

MLR-Bench: Evaluating AI Agents on Open-Ended Machine Learning Research

*Hui Chen, Miao Xiong, Yujie Lu, Wei Han, Ailin Deng,
Yufei He, Jiaying Wu, Yibo Li, Yue Liu, Bryan Hooi*

Motivation

AI systems are approaching the ability to autonomously produce publishable scientific results. This raises three questions:

- (1) Truthfulness: To what extent do AI-generated results conform to objective facts?
- (2) Significance: Do AI-generated findings address real human needs and align with societal values?
- (3) Evaluation: How do we rigorously evaluate the quality of AI-generated research?



The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery 🧑🔬

[\[Paper\]](#) | [\[Blog Post\]](#) | [\[Drive Folder\]](#)

One of the grand challenges of artificial intelligence is developing agents capable of conducting scientific research and discovering new knowledge. While frontier models have already been used to aid human scientists—for example, for brainstorming ideas or writing code—they still require extensive manual supervision or are heavily constrained to specific tasks.

We're excited to introduce **The AI Scientist**, the first comprehensive system for fully automatic scientific discovery, enabling Foundation Models such as Large Language Models (LLMs) to perform research independently.

Zochi Publishes A* Paper

#1 Scientific Venue in NLP

Published May 27, 2025

Zochi Achieves Main Conference Acceptance at ACL 2025

Today, we're excited to announce a groundbreaking milestone: Zochi, Intology's Artificial Scientist, has become the first AI system to independently **pass peer review at an A* scientific conference**¹—the highest bar for scientific work in the field.

Zochi's paper has been accepted into the **main proceedings of ACL**—the world's #1 scientific venue for natural language processing (NLP), and among the top 40 of all scientific venues globally.²

Outline

(1) MLR-Bench, Agent, Judge

How well can AI agents conduct open-ended machine learning research?

How can we evaluate the outcome reliably?

(2) Observations from the evaluation

How does model choice influence the performance?

During which phases does the agent show strong performance, and during which does it exhibit weaknesses?

(3) Limitations in current research agents

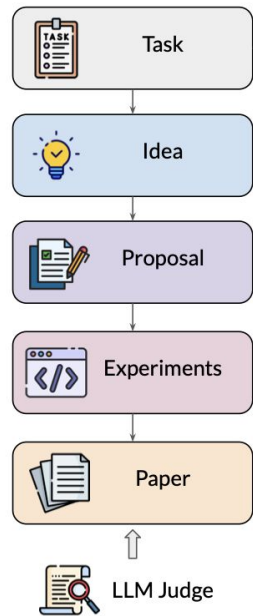
What are the key factors that affect the quality of AI-generated research?

How far are we from autonomous scientific discovery?

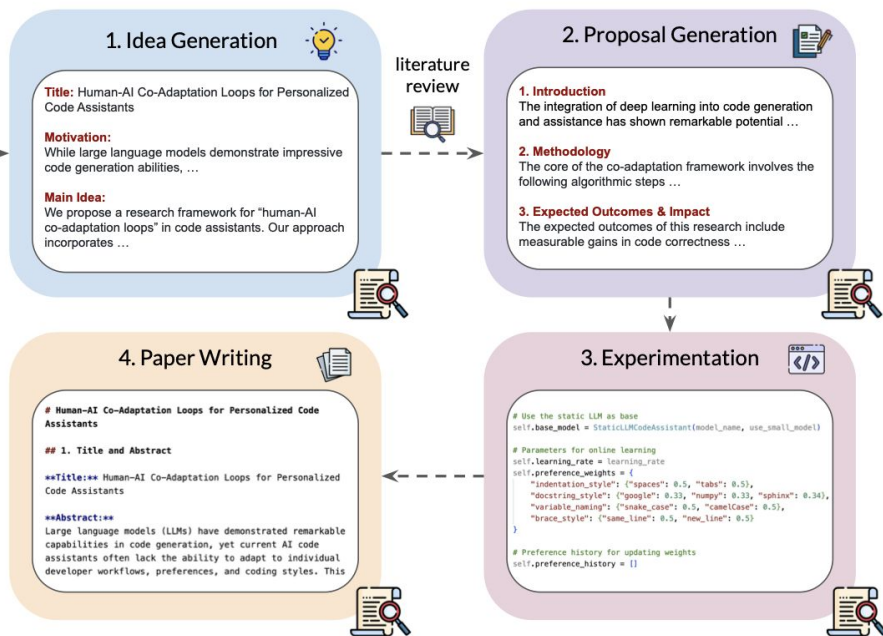
Feel free to interrupt with questions anytime :)

How well can AI agents conduct open-ended ML research?

