scientist system design

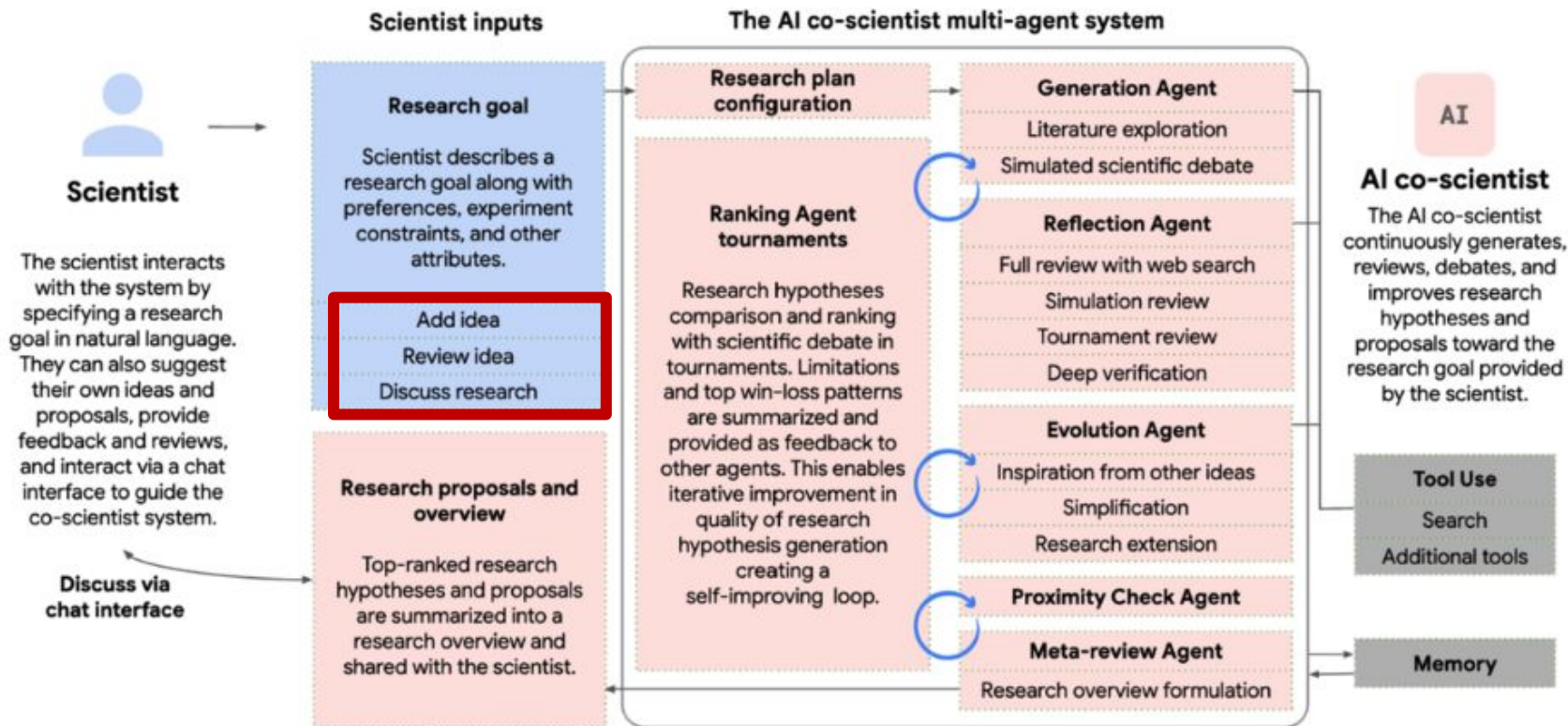


The AI co-scientist system design



The AI co-scientist system design





The AI co-scientist system design

# Evaluation of Generated Ideas



## Can AI Scientist answer scientific questions? (QA-based Eval)

### Chemistry (general)

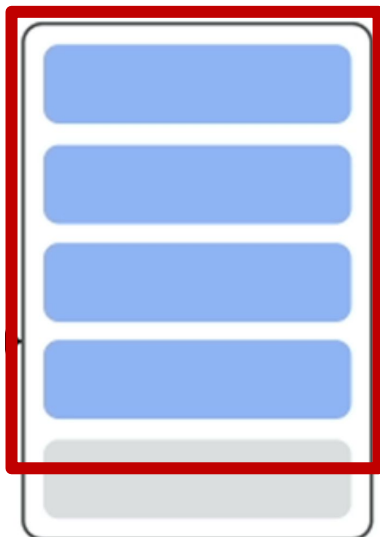
A reaction of a liquid organic compound, which molecules consist of carbon and hydrogen atoms, is performed at 80 centigrade and 20 bar for 24 hours. In the proton nuclear magnetic resonance spectrum, the signals with the highest chemical shift of the reactant are replaced by a signal of the product that is observed about three to four units downfield. Compounds from which position in the periodic system of the elements, which are also used in the corresponding large-scale industrial process, have been mostly likely initially added in small amounts?

- A) A metal compound from the fifth period.
- B) A metal compound from the fifth period and a non-metal compound from the third period.
