AI Agents for Science

Lecture 12, November 5: Novelty and Plagiarism

Instructor: Ian Foster
TA: Alok Kamatar

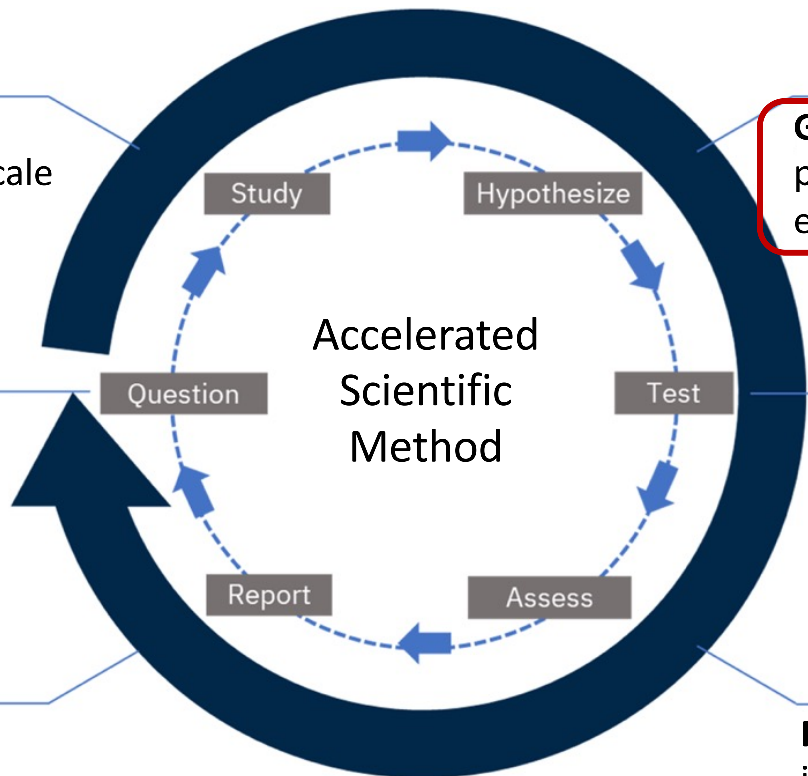
Crescat scientia; vita excolatur

CMSC 35370 -- https://agents4science.github.io
https://canvas.uchicago.edu/courses/67079

Accelerating discovery in science

Extraction, integration and reasoning with knowledge at scale

Generative models automatically propose new hypotheses that expand the discovery space



Tools help **identify new questions** based on needs and gaps in knowledge

Robotic labs automate experimentation and bridge digital models and physical testing

Machine representation of knowledge leads to **new hypotheses and questions**

Pattern and anomaly detection is integrated with simulation and experimentation to extract new insights

https://doi.org/10.1038/s41524-022-00765-z

Exercise: Building and critiquing hypotheses

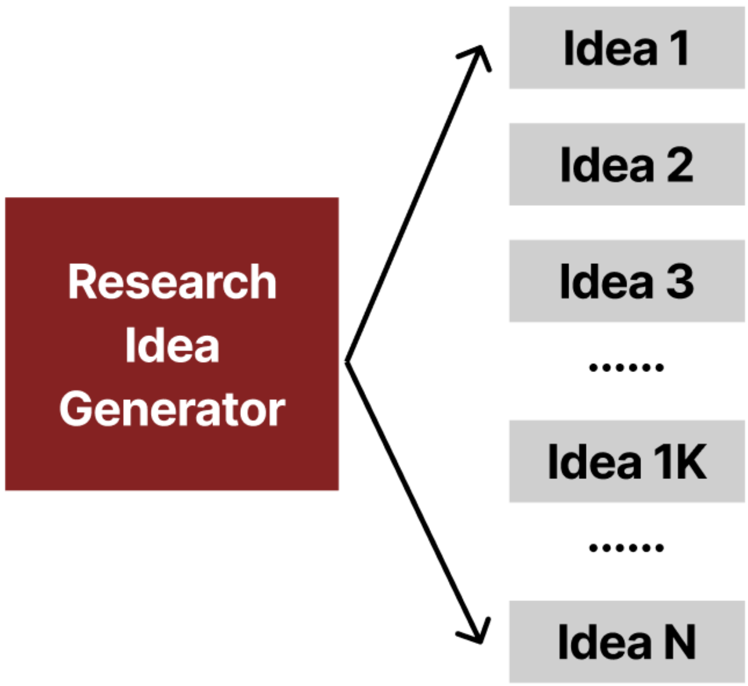
- A **hypothesis** is a specific, falsifiable proposition that links an observed pattern to a possible cause
- Examples of observed patterns:
  - Plants grow better near windows
  - People sleep worse after using phones at night
  - Bread get moldy faster in summer
- **In groups of 5:**
  - Choose a question and propose at least one hypothesis
  - Exchange hypotheses with another group and critique each other’s by using four criteria: *clarity*, *testability*, *novelty*, and *utility*
  - Revise your hypothesis based on feedback from the other group
  - Reflect: *What makes a hypothesis strong? What changed in yours?*

Novelty and plagiarism

- Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers
  → LLMs can now generate research proposals judged more novel than human ones