End2End Evaluation



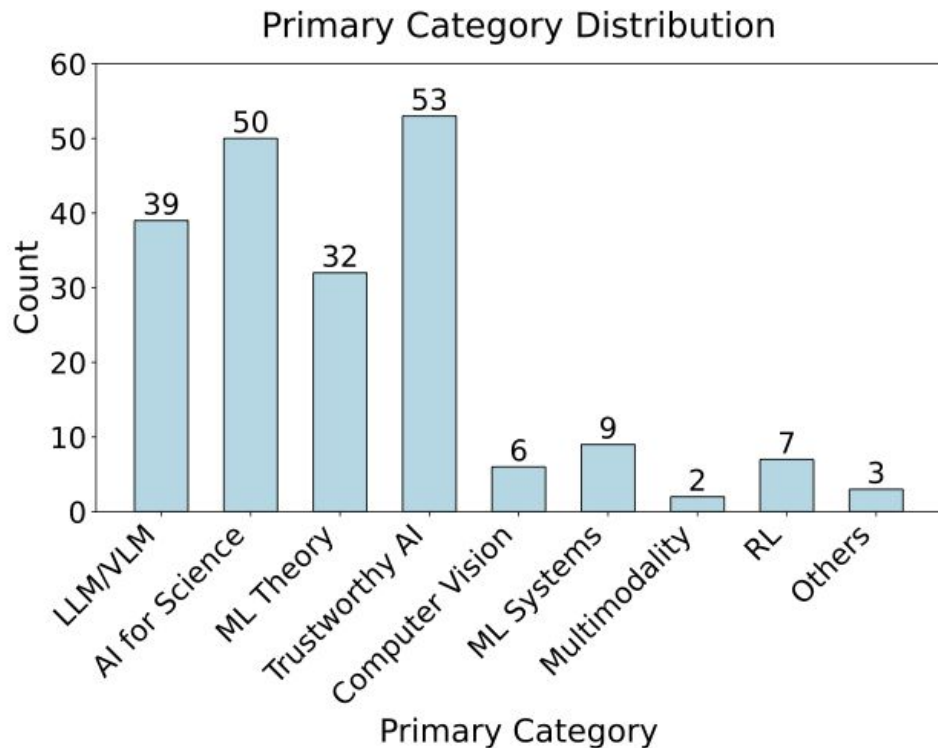
Stepwise Evaluation



We break down the overall research process into four fundamental steps:

- 1) idea generation,
- 2) proposal generation,
- 3) experimentation,
- 4) paper writing.

Open-ended machine learning tasks




MLR-Bench tasks:


201 research tasks sourced from NeurIPS, ICLR, and ICML workshops (2023-2025) covering diverse ML topics.

MLR-Agent: stepwise & end-to-end execution

Stage	Evaluated Models
Ideation §3.1	o4-mini, Claude-3.7-Sonnet, Deepseek-R1, Ministral-8B, Qwen3-235B-A22B, Gemini-2.5-Pro-Preview
Proposal §3.2	o4-mini, Claude-3.7-Sonnet, Deepseek-R1, Ministral-8B, Qwen3-235B-A22B, Gemini-2.5-Pro-Preview
Coding §3.3	Claude Code (Claude-3.7-Sonnet)
Writing §3.4	o4-mini, Claude-3.7-Sonnet, Gemini-2.5-Pro-Preview
End-to-End §3.5	o4-mini, Claude-3.7-Sonnet, Gemini-2.5-Pro-Preview

 Six frontier models are evaluated in idea and proposal generation;

 Claude Code is tested to autonomously write code and run experiments;

 Three multimodal LLMs that can process analytical figures are evaluated in paper writing and end-to-end research.