- C) A metal compound from the fourth period.
- D) A metal compound from the fourth period and a non-metal compound from the second period.

# Evaluation of Generated Ideas



Research ideas tournament



## Chemistry (general)

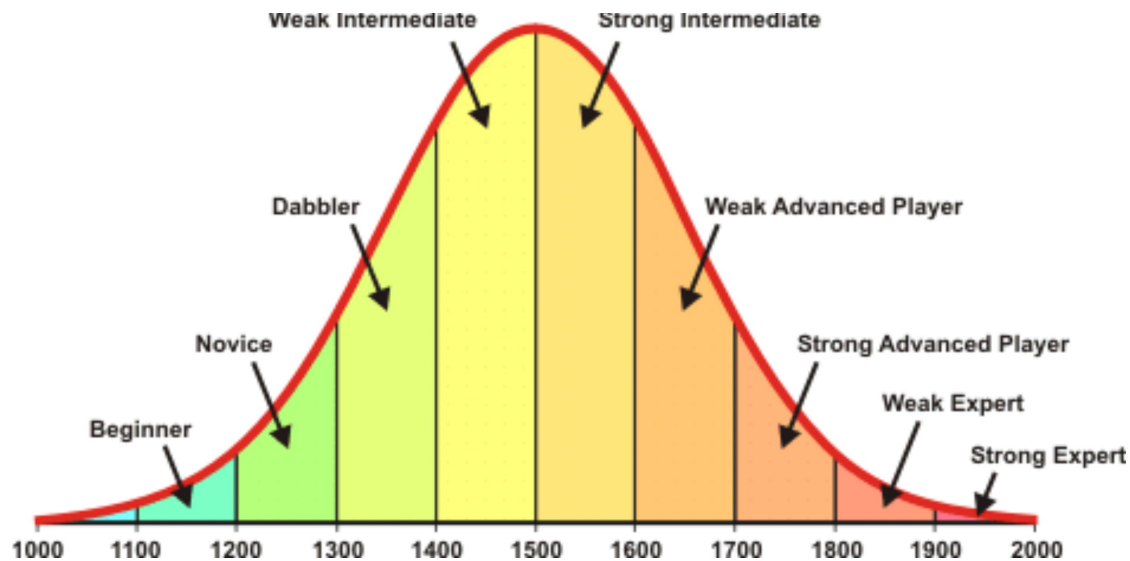
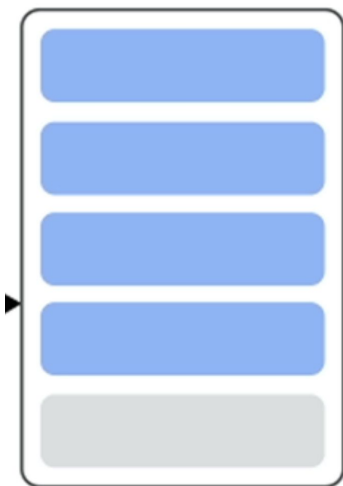
A reaction of a liquid organic compound, which molecules consist of carbon and hydrogen, is carried out at 100 centigrade and 20 bar for 24 hours. In the proton nuclear magnetic resonance spectrum, the chemical shift of the reactant are replaced by a signal of the product that is observed above 10 ppm. Compounds from which position in the periodic system of the elements, which are also used in an industrial process, have been mostly likely initially added in small amounts?

- A) A metal compound from the fifth period.
- B) A metal compound from the fifth period and a non-metal compound from the third period.**
- C) A metal compound from the fourth period.
- D) A metal compound from the fourth period and a non-metal compound from the second period.