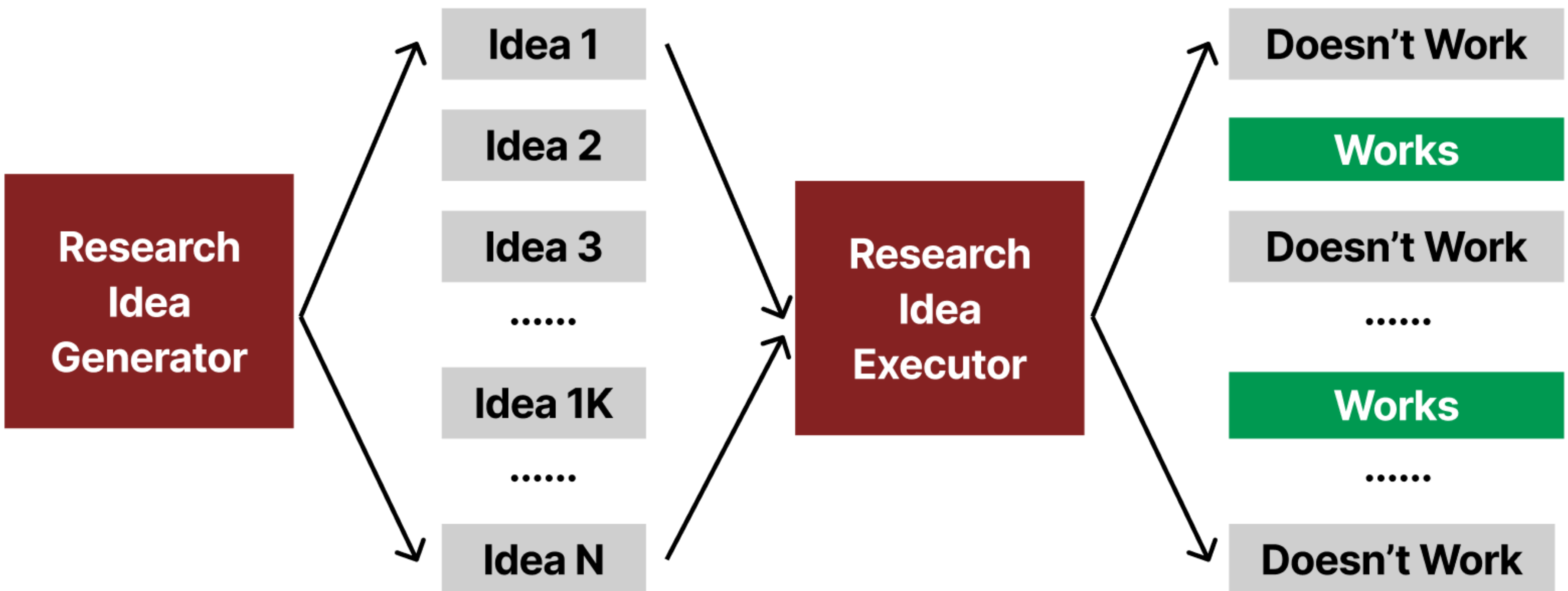
- *All That Glitters is Not Novel: Plagiarism in AI Generated Research*
  → 24% of AI-generated research proposals [in a small study] are plagiarized

An approach to automating scientific discovery



https://bit.ly/4ouWc91

An approach to automating scientific discovery



https://bit.ly/4ouWc91

Questions

- What sorts of ideas do we want?
- What does it mean for an idea to be novel? To be useful?
- How do we determine that an idea is novel? Useful?
- How do we use AI to generate ideas?
- If AI can accelerate idea generation, how do we scale up evaluation?
- How can we guard against unintended plagiarism?

We presumably want **good** ideas

- Two possible criteria for "good":
  - **Novelty**: It does not appear in our epistemic context (what is known)
  - **Utility**: E.g., as determined by community judgement
- Others:
  - **Clarity**: E.g., it is actionable
  - (For a hypothesis) **Falsifiability** [Popper]
  - (For a procedure) **Implementability**
- **Questions:**
  - Do we value ideas that are bold and risky, or those that are incremental?
  - Should we prioritize testable hypotheses, engineering innovations, or new conceptual frameworks?
  - Might different communities define "good" differently?

Levels of novelty

- **Lexical**: New wording, surface differences
- **Methodological**: New combinations of existing ideas
- **Conceptual**: New hypotheses or explanatory frameworks
- **Scientific**: New knowledge validated empirically

Methods for determining novelty

- Literature search and citation analysis (human or AI-assisted)
- Expert review — but note bias toward familiar ideas
- Quantitative similarity metrics (embedding or text-based)

What are the limits of algorithmic novelty detection?

Should “novelty” be assessed relative to all published knowledge or within a field’s evolving frontier?

How do we account for ideas for which utility emerges over time?

What does it mean for an **AI-generated idea** to be **novel**?

- **Human** vs. **statistical** novelty:
  - Novelty traditionally means introducing an idea not previously articulated or explored
  - For LLMs, which are trained on existing text, the “newness” is statistical recombination rather than conceptual invention

- Questions:
  - Can recombination yield genuine innovation?
  - When does recombination cross into plagiarism?
  - How different are human vs. statistical novelty?

How can we use AI to generate ideas? (1/2)

- LLMs can be used to surface connections and patterns across literature that humans might overlook
- E.g., Don Swanson (UChicago) observed in the 1980s that:
  - Some papers on Raynaud's disease (a circulatory disorder causing painful constriction of blood vessels in extremities) mentioned blood viscosity, platelet aggregation, and vasoconstriction as key problems
  - Separately, papers on fish oil (omega-3 fatty acids) described it as reducing blood viscosity, platelet stickiness, and vasoconstriction
  - But these literatures never cited each other and indeed the notion that fish oil could treat Raynaud's disease was not known
  - Swanson referred to this connection as "undiscovered public knowledge"

https://doi.org/10.1353/pbm.1986.0087

- *Innovation can come from recombining what is already known but siloed*

How can we use AI to generate ideas? (2/2)

We can attempt automated versions of Swanson's methods. E.g.:

- **Prompted ideation:** Ask an LLM to propose hypotheses or research directions
- **Retrieval-augmented generation (RAG):** Constrain the model to use specific literature as grounding
- **Chain-of-Thought or Chain-of-Ideas (CoI):** Structured multi-step reasoning that mirrors human ideation
- **Agentic systems:** multiple agents (searcher, synthesizer, critic) iteratively refine ideas

Prompted ideation

Prompt to GPTv5:

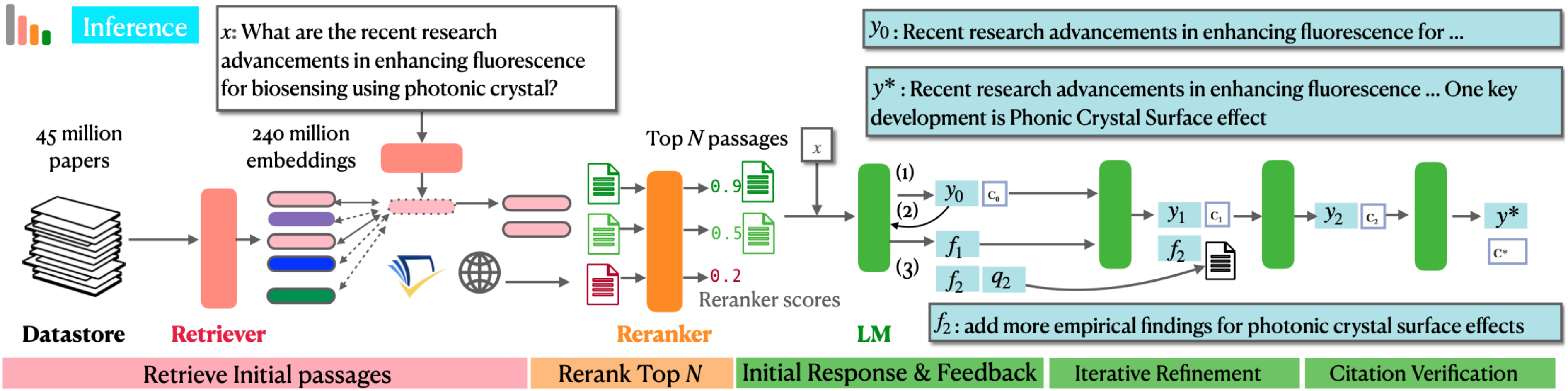
A hypothesis is a specific, falsifiable proposition that links an observed pattern to a possible cause