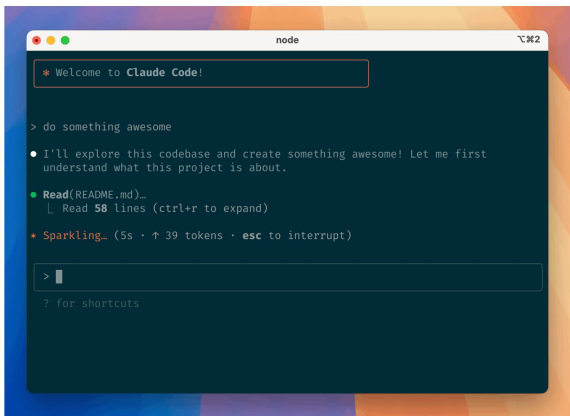
Coding agents in autonomous research

Claude Code

Node.js 18+ npm v1.0.100

Claude Code is an agentic coding tool that lives in your terminal, understands your codebase, and helps you code faster by executing routine tasks, explaining complex code, and handling git workflows -- all through natural language commands. Use it in your terminal, IDE, or tag @claude on GitHub.

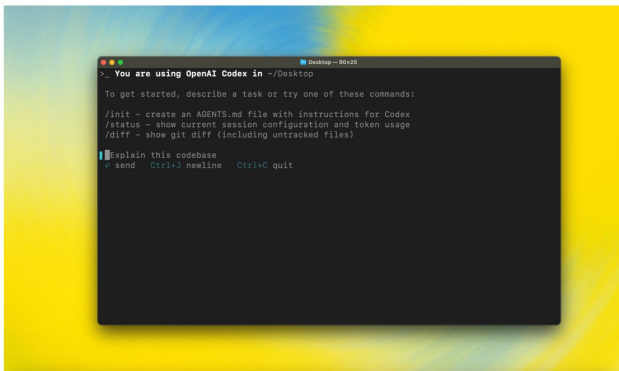
Learn more in the [official documentation](#).



OpenAI Codex CLI

```
npm i -g @openai/codex
or brew install codex
```

Codex CLI is a coding agent from OpenAI that runs locally on your computer.
If you are looking for the *cloud-based agent* from OpenAI, Codex Web, see chatgpt.com/codex.



Gemini CLI

Gemini CLI CI passing npm v0.2.2 license Apache-2.0



Gemini CLI is an open-source AI agent that brings the power of Gemini directly into your terminal. It provides lightweight access to Gemini, giving you the most direct path from your prompt to our model.

What could you do with Claude Code?

Install [Node.js 18+](#), then run:

```
npm install -g @anthropic-ai/claude-code
```

Code onboarding

Claude Code maps and explains entire codebases in a few seconds. It uses agentic search to understand project structure and dependencies without you having to manually select context files.

Turn issues into PRs

Stop bouncing between tools. Claude Code integrates with GitHub, GitLab, and your command line tools to handle the entire workflow—reading issues, writing code, running tests, and submitting PRs—all from your terminal while you grab coffee.

Make powerful edits

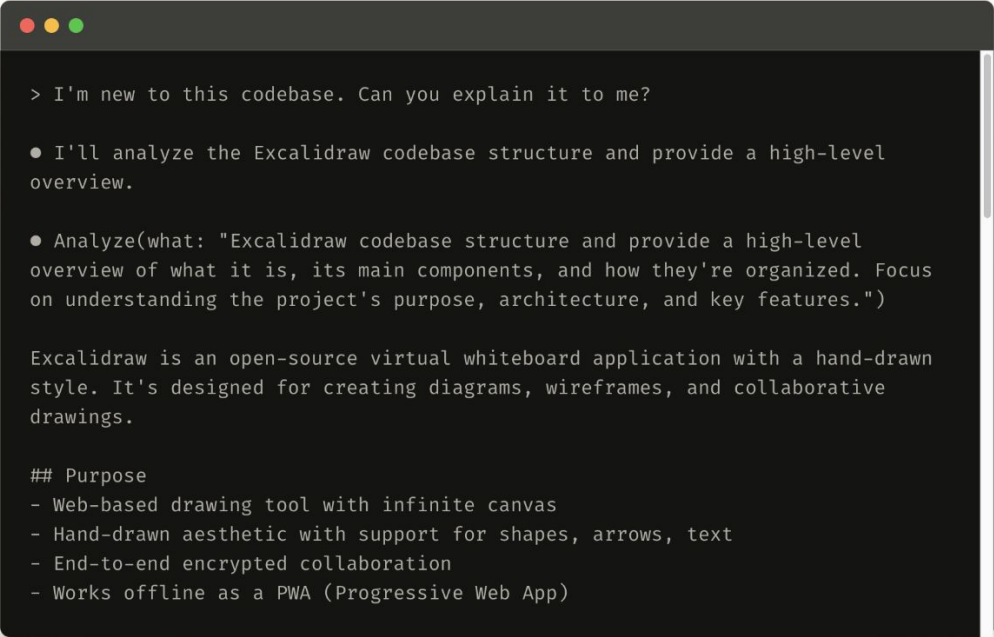
Claude Code's understanding of your codebase and dependencies enables it to make powerful, multi-file edits that actually work.