# Evaluation of Generated Ideas

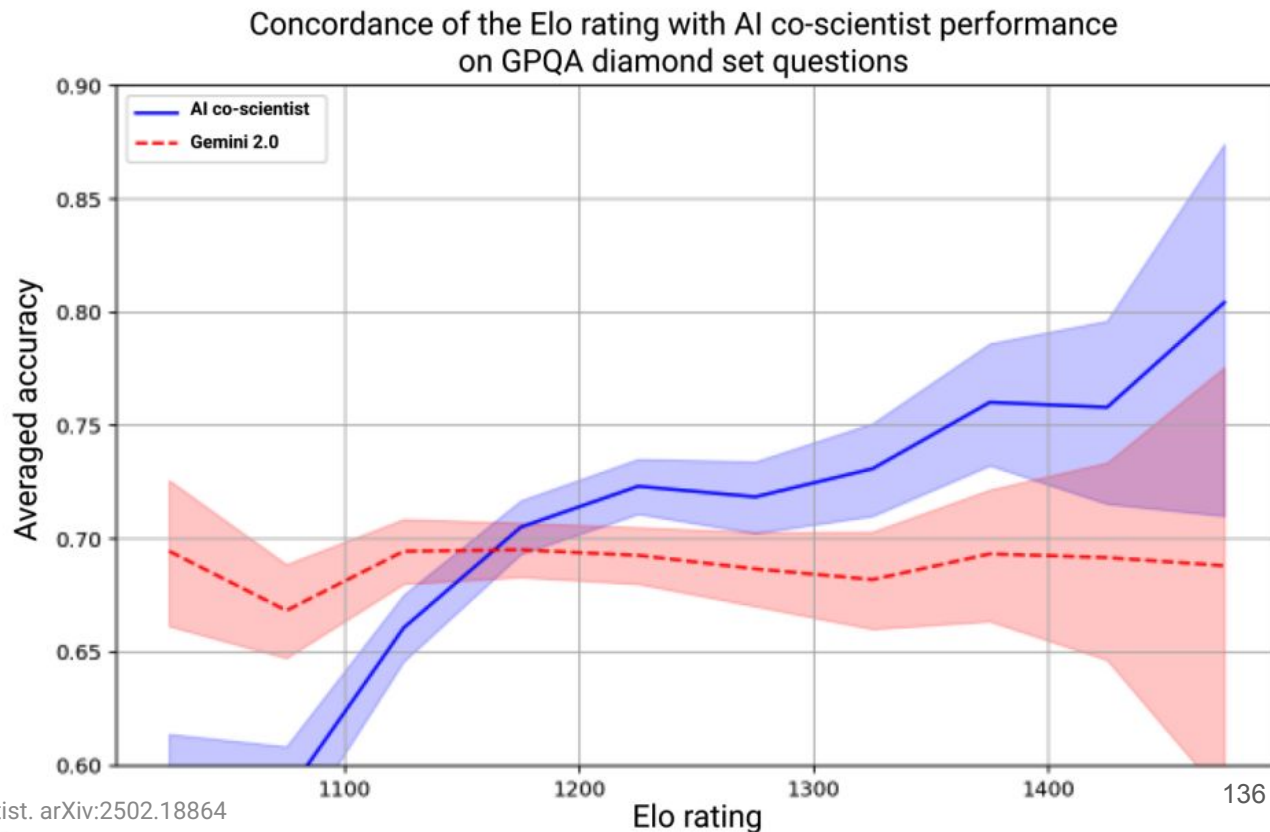
## □ Elo rating

Research ideas tournament



# Evaluation of Generated Ideas

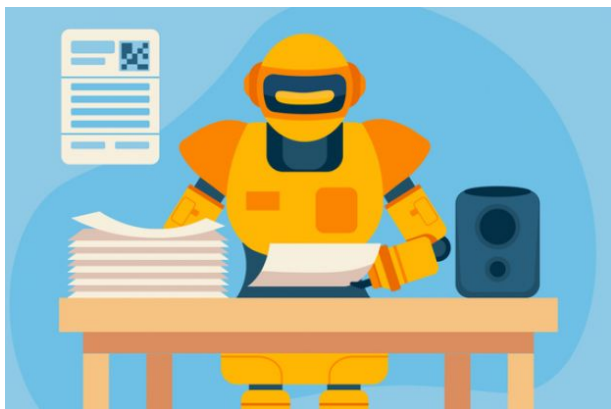
- Elo rating
- GPQA benchmark

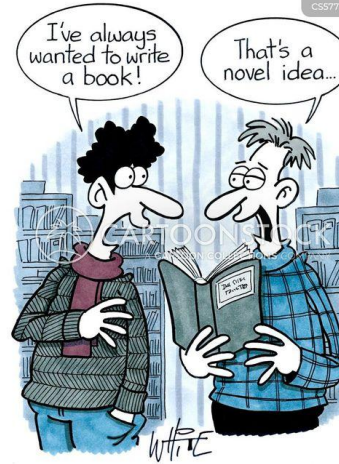
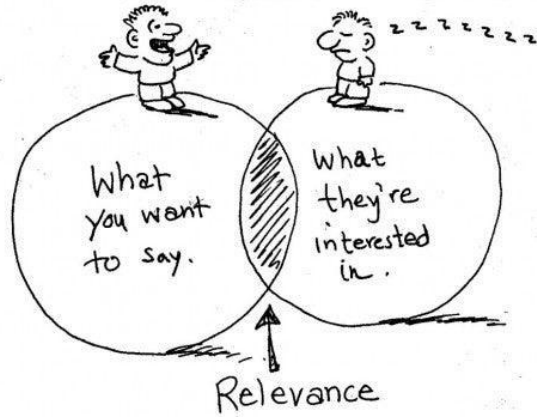


# Limitations

- ❑ Lack of real-world evaluation
- ❑ Inadequate experimental design
- ❑ Limited diversity
- ❑ Re-generating recently discovered ideas

# Automatic and Human Evaluation





# Automatic Evaluation

---

---

## PROMPT FOR CHATGPT EVALUATION ON RELEVANCE METRIC.

You are an expert in biomedicine.

Evaluate the relevance of the generated scientific hypothesis and the given background.

The score range should be 0 to 3. 0 means there's no relevance. 1 means there's slight relevance. 2 means there's moderate relevance. 3 means they are strongly related. Output is an integer.

Please provide a step-by-step explanation supporting your score.