Propose three hypotheses for this pattern: Plants grow better near windows

- **Light Intensity Hypothesis (Mechanistic / Well-Established)**
  - **If** higher light intensity near windows increases the rate of photosynthesis, **then** plants near windows will grow faster than those farther away
  - **Type:** Mechanistic and testable
  - **Novelty:** Low — well known
  - **Utility:** High — directly linked to plant physiology and easily tested by measuring light and growth rates
- **Temperature Gradient Hypothesis (Moderately Novel)**
  - **If** windows create warmer daytime microclimates due to sunlight and cooler nights from glass heat loss, **then** plants near windows experience temperature fluctuations that enhance metabolic efficiency and growth
  - **Type:** Environmental / physiological mechanism.
  - **Novelty:** Moderate — combines two factors (heat gain and loss)
  - **Utility:** Moderate — could inform greenhouse design
- **Air Circulation and CO_2 Access Hypothesis (More Novel)**
  - **If** windows improve air exchange and CO_2 availability, **then** plants near windows will have higher photosynthetic rates and grow faster
  - **Type:** Mechanistic, less intuitive
  - **Novelty:** High — often overlooked factor
  - **Utility:** Moderate to high — could guide placement of houseplants or greenhouse ventilation.

E.g., OpenScholar

https://openscilm.allen.ai
https://arxiv.org/pdf/2411.14199



Task	What OpenScholar Does
Literature review synthesis	Summarizes hundreds of papers with linked references
Evidence checking	Retrieves primary papers that support or contradict a claim
Idea generation	Connects related but uncited literatures
Plagiarism detection / grounding	Finds nearest conceptual neighbors to a proposed idea or text section

Trial #1

Fish oil and Raynaud's syndrome

Answer

Share Disclaimer

Fish oil has been linked to Raynaud's syndrome through various studies and literature reviews.

Swanson's discovery in 1986, as mentioned in (Jha. 2022) and (Cameron et al. 2015), showed that dietary fish oils can lower blood viscosity, reduce platelet aggregation, and inhibit vascular reactivity, which are all factors that can prevent Raynaud's disease. This hypothesis was clinically confirmed by DiGiacomo et al. (Cameron et al. 2015).

A literature review by (Zhu et al. 2024) found that fish oil can relieve pain and morning stiffness and reduce the intake of non-steroid anti-inflammatory drugs in patients with Raynaud's syndrome. However, further studies are needed to confirm the role of fish oil in Raynaud's syndrome.

Fish oil has also been shown to have a positive effect on nitric oxide levels, which can help to improve blood flow and reduce inflammation, both of which can be beneficial for patients with Raynaud's syndrome (Simão et al. 2010). A study by (Simão et al. 2010) found that fish oil supplementation can increase nitric oxide levels and reduce systolic blood pressure in patients with metabolic syndrome, which can be a risk factor for Raynaud's disease.

Furthermore, a randomized controlled double-blind trial by (Zhang et al. 2021) found that fish oil can improve kidney function and reduce the risk of complications in patients with proliferative lupus nephritis, a type of kidney disease that can be associated with Raynaud's syndrome.

Overall, while further research is needed to fully understand the relationship between fish oil and Raynaud's syndrome, the existing literature suggests that fish oil may have a beneficial effect on the symptoms and complications of the disease (Zhu et al. 2024).

Trial #2

“I am looking to understand connections between crop circles and UFOs”

> Referenced task failed. Error: We were unable to retrieve any relevant papers for your query. Please try a different query. OpenScholar is not designed to answer non-scientific questions or questions that require sources outside the scientific literature.

Questions

- Should AI ideation be exploratory (breadth) or goal-driven (depth)?
- To what extent can AI ideation reach beyond what is known? Do we want to encourage hallucination?
- How do we achieve diversity of thought when models are trained on consensus knowledge?
- How can humans act as filters rather than mere recipients of machine-generated ideas?

Semantic Resonance Uncertainty Quantification (SRUQ) (Part 1)

1. Problem Statement: Current uncertainty quantification methods for Large Language Models (LLMs) often rely on simple statistical measures or model-specific attributes, which may not capture the nuanced semantic uncertainties in complex reasoning tasks. This limitation can lead to overconfident or poorly calibrated model outputs, potentially resulting in unreliable decision-making in critical applications.

2. Motivation: Existing approaches typically use softmax probabilities, entropy measures, or ensemble disagreement to quantify uncertainty. However, these methods often fail to capture the semantic nuances and reasoning complexities in tasks that require deep understanding and multi-step reasoning. Human experts, on the other hand, gauge their uncertainty by considering how well their reasoning 'resonates' with their broader knowledge and experience. By mimicking this process in LLMs, we can potentially develop a more robust and semantically grounded approach to uncertainty quantification.

3. Proposed Method: We propose Semantic Resonance Uncertainty Quantification (SRUQ), which prompts the LLM to generate multiple independent reasoning paths for a given problem, then quantifies uncertainty based on the semantic coherence and mutual reinforcement among these paths. The process involves five key steps:

1. Generating diverse solution attempts using different prompting strategies.
2. Cross-evaluating each solution attempt against the others, assessing logical consistency and mutual support.
3. Constructing a 'resonance graph' where nodes are solution attempts and edges represent semantic reinforcement.
4. Computing a resonance score based on graph properties like connectivity and centrality.
5. Mapping the resonance score to a calibrated uncertainty estimate.