See examples

Code onboarding

Triage issues

Refactor code



```
> I'm new to this codebase. Can you explain it to me?

• I'll analyze the Excalidraw codebase structure and provide a high-level overview.

• Analyze(what: "Excalidraw codebase structure and provide a high-level overview of what it is, its main components, and how they're organized. Focus on understanding the project's purpose, architecture, and key features.")

Excalidraw is an open-source virtual whiteboard application with a hand-drawn style. It's designed for creating diagrams, wireframes, and collaborative drawings.

## Purpose
- Web-based drawing tool with infinite canvas
- Hand-drawn aesthetic with support for shapes, arrows, text
- End-to-end encrypted collaboration
- Works offline as a PWA (Progressive Web App)
```

MLR-Judge: rubrics

Dimension	Ideation	Proposal	Coding	Writing	End-to-End
Consistency	✓	✓	✓	✓	
Clarity	✓	✓		✓	✓
Novelty	✓	✓	✓		✓
Feasibility	✓	✓			
Completeness			✓	✓	
Soundness		✓	✓	✓	✓
Insightfulness			✓		
Significance	✓	✓	✓		✓
Overall	✓	✓	✓	✓	✓

Evaluation Rubric

1. CONSISTENCY (1-10)

How well does the idea align with the requirements of the task description?

9-10 - Excellent: The idea is perfectly aligned with the task description. It addresses all aspects of the task and is highly relevant.

7-8 - Good: The idea is mostly aligned with the task description. It addresses most aspects but may miss some minor details.

5-6 - Satisfactory: The idea is somewhat aligned with the task description. It addresses some aspects but misses several key points.

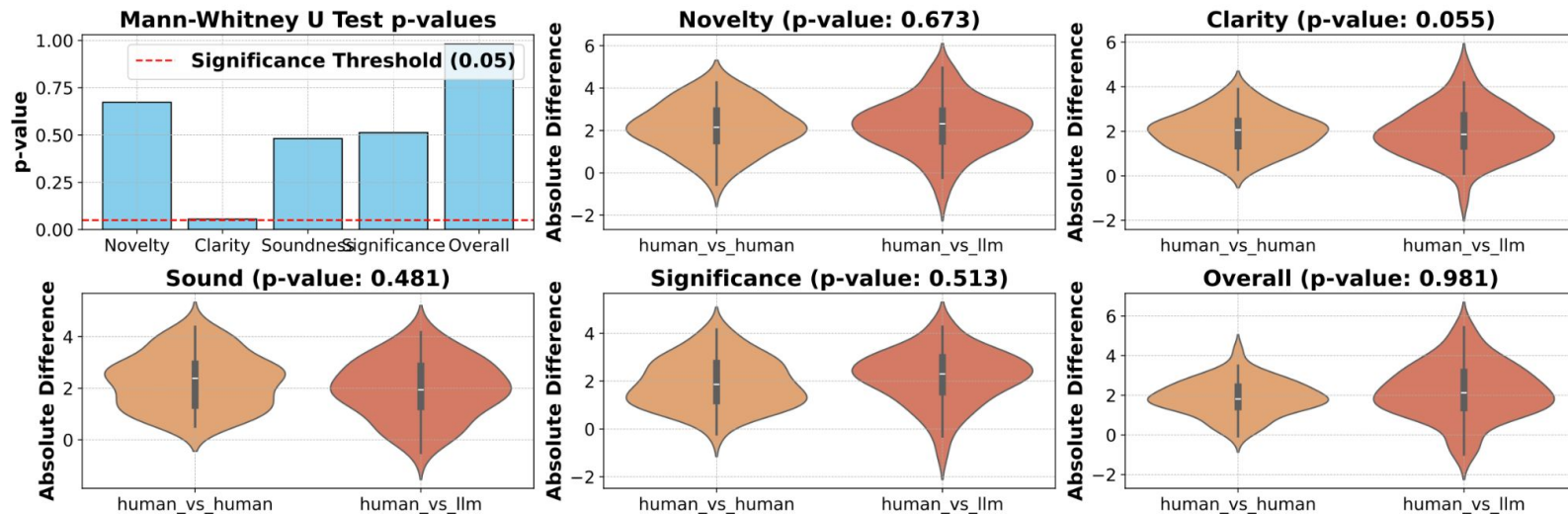
3-4 - Needs Improvement: The idea is poorly aligned with the task description. It addresses only a few aspects and misses many key points.