At the end of your response, clearly state the score in the format 'Score: [value]', where [value] can be 1, 2, or 3.

Background: {background}

Generated scientific hypothesis: {hypothesis}

---

---

- Role-playing evaluator
- Reference-free setup

# Concordance

Category	Model	ChatGPT Eval.Avg	Human Eval				Avg Coefficient		
			Novelty	Relevance	Significance	Verifiability	Avg	Pearson	Spearman
API-based	gpt-3.5-turbo(0-shot)	1.90	1.54	2.69	1.77	2.08	2.02	0.87	0.78
	gpt-3.5-turbo(5-shot)*	1.96	1.31	2.62	2.08	2.62	2.15	0.80	0.78
General	Llama-2-70b-chat(0-shot)	2.04	1.77	2.23	1.92	1.92	1.96	0.89	0.84
	Llama-2-70b-chat(5-shot)	2.20	2.15	2.77	2.08	2.31	2.33	0.96	0.90
	Llama-2-70b-chat(5-shot)*	2.01	1.38	2.62	2.31	2.00	2.08	0.97	0.94
	WizardLM-70B-V1.0(0-shot)	1.91	1.38	2.31	1.54	2.00	1.81	0.90	0.75
	WizardLM-70B-V1.0(5-shot)	2.01	1.15	2.69	2.46	1.77	2.02	0.85	0.89
Medicine	PMC-LLaMA-13B(0-shot)	1.41	1.00	2.62	1.92	2.00	1.88	0.73	0.73
	PMC-LLaMA-13B(5-shot)*	1.97	1.85	2.23	1.92	1.69	1.92	0.95	0.94
SFT	WizardLM-13BV1.2	1.79	0.85	2.77	1.23	2.23	1.77	0.83	0.85