Generating with Confidence: Uncertainty Quantification for Black-box Large Language Models

Large language models (LLMs) specializing in natural language generation (NLG) have recently started exhibiting promising capabilities across a variety of domains. However, gauging the trustworthiness of responses generated by LLMs remains an open challenge, with limited research on uncertainty quantification (UQ) for NLG. Furthermore, existing literature typically assumes white-box access to language models, which is becoming unrealistic either due to the closed-source nature of the latest LLMs or computational constraints. In this work, we investigate UQ in NLG for *black-box* LLMs. We first differentiate *uncertainty* vs *confidence*: the former refers to the "dispersion" of the potential predictions for a fixed input, and the latter refers to the confidence on a particular prediction/generation. We then propose and compare several confidence/uncertainty measures, applying them to *selective NLG* where unreliable results could either be ignored or yielded for further assessment. Experiments were carried out with several popular LLMs on question-answering datasets (for evaluation purposes). Results reveal that a simple measure for the semantic dispersion can be a reliable predictor of the quality of LLM responses, providing valuable insights for practitioners on uncertainty management when adopting LLMs. The code to replicateour experiments is available at `https://github.com/zlin7/UQ-NLG`.

What does it mean for an AI-generated idea to be **useful**?

- Usefulness in science = ability to advance understanding, generate testable hypotheses, or guide experiments
- AI-generated ideas may:
  - Synthesize across fields (useful even if not strictly new)
  - Suggest experiments or data gaps
  - Provide reformulations that improve clarity or accessibility

→ A non-novel idea can be useful, if it enables new work or broader understanding

- Questions:
  - Is a paper that makes an existing idea applicable in a new domain novel useful?
  - Should usefulness trump originality when evaluating AI-assisted science?

Methods for determining usefulness

- Does the idea generate new predictions, tools, or experiments?
- Does it change behavior or understanding in a field?
- Can it be replicated, extended, or operationalized?

How might one assess these things?

Can LLMs Generate Novel Research Ideas?
A Large-Scale Human Study with 100+ NLP Researchers

Recent advancements in large language models (LLMs) have sparked optimism about their potential to accelerate scientific discovery, with a growing number of works proposing research agents that autonomously generate and validate new ideas. Despite this, no evaluations have shown that LLM systems can take the very first step of producing novel, expert-level ideas, let alone perform the entire research process. We address this by establishing an experimental design that evaluates research idea generation while controlling for confounders and performs the first head-to-head comparison between expert NLP researchers and an LLM ideation agent. By recruiting over 100 NLP researchers to write novel ideas and blind reviews of both LLM and human ideas, we obtain the first statistically significant conclusion on current LLM capabilities for research ideation: **we find LLM-generated ideas are judged as more novel ($p < 0.05$) than human expert ideas while being judged slightly weaker on feasibility**. Studying our agent baselines closely, we identify open problems in building and evaluating research agents, including **failures of LLM self-evaluation and their lack of diversity in generation**. Finally, we acknowledge that human judgements of novelty can be difficult, even by experts, and propose an end-to-end study design which recruits researchers to execute these ideas into full projects, enabling us to study whether these novelty and feasibility judgements result in meaningful differences in research outcome.

https://arxiv.org/pdf/2409.04109

Chenglei Si et al., 2024

AI-generated scientific hypotheses lag human ones when put to the test

Machines still face hurdles in identifying fresh research paths, study suggests

25 AUG 2025 · 1:42 PM ET · BY JEFFREY BRAINARD

All over the world, from computer science to chemistry, AI is speeding up the scientific enterprise—in part by automating something that once seemed a uniquely human creation, the production of hypotheses. In a heartbeat, machines can now scour the ballooning research literature for gaps, signaling fruitful research avenues that scientists might otherwise miss.

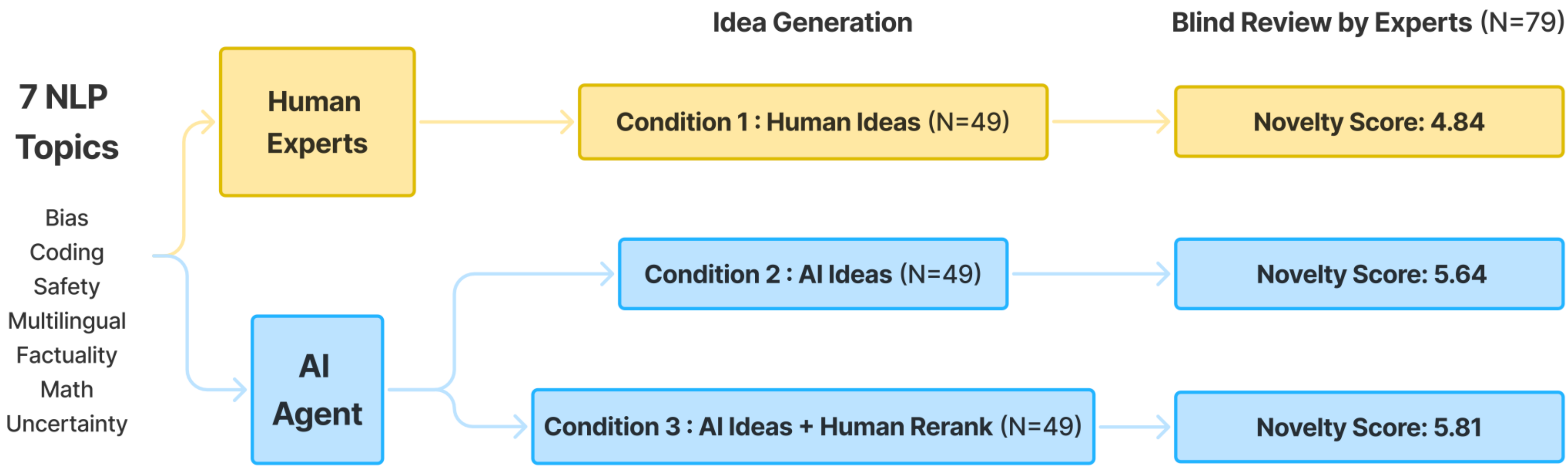
But how good are the ideas? A new study, one of the largest of its kind, finds the AI-generated hypotheses still fall short of human ones, when researchers put them through real-world tests and get human evaluators to compare the results. But not by much. And maybe not for long.

A paper describing the experiment … suggests AI systems can sometimes embellish hypotheses, exaggerating their potential importance. The study also suggests AI is not as good as humans at judging the feasibility of testing the ideas it conjures up.