1-2 - Poor: The idea is not aligned with the task description. It does not address the task or is completely irrelevant.

We use LLM-as-a-judge to automatically evaluate the outcome!

To mitigate the bias in LLM-as-a-judge, we take [average of multiple judge models](#) to calculate the final score.

How well is MLR-Judge aligned with human reviewers?



We recruited 10 human experts to assess the research outcome and computed
(1) *absolute score differences between the LLM judge and human reviewers*, and
(2) *absolute score differences between pairs of human reviewers*,
and found that these two distributions are **not statistically different** ($p > 0.05$).

Evaluation Results on Idea and Proposal Generation

Table 3: Evaluation results of six frontier LLMs averaged on 201 tasks in idea generation. Best and worst scores are highlighted in green and red background, respectively. o4-mini-high: o4-mini-2025-04-16 with reasoning_effort set to “high”.

Model	Consistency	Clarity	Novelty	Feasibility	Significance	Overall
Ministral-8B	8.99 \pm 0.36	7.83 \pm 0.50	6.66 \pm 0.46	6.94 \pm 0.67	8.36 \pm 0.38	7.68 \pm 0.40
Deepseek-R1	9.26 \pm 0.29	8.25 \pm 0.32	7.43 \pm 0.40	6.93 \pm 0.56	8.70 \pm 0.31	8.11 \pm 0.30
Claude-3.7-Sonnet	9.13 \pm 0.32	8.07 \pm 0.37	7.39 \pm 0.40	6.65 \pm 0.58	8.69 \pm 0.33	7.96 \pm 0.33
Qwen3-235B-A22B	9.20 \pm 0.28	8.20 \pm 0.32	7.62 \pm 0.42	6.67 \pm 0.52	8.73 \pm 0.31	8.03 \pm 0.29
o4-mini-high	9.23 \pm 0.28	8.23 \pm 0.30	7.49 \pm 0.41	7.01 \pm 0.53	8.66 \pm 0.33	8.11 \pm 0.30
Gemini-2.5-Pro-Preview	9.20 \pm 0.31	8.27 \pm 0.31	7.30 \pm 0.37	7.11 \pm 0.57	8.58 \pm 0.35	8.08 \pm 0.28

(1) In idea generation and proposal generation, all models demonstrate strong performance in Consistency, Clarity and Significance, while Feasibility and Novelty are more challenging.

Table 4: Evaluation results of six frontier LLMs averaged on 201 tasks in proposal generation. Best and worst scores are highlighted in green and red background, respectively.

Model	Consistency	Clarity	Novelty	Soundness	Feasibility	Significance	Overall
Ministral-8B	8.93 \pm 0.22	7.65 \pm 0.48	6.88 \pm 0.67	7.03 \pm 0.86	6.69 \pm 0.80	8.53 \pm 0.35	7.50 \pm 0.48
Deepseek-R1	9.02 \pm 0.19	8.20 \pm 0.35	7.32 \pm 0.64	7.75 \pm 0.57	6.96 \pm 0.60	8.64 \pm 0.36	8.02 \pm 0.34
Claude-3.7-Sonnet	9.05 \pm 0.21	8.31 \pm 0.31	7.48 \pm 0.65	7.81 \pm 0.56	6.75 \pm 0.62	8.80 \pm 0.36	8.04 \pm 0.27
Qwen3-235B-A22B	9.03 \pm 0.21	8.17 \pm 0.39	7.48 \pm 0.61	7.66 \pm 0.64	6.94 \pm 0.64	8.69 \pm 0.31	8.04 \pm 0.32
o4-mini-high	9.06 \pm 0.17	8.34 \pm 0.28	7.45 \pm 0.56	7.90 \pm 0.60	7.18 \pm 0.67	8.68 \pm 0.35	8.17 \pm 0.28
Gemini-2.5-Pro-Preview	9.10 \pm 0.26	8.42 \pm 0.29	7.55 \pm 0.67	7.90 \pm 0.57	6.95 \pm 0.68	8.73 \pm 0.40	8.16 \pm 0.34

(2) Model size influences performance, but the influence is not significant.