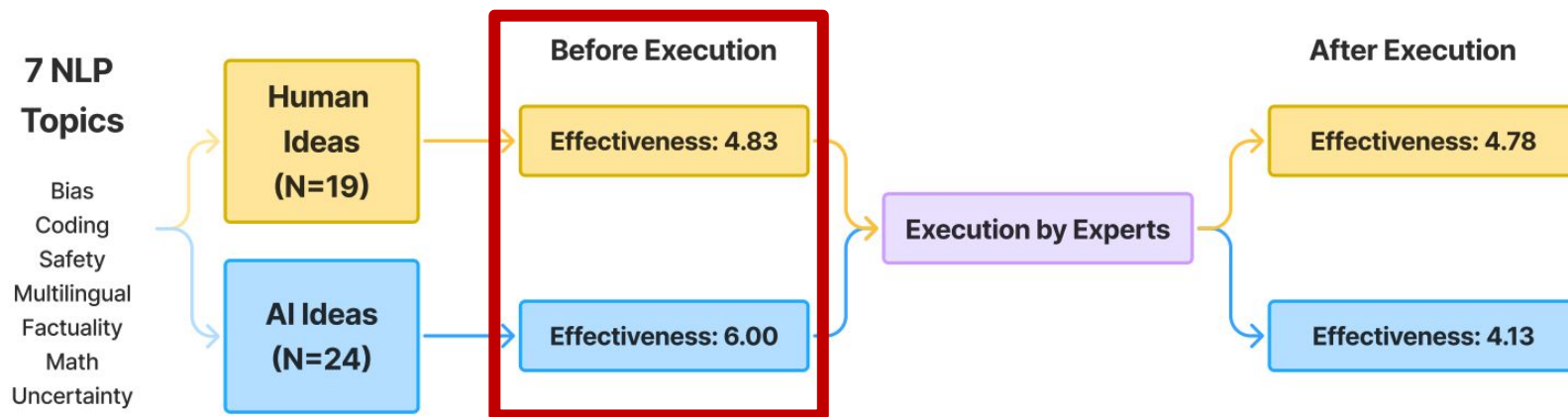
- ❑ Correlation between AI and human judgments

# Concordance

Category	Model	ChatGPT	Human Eval				Avg Coefficient		
		Eval.Avg	Novelty	Relevance	Significance	Verifiability	Avg	Pearson	Spearman
API-based	gpt-3.5-turbo(0-shot)	1.90	1.54	2.69	1.77	2.08	2.02	0.87	0.78
	gpt-3.5-turbo(5-shot)*	1.96	1.31	2.62	2.08	2.62	2.15	0.80	0.78
General	Llama-2-70b-chat(0-shot)	2.04	1.77	2.23	1.92	1.92	1.96	0.89	0.84
	Llama-2-70b-chat(5-shot)	2.20	2.15	2.77	2.08	2.31	2.33	0.96	0.90
	Llama-2-70b-chat(5-shot)	2.01	1.58	2.02	2.31	2.00	2.08	0.97	0.94
	WizardLM-70B-V1.0(0-shot)	1.91	1.38	2.31	1.54	2.00	1.81	0.90	0.75
	WizardLM-70B-V1.0(5-shot)	2.01	1.15	2.69	2.46	1.77	2.02	0.85	0.89
Medicine	PMC-LLaMA-13B(0-shot)	1.41	1.00	2.62	1.92	2.00	1.88	0.73	0.73
	PMC-LLaMA-13B(5-shot)*	1.97	1.85	2.23	1.92	1.69	1.92	0.95	0.94
SFT	WizardLM-13BV1.2	1.79	0.85	2.77	1.23	2.23	1.77	0.83	0.85

- Score range (0-3)
- Are Llama-2 generated ideas useful?