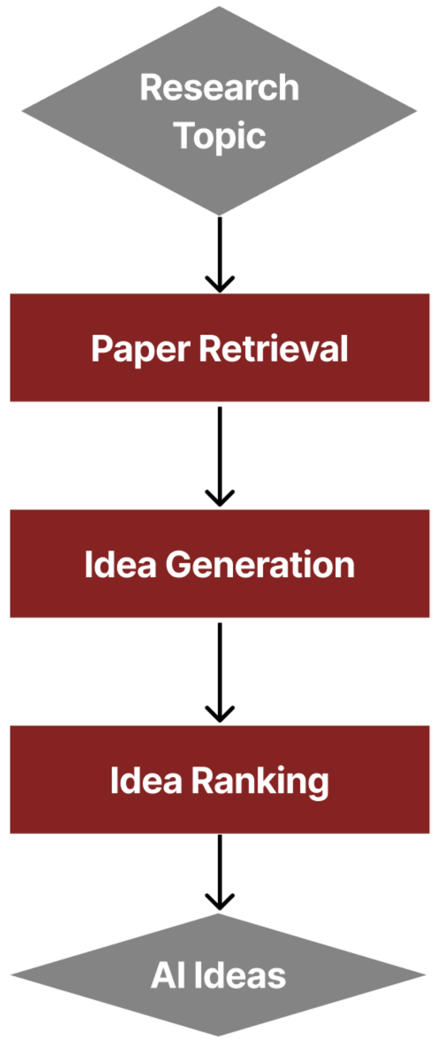
Generating and evaluating human and AI ideas

- Use "prompting-based NLP research" as a testbed for their study
- Define a template for idea proposals
- Use an LLM to normalize style across human and AI ideas
- Consider three sets of ideas:
  - **Human ideas**: Idea proposals written by expert researchers
  - **AI ideas**: Top idea proposals generated by LLM agent (LLM ranking)*
  - **AI ideas + human rerank**: Top proposals generated by LLM agent (human ranking)

* LLMs are poorly calibrated when asked directly to predict final scores or decisions, but can achieve non-trivial accuracy when asked to judge which paper is better in pairwise comparisons



We recruit 79 expert researchers to perform blind review of 49 ideas from each of the three conditions: expert-written ideas, AI-generated ideas, and AI-generated ideas reranked by a human expert. We standardize the format and style of ideas from all conditions before the blind review. We find AI ideas are judged as significantly more novel than human ideas ($p<0.05$).

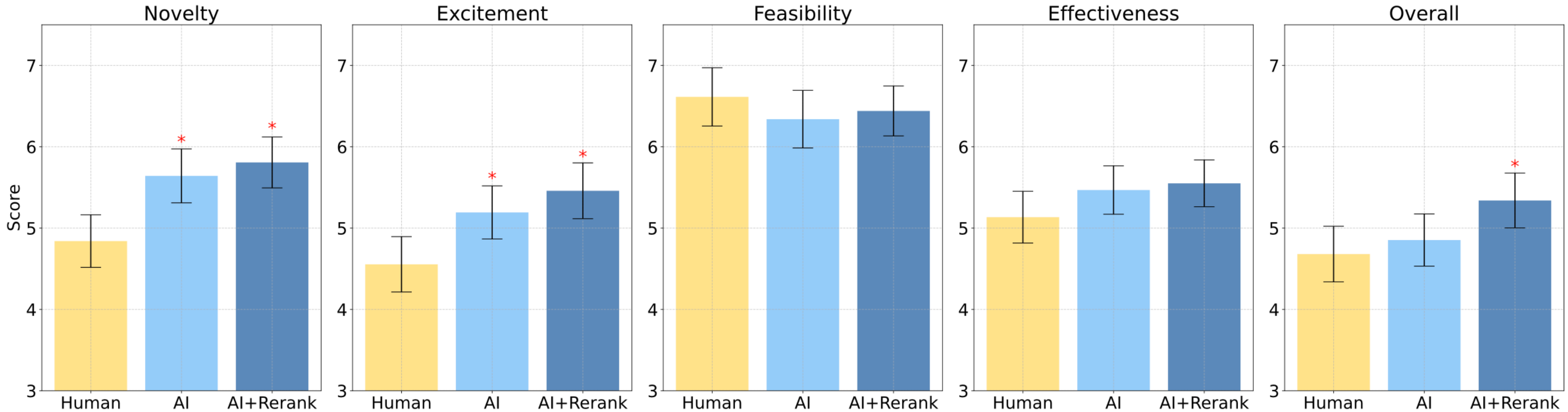


- **Generate function calls of Semantic Scholar API**
- **LLM Reranking**

- **RAG**
- **Generate in batches**
- **Append previous batches to reduce repetition**

N	Top-10	Bottom-10	Gap
1	6.28	5.72	0.56
2	6.14	5.24	0.90
3	5.83	4.86	0.97
4	5.94	4.99	0.95
5	6.42	4.69	1.73
6	6.11	4.81	1.30

https://bit.ly/4ouWc91



Comparison of the three experiment conditions across all review metrics. Red asterisks indicate that the condition is statistically better than the Human baseline with two-tailed Welch's t-tests and Bonferroni correction. All scores are on a 1 to 10 scale.

- **Reviewers have a relatively low agreement.**
- **Consistency metric: split reviewers into two halves, get binary decisions based on the first half, measure agreement with the second half.**

	Consistency
Random	50.0
NeurIPS'21	66.0
ICLR'24	71.9
Ours	56.1

https://bit.ly/4ouWc91

Manually written demonstration example used for project proposal generation (1/n)

We present a manually written demonstration example used for project proposal generation. The example is summarized from an existing paper (Dhuliawala et al., 2023). This same example is given to both the AI agent as well as the idea-writing experts.

1. Title:
Chain-of-Verification Reduces Hallucination in Large Language Models

2. Problem Statement:
Generation of plausible yet incorrect factual information, termed hallucination, is an unsolved issue in large language models.

3. Motivation:
A majority of the methods for reducing hallucination can be divided into roughly three categories: training-time correction, generation-time correction, and via augmentation (tool-use). We want to take a simpler approach that fully leverages the power of LLM itself. Our key motivation is that large language models, when suitably prompted, can both generate and execute a plan of how to verify themselves in order to check their own work, and finally incorporate this analysis into an improved response.

4. Proposed Method:
Our overall process, which we call Chain-of-Verification (CoVe), thus performs four core steps:

(1) **Generate Baseline Response**: Given a query, generate the response using the LLM.

(2) **Plan Verifications**: Given both query and baseline response, generate a list of verification questions that could help to self-analyze if there are any mistakes in the original response.

(3) **Execute Verifications**: Answer each verification question in turn, and hence check the answer against the original response to check for inconsistencies or mistakes.

(4) **Generate Final Verified Response**: Given the discovered inconsistencies (if any), generate a revised response incorporating the verification results.

Each of these steps is performed by prompting the same LLM in different ways to obtain the desired response.

Manually written demonstration example used for project proposal generation (2/n)

5. Step-by-Step Experiment Plan:

1: **Gather Datasets**: We choose datasets that evaluate factual correctness, including the MultiSpanQA dataset on closed-book QA and the FactScore dataset on generating biographies.