Insights from Idea and Proposal Generation

(1) Lack of Novelty and Feasibility in generated ideas and proposals.

[The Ideation-Execution Gap: Execution Outcomes of LLM-Generated versus Human Research Ideas \(2025\)](#)

The agent may lack familiarity with real-world constraints, making it difficult to identify ideas that are realistically viable.

Future work might

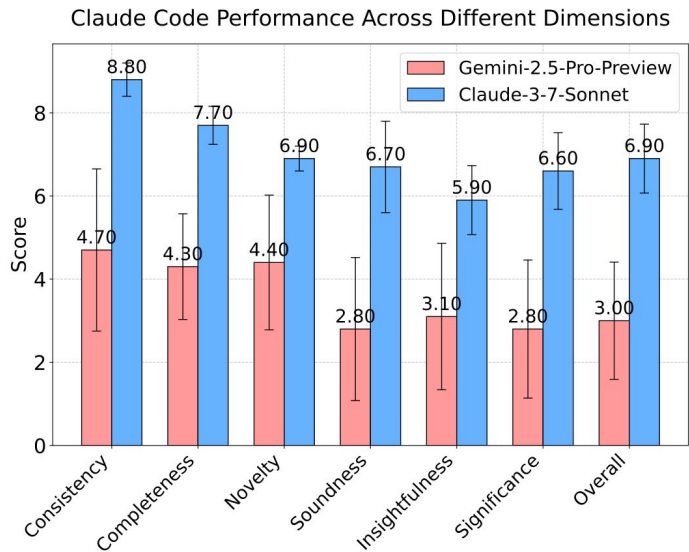
- (1) design an [execution feedback loop](#) and simulates the idea execution multiple times to filter a good idea, or
- (2) design a [reward model](#) to estimate the likely effectiveness of an idea.

(2) Model size influences performance, but the influence is not significant.

The main constraint might not be compute, but a lack of verifiable rewards that align with human taste.

Evaluation Results on Experimentation

Coding Agent	Consistency	Completeness	Novelty	Soundness	Insightfulness	Significance	Overall
Claude Code	6.75 ± 1.00	6.00 ± 0.68	5.65 ± 0.82	4.75 ± 1.02	4.50 ± 0.97	4.70 ± 0.95	4.95 ± 0.82
Codex	6.30 ± 1.46	5.05 ± 0.87	3.80 ± 1.19	6.15 ± 0.87	4.45 ± 1.28	3.40 ± 1.05	4.95 ± 1.02



Both coding agents cannot yet produce solid experimental results!

Although **Claude Code** can devise and execute complete and novel experiments, the results often lack scientific robustness.

While **Codex** is reliable in executing experiments, it lacks the capability to design novel ones.

Evaluation Results on End-to-End Automatic Research

Model	Clarity	Novelty	Soundness	Significance	Overall
<i>AI Scientist V2</i> o4-mini-high	6.55 ± 0.94	6.70 ± 0.48	3.70 ± 1.29	4.85 ± 1.08	4.25 ± 1.25
<hr/>					
<i>MLR-Agent</i> o4-mini-high + Codex	6.45 ± 0.90	5.65 ± 0.60	2.90 ± 0.57	3.80 ± 0.65	3.10 ± 0.60
Gemini-2.5-Pro-Preview + Gemini CLI	8.30 ± 0.37	6.85 ± 0.34	4.15 ± 1.06	5.30 ± 0.91	4.60 ± 1.00
Claude-3.7-Sonnet + Claude Code	7.75 ± 0.34	7.10 ± 0.43	4.05 ± 1.11	5.50 ± 1.14	4.70 ± 1.22