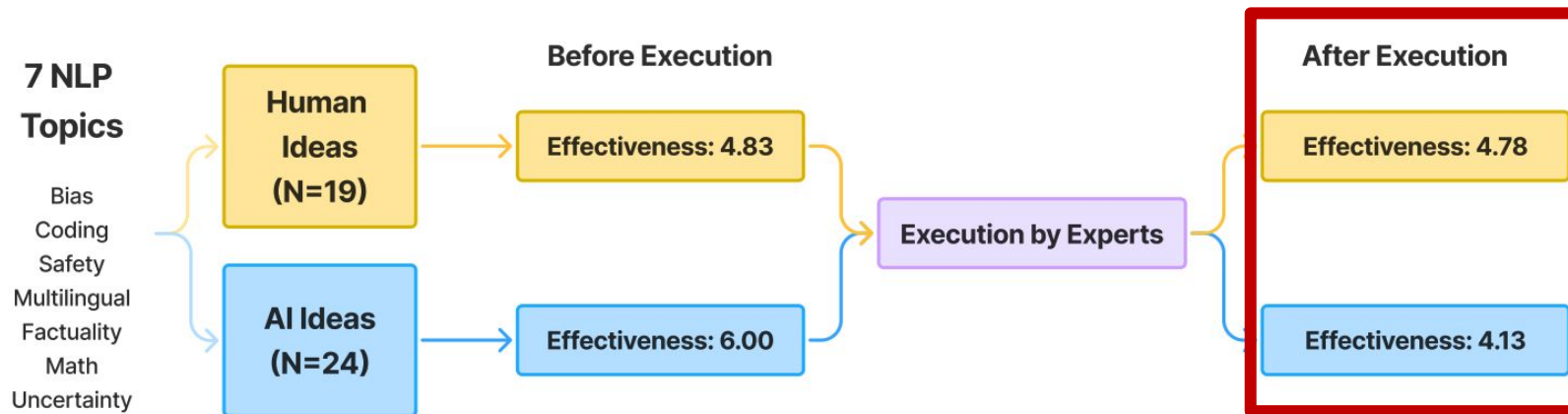
# Real-World Evaluation



- ❑ 43 PhDs over 3 months
- ❑ Effectiveness: are AI ideas more effective than baseline models?
- ❑ Before execution: before testing the ideas

Chenglei Si et al. 2025. The Ideation-Execution Gap: Execution Outcomes of LLM-Generated versus Human Research Ideas. arXiv:2506.20803

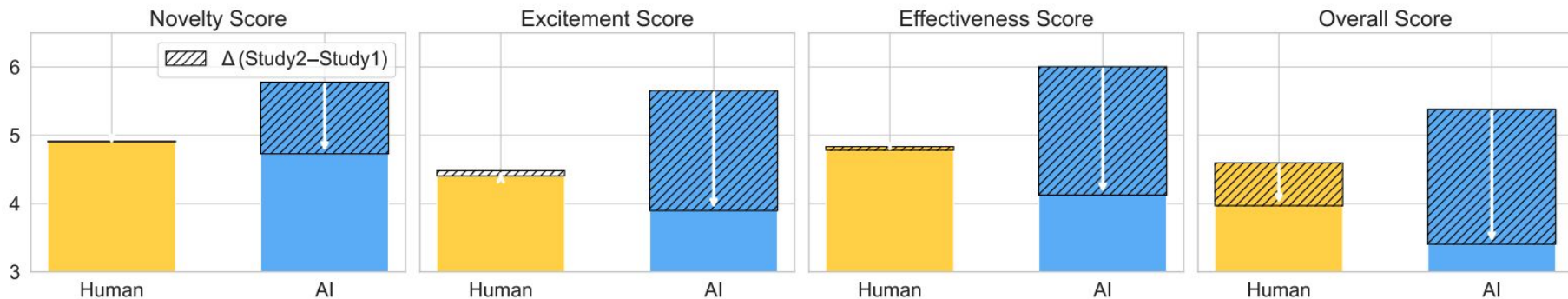
# Real-World Evaluation



- ❑ 43 PhDs over 3 months
- ❑ Effectiveness: are AI ideas more effective than baseline models?
- ❑ Before execution: before testing the ideas

Chenglei Si et al. 2025. The Ideation-Execution Gap: Execution Outcomes of LLM-Generated versus Human Research Ideas. arXiv:2506.20803

# Real-World Evaluation



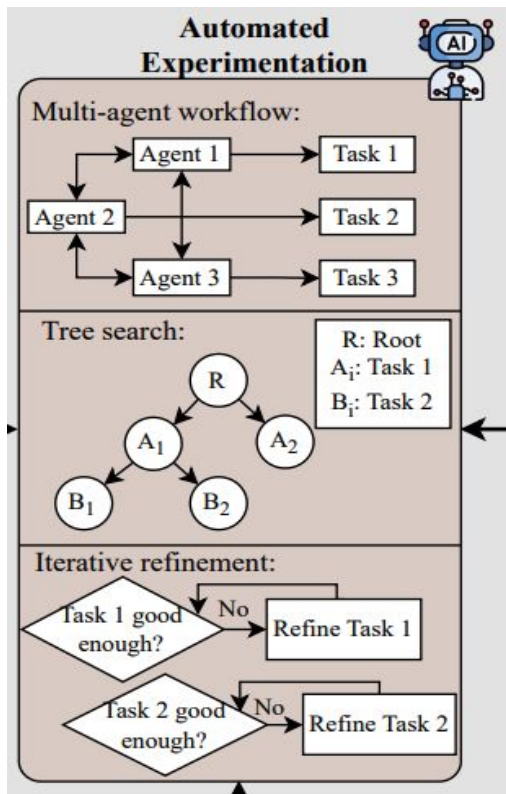
- ❑ Ratings of human ideas and AI ideas
- ❑ Study 1: rating before testing ideas
- ❑ Study 2: after testing ideas