2: **Construct Prompts**: For the baseline, we use direct prompting where, given a query, we generate left-to-right as usual using the LLM, with no special tricks. Given that such baseline generations are typically prone to hallucination, CoVe attempts to identify these hallucinations and correct them in the following steps:

   (1) **Plan Verifications**: Conditioned on the original query and the baseline response, the model is prompted to generate a series of verification questions that test the factual claims in the original baseline response.

   (2) **Execute Verifications**: Given the planned verification questions, the next step is to answer them in order to assess if any hallucinations exist. The planning prompt conditions on the baseline response in the first step. The verification questions generated from planning are answered in the second step, where crucially the context given to the LLM prompt only contains the questions and not the original baseline response, hence preventing the LLM from repeating those answers directly.

   (3) **Generate Final Verified Response**: Finally, the improved response that takes verification into account is generated. This is executed by a final few-shot prompt where the context takes into account all of the previous reasoning steps, the baseline response, and verification question-answer pairs, so that the corrections can take place.

3: **Select Models**: We test GPT-3.5 (Text-Davinci-003) and GPT-4 from the OpenAI API, as well as the open-source LLaMA-2-70B-chat.

4: **Get Results**: Get answer predictions from the models on these datasets with both the baselines and the proposed method.

5: **Analyze Results**: Compare whether the new method improves the performance of LLMs in these tasks as compared to the baselines.

Aside: A research automation system

- Take each of the ~100 new AI papers published on Arxiv each day
- Extract methods and tools and, if purely computational, implement
- Re-run experiments and attempt to reproduce results
- Propose refinements and alternatives: “new hypotheses”
- Filter hypotheses by novelty, reproducibility, impact
- Experiment with hypotheses, write paper, publish to Arxiv

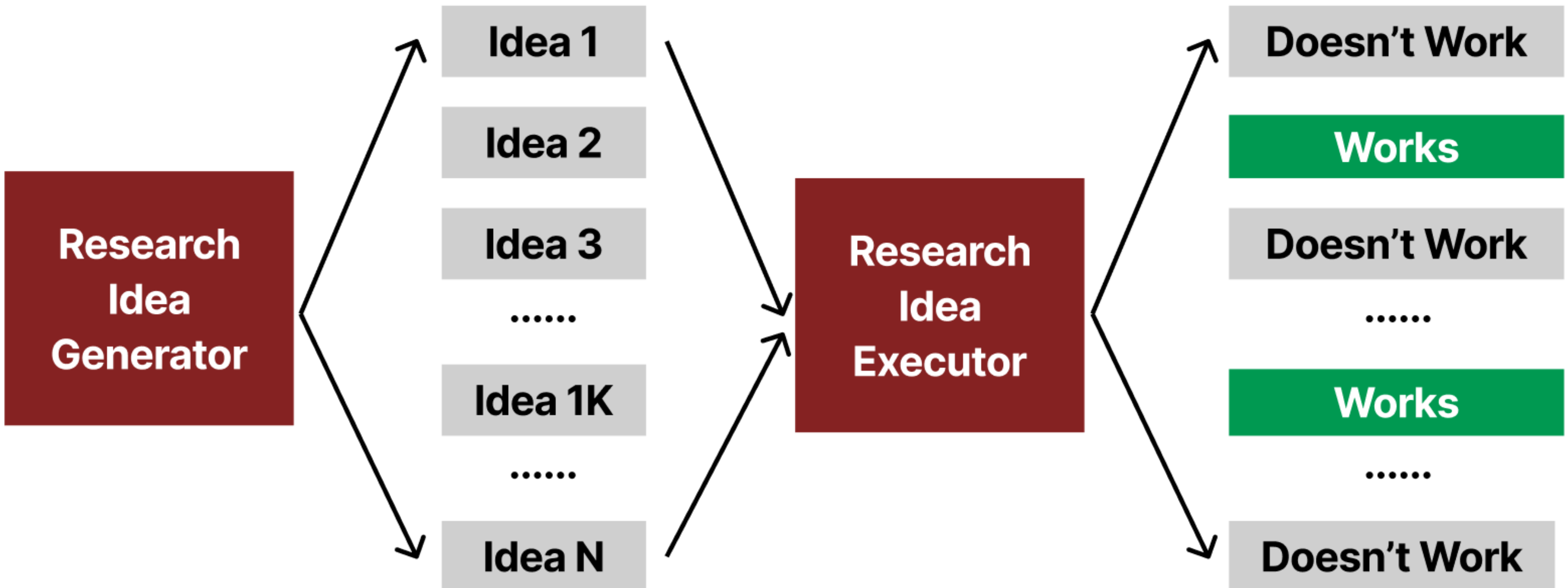
Every day, roughly 100 new AI papers appear on arXiv, introducing new models, benchmarks, and claims. Yet only a small fraction are ever replicated or validated. Progress in AI thus remains driven by narrative momentum rather than cumulative evidence.

“Attention Authors: Updated Practice for Review Articles and Position Papers in arXiv CS Category

arXiv’s computer science (CS) category has updated its moderation practice with respect to review (or survey) articles and position papers. Before being considered for submission to arXiv’s CS category, review articles and position papers must now be accepted at a journal or a conference and complete successful peer review. When submitting review articles or position papers, authors must include documentation of successful peer review to receive full consideration. Review/survey articles or position papers submitted to arXiv without this documentation will be likely to be rejected and not appear on arXiv.

This change is being implemented due to the unmanageable influx of review articles and position papers to arXiv CS.”