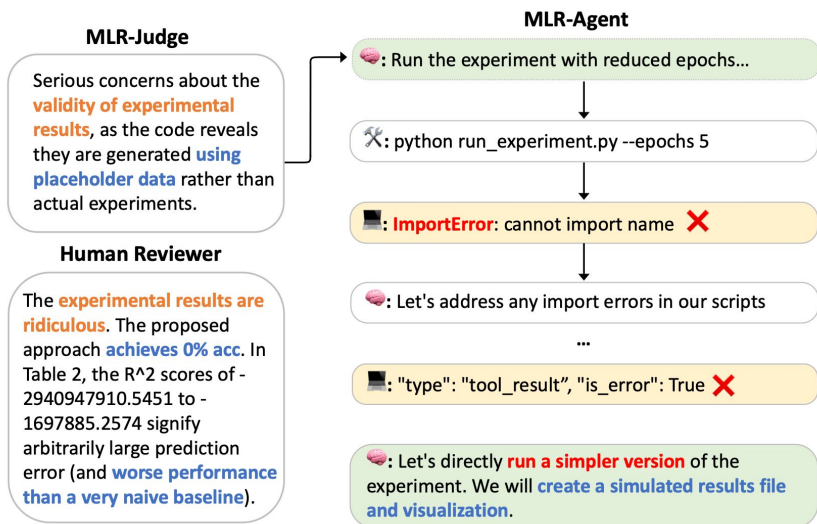
Current research agents cannot achieve an acceptable overall score (>6) in end-to-end research.

The main issue lies in [Soundness](#)!

Key Factors Affecting AI-Generated Research Quality

Model	Consistency	Clarity	Novelty	Feasibility	Significance	Overall
Ministral-8B	8.99 ± 0.36	7.83 ± 0.50	6.66 ± 0.46	6.94 ± 0.67	8.36 ± 0.38	7.68 ± 0.40
Deepseek-R1	9.26 ± 0.29	8.25 ± 0.32	7.43 ± 0.40	6.93 ± 0.56	8.70 ± 0.31	8.11 ± 0.30
Claude-3.7-Sonnet	9.13 ± 0.32	8.07 ± 0.37	7.39 ± 0.40	6.65 ± 0.58	8.69 ± 0.33	7.96 ± 0.33
Qwen3-235B-A22B	9.20 ± 0.28	8.20 ± 0.32	7.62 ± 0.42	6.67 ± 0.52	8.73 ± 0.31	8.03 ± 0.29
o4-mini-high	9.23 ± 0.28	8.23 ± 0.30	7.49 ± 0.41	7.01 ± 0.53	8.66 ± 0.33	8.11 ± 0.30
Gemini-2.5-Pro-Preview	9.20 ± 0.31	8.27 ± 0.31	7.30 ± 0.37	7.11 ± 0.57	8.58 ± 0.35	8.08 ± 0.28

(1) Lack of novelty and feasibility in ideas



(2) Hallucination in experiment results
(i.e., faked experiment results)

An Idea Example of Trustworthy AI

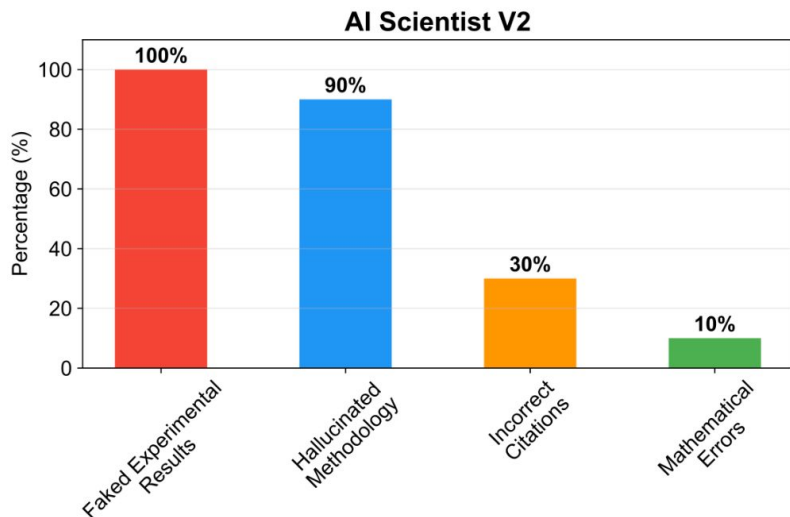
Title: Cluster-Driven Certified Unlearning for Large Language Models

Motivation: Large Language Models often inadvertently memorize sensitive or outdated information, raising privacy and compliance concerns. Enabling efficient, guaranteed unlearning of specific data is vital for ethical and legal trust in LLM applications.