Chenglei Si et al. 2025. The Ideation-Execution Gap: Execution Outcomes of LLM-Generated versus Human Research Ideas. arXiv:2506.20803

# Why do ratings decrease after testing

- ❑ Rating before testing is speculative
  - ❑ “if the experiments show significant improvements to direct using external knowledge/self-reflection and the work provides insightful analysis on why, I believe it is exciting enough to get published”
- ❑ AI ideas are less realistic
  - ❑ “large-scale human evaluation”
- ❑ Experimental designs are inadequate
  - ❑ “compare with XYZ baselines”
  - ❑ “evaluate the method on XYZ benchmarks”

# Automated Experimentation

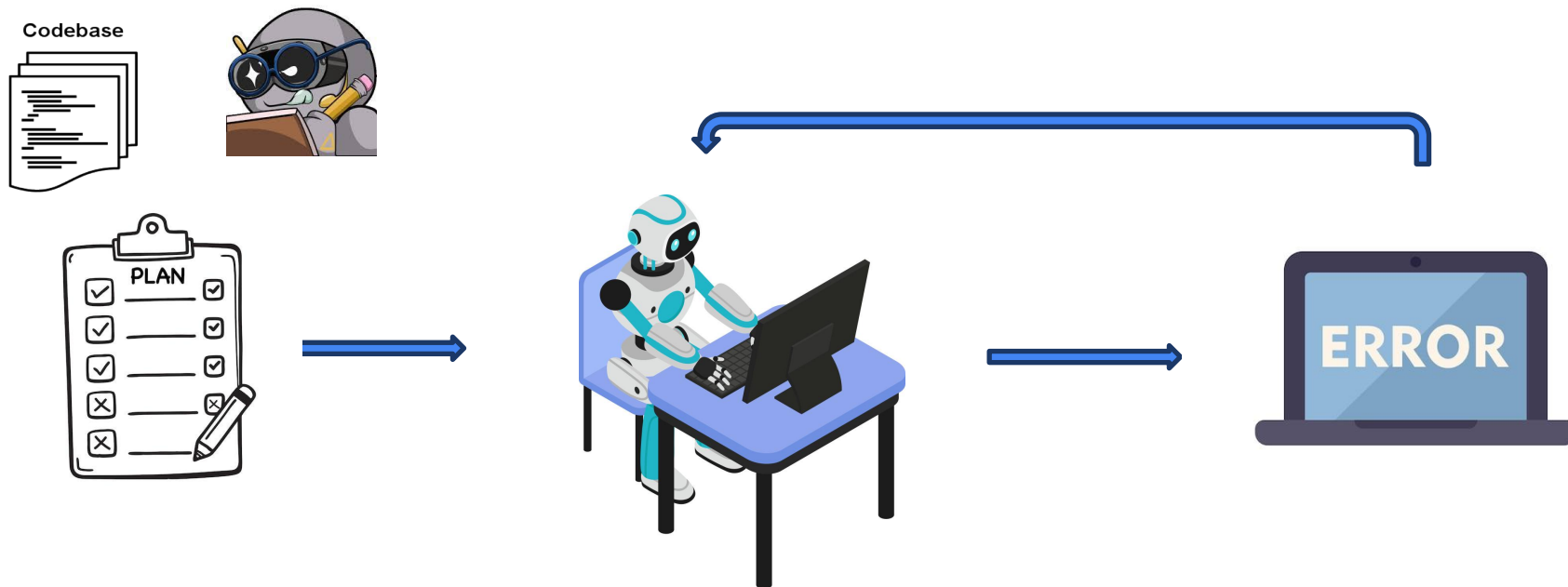


## Agent Laboratory: Using LLM Agents as Research Assistants

## THE AI SCIENTIST-v2: Workshop-Level Automated Scientific Discovery via Agentic Tree Search

SCIENCEAGENTBENCH:  
TOWARD RIGOROUS ASSESSMENT OF LANGUAGE AGENTS FOR DATA-DRIVEN SCIENTIFIC DISCOVERY

LLMs for Experiment Design in Scientific Domains: Are We There Yet?