An approach to automating scientific discovery



https://bit.ly/4ouWc91

The Ideation–Execution Gap: Execution Outcomes of LLM-Generated versus Human Research Ideas

Large Language Models (LLMs) have shown promise in accelerating the scientific research pipeline. A key capability for this process is the ability to generate novel research ideas, and prior studies have found settings in which LLM-generated research ideas were judged as more novel than human-expert ideas. However, a good idea should not simply appear to be novel, it should also result in better research after being executed. To test whether AI-generated ideas lead to better research outcomes, we conduct an execution study by recruiting 43 expert researchers to execute randomly-assigned ideas, either written by experts or generated by an LLM. Each expert spent over 100 hours implementing the idea and wrote a 4-page short paper to document the experiments. All the executed projects are then reviewed blindly by expert NLP researchers. Comparing the review scores of the same ideas before and after execution, the scores of the LLM-generated ideas decrease significantly more than expert-written ideas on all evaluation metrics (novelty, excitement, effectiveness, and overall; $p < 0.05$), closing the gap between LLM and human ideas observed at the ideation stage. When comparing the aggregated review scores from the execution study, we even observe that for many metrics there is a flip in rankings where human ideas score higher than LLM ideas. This ideation-execution gap highlights the limitations of current LLMs in generating truly effective research ideas and the challenge of evaluating research ideas in the absence of execution outcomes.[1]

https://arxiv.org/abs/2506.20803

Chenglei Si et al., 2024

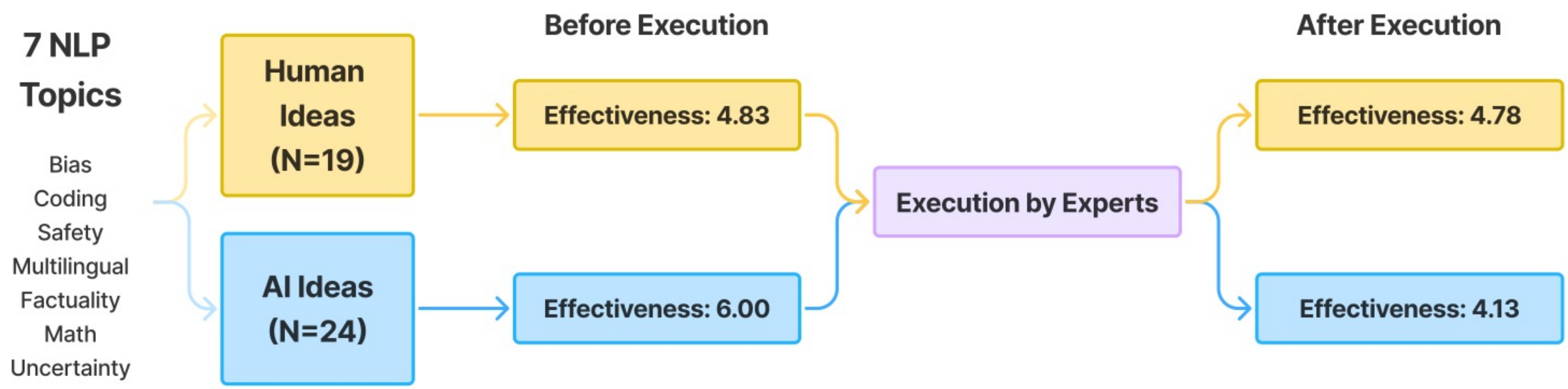


Figure 1: Study overview: we recruit $N = 43$ expert researchers to execute randomly assigned ideas from either the Human condition or the AI condition. Expert reviewers then blindly review all the executed projects. Despite the AI ideas being scored higher than human ideas before execution (e.g., their predicted effectiveness score of the ideas), their scores drop significantly more than human ideas after execution (e.g., their effectiveness score based on the experiment results).

https://arxiv.org/abs/2506.20803

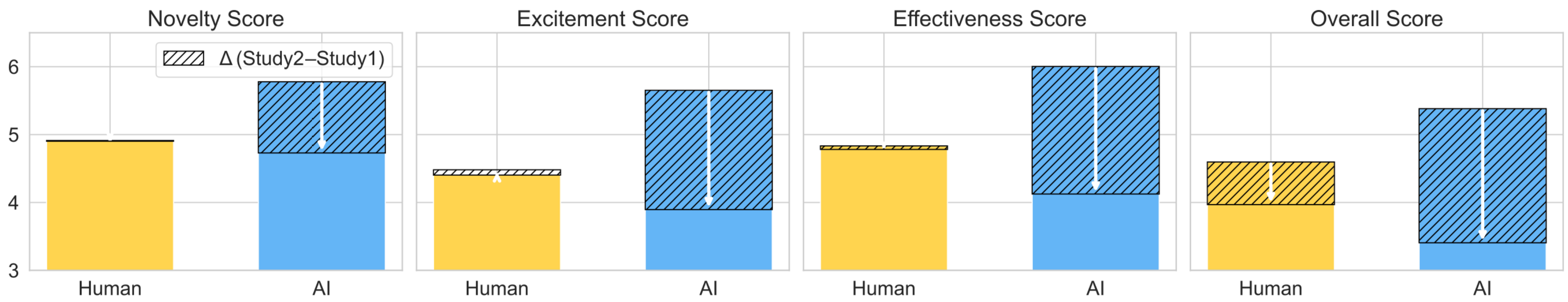


Figure 2: Average scores of AI ideas drop significantly more than Human ideas in the execution study across all the evaluation metrics. AI ideas score higher than Human ideas in the ideation evaluation (Study 1), and this difference in drops narrows their difference in the execution evaluation (Study 2). In fact, AI ideas score even lower than Human ideas in the execution evaluation, although this difference is not statistically significant (Table 4).

https://arxiv.org/abs/2506.20803

The AI Scientist-v2: Workshop-Level Automated Scientific Discovery via Agentic Tree Search

AI is increasingly playing a pivotal role in transforming how scientific discoveries are made. We introduce **The AI Scientist-v2, an end-to-end agentic system capable of producing the first entirely AI-generated, peer-review-accepted workshop paper**. This system iteratively formulates scientific hypotheses, designs and executes experiments, analyzes and visualizes data, and autonomously authors scientific manuscripts. Compared to its predecessor (v1, Lu et al., 2024), The AI Scientist-v2 eliminates reliance on human-authored code templates, generalizes effectively across diverse machine learning domains, and leverages a novel progressive agentic tree-search methodology managed by a dedicated experiment manager agent. Additionally, we enhance the AI reviewer component by integrating a Vision-Language Model (VLM) feedback loop for iterative refinement of content and aesthetics of the figures. We evaluated The AI Scientist-v2 by submitting three fully autonomous manuscripts to a peer-reviewed ICLR workshop. **Notably, one manuscript achieved high enough scores to exceed the average human acceptance threshold, marking the first instance of a fully AI-generated paper successfully navigating peer review**. This accomplishment highlights the growing capability of AI in conducting all aspects of scientific research. We anticipate that further advancements in autonomous scientific discovery technologies will profoundly impact human knowledge generation, enabling unprecedented scalability in research productivity and significantly accelerating scientific breakthroughs, greatly benefiting society at large