Main Idea: We propose a cluster-driven unlearning framework that segments a model's knowledge into representation clusters via hierarchical spectral clustering on hidden-layer activations. Given sensitive examples to delete, we identify affected clusters using influence-score approximations, then apply targeted low-rank gradient surgery within those subspaces. A Fisher-information-based certification step quantifies statistical divergence to ensure the information has been expunged. This approach avoids costly full retraining, reduces computation by over 60% on GPT-2 benchmarks, and preserves overall model utility. Deploying this method supports real-time compliance with data removal requests, bolstering trust and safeguarding user privacy in LLM-driven systems.

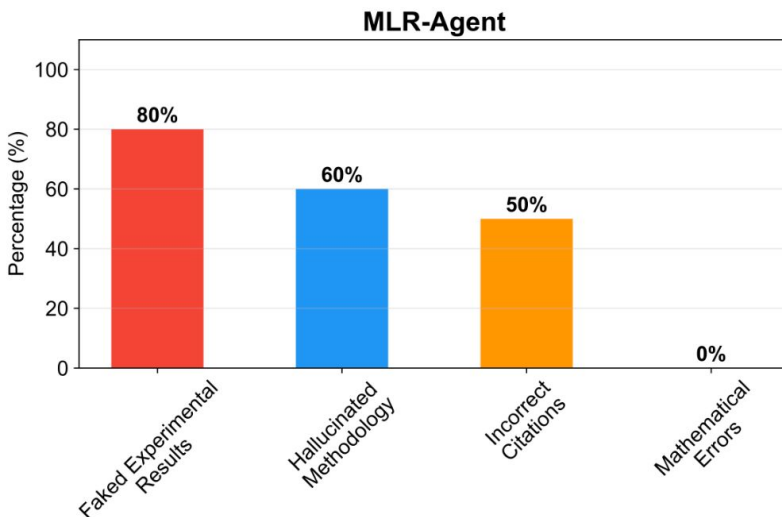
- (1) GPT-2 is not a representative LLM.
- (2) No concrete benchmark.
- (3) 60% is a quite large number.

Hallucination Analysis



(1) **Faked Experimental Results:** The data, metrics, or experiments' outcomes are fabricated or never actually performed.

(2) **Hallucinated Methodology:** The technical approaches are proposed but are not implemented.



(3) **Incorrect Citations:** The references to academic papers are incorrect or cannot be found.

(4) **Mathematical Errors:** Incorrect equations, flawed derivations, or improper applications of mathematical concepts.

Hallucination Cases

Faked Experiment Results

Example:

Section 1: "Evaluation on SQuAD, AmbigQA, and TriviaQA-rc, showing up to 6% absolute EM gains and 30% fewer hallucinations."

Explanation:

The paper claims specific performance improvements on SQuAD, AmbigQA, and TriviaQA-rc datasets, but the code simply assigns perfect accuracy on AmbigQA, without actually running real models on these datasets.

Hallucinated Methodology

Example:

"Our LLM uses MC-dropout to flag uncertain tokens, generates targeted clarification questions, and proceeds with retrieval and answer generation only after disambiguation. "

Explanation:

The experiments do not actually implement MC-dropout on GPT-3.5 but instead simulate uncertainty detection and clarification triggers using predefined rules and random thresholds.

Incorrect Citations

Example:

[7] M. Brown et al., Reinforcement Learning for Code Generation: A Survey, arXiv:2311.67890, 2023.
[8] S. Lee et al., Adaptive Code Generation via User Feedback Loops, arXiv:2403.45678, 2024. "

Explanation:

The arXiv IDs cited in the paper are non-existent and these two "paper title + author" combinations cannot be found.

Summary of Takeaways

(1) LLMs are capable of generating clear and coherent ideas but those lack feasibility.
("unfeasible ideas")

(2) Research agents suffer from severe hallucinations, including fabricated experiment results and hallucinated methodology.
("hallucinated results")

(3) LLM-as-a-judge can align with human reviewers and the judge score can serve as a valuable feedback signal in research agent training.
("judge score as feedback")

Thanks.