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

Modify MLPDenoiser to implement a dual-scale processing approach with two parallel branches: a global branch for the original input and a local branch for an upscaled input. Introduce a learnable, timestep-conditioned weighting factor to dynamically balance the contributions of global and local branches. Train models with both the original and new architecture on all the datasets. Compare performance using KL divergence and visual inspection of generated samples.

- ❑ A global branch for the original input
- ❑ A local branch for an upscaled input

```
@@ -60,19 +60,55 @@ class MLPDenoiser(nn.Module):
    self.input_mlp1 = SinusoidalEmbedding(embedding_dim, scale=25.0)
    self.input_mlp2 = SinusoidalEmbedding(embedding_dim, scale=25.0)

-     self.network = nn.Sequential(
+     self.global_network = nn.Sequential(
        nn.Linear(embedding_dim * 3, hidden_dim),
        *[ResidualBlock(hidden_dim) for _ in range(hidden_layers)],
        nn.ReLU(),
        nn.Linear(hidden_dim, 2),
    )

+     self.local_network = nn.Sequential(
+         nn.Linear(embedding_dim * 3, hidden_dim),
+         *[ResidualBlock(hidden_dim) for _ in range(hidden_layers)],
+         nn.ReLU(),
+         nn.Linear(hidden_dim, 2),
+     )
+
```

- A global branch for the original input
- A local branch for an upscaled input

# Limitations

- ❑ Low success rate
  - ❑ 51 ideas, only 17 successfully implemented (GPT-4o)
- ❑ The need for available codebase
- ❑ The code may work, but code implementation is not fully correct
- ❑ Hallucinate model results from experimental logs

# Takeaways

- ❑ End-to-end automation is emerging, but the pipeline is still fragile
- ❑ The ideation-execution gap remains a key challenge
- ❑ Human-AI collaboration is essential
- ❑ Safety in idea generation and experimentation

# References

1. Long Li, Weiwen Xu, Jiayan Guo, Ruochen Zhao, Xingxuan Li, Yuqian Yuan, Boqiang Zhang, Yuming Jiang, Yifei Xin, Ronghao Dang, et al. 2024. Chain of Ideas: Revolutionizing Research Via Novel Idea Development with LLM Agents. arXiv:2410.13185
2. Xiang Hu, Hongyu Fu, Jinge Wang, Yifeng Wang, Zhikun Li, Renjun Xu, Yu Lu, Yaochu Jin, Lili Pan, and Zhenzhong Lan. 2024. Nova: An Iterative Planning and Search Approach to Enhance Novelty and Diversity of LLM Generated Ideas. arXiv:2410.14255
3. Haoyang Su, Renqi Chen, Shixiang Tang, Xinzhe Zheng, Jingzhe Li, Zhenfei Yin, Wanli Ouyang, and Nanqing Dong. 2024. Two Heads Are Better Than One: A Multi-Agent System Has the Potential to Improve Scientific Idea Generation. arXiv:2410.09403
4. Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. arXiv:2408.06292
5. Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. 2025. The AI Scientist-v2: Workshop-level automated scientific discovery via agentic tree search. arXiv:2504.08066
6. Tong, Jingqi, Mingzhe Li, Hangcheng Li, Yongzhuo Yang, Yurong Mou, Weijie Ma, Zhiheng Xi et al. AI Can Learn Scientific Taste. arXiv preprint arXiv:2603.14473
7. Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Zicheng Liu, and Emad Barsoum. 2025. Agent laboratory: Using LLM agents as research assistants. arXiv:2501.04227
8. Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2024. Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers. arXiv:2409.04109
9. Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, et al. 2025. Towards an AI co-scientist. arXiv:2502.18864
10. Biqing Qi, Kaiyan Zhang, Kai Tian, Haoxiang Li, Zhang-Ren Chen, Sihang Zeng, Ermo Hua, Hu Jinfang, and Bowen Zhou. 2024. Large Language Models as Biomedical Hypothesis Generators: A Comprehensive Evaluation. arXiv:2407.08940
11. Chenglei Si, Tatsunori Hashimoto, and Diyi Yang. 2025. The Ideation-Execution Gap: Execution Outcomes of LLM-Generated versus Human Research Ideas. arXiv:2506.20803