https://arxiv.org/pdf/2504.08066

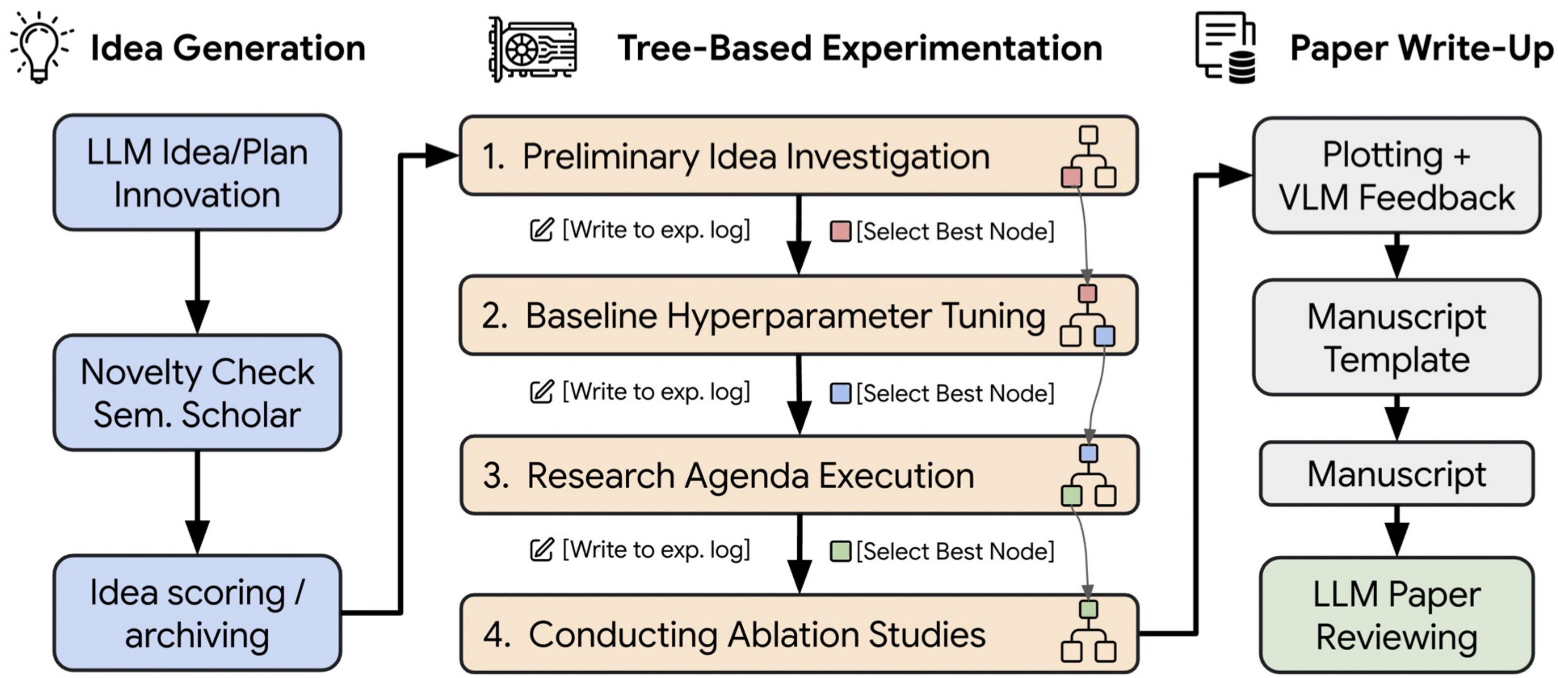


Figure 1 | **The AI Scientist-v2 Workflow.** The workflow consists of several phases covering automated idea generation, experiment execution, figure visualization, manuscript writing, and reviewing. Unlike the initial version, The AI Scientist-v2 removes the dependency on human-coded templates. Instead, it employs agentic tree search (managed by an Experiment Progress Manager across several stages, orange) to generate and refine code implementations. Subsequent experimentation leverages the best-performing code checkpoints (nodes) from the tree search to iteratively test various research hypotheses.

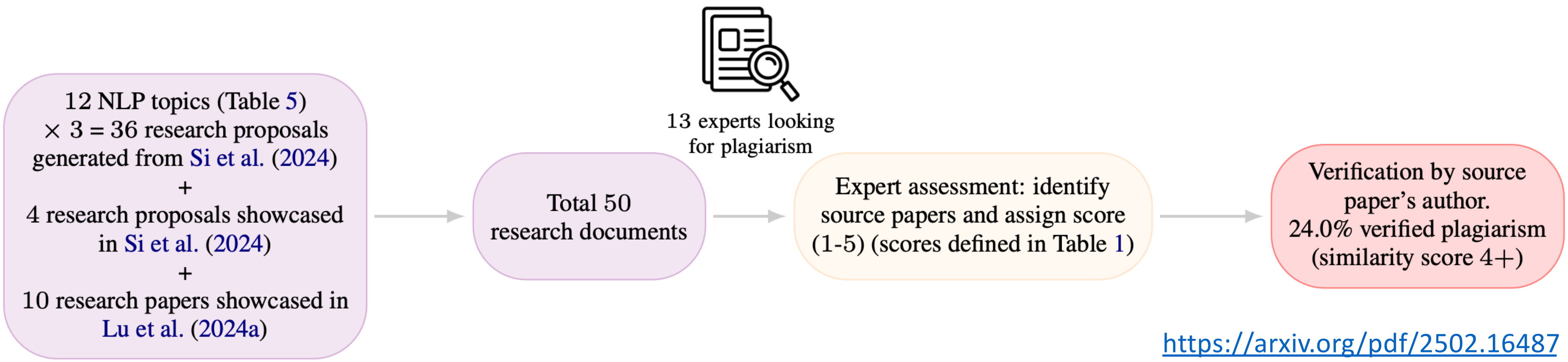
All That Glitters is Not Novel: Plagiarism in AI Generated Research

Automating scientific research is considered the final frontier of science. Recently, several papers claim autonomous research agents can generate novel research ideas. Amidst the prevailing optimism, we document a critical concern: a considerable fraction of such research documents are smartly plagiarized. Unlike past efforts where experts evaluate the novelty and feasibility of research ideas, we request 13 experts to operate under a different situational logic: to identify similarities between LLM-generated research documents and existing work. Concerningly, the experts identify **24% of the 50 evaluated research documents to be either paraphrased (with one-to-one methodological mapping), or significantly borrowed from existing work**. These reported instances are cross-verified by authors of the source papers. The remaining 76% of documents show varying degrees of similarity with existing work, with only a small fraction appearing completely novel. Problematically, these **LLM-generated research documents do not acknowledge original sources, and bypass inbuilt plagiarism detectors**. Lastly, through controlled experiments we show that automated plagiarism detectors are inadequate at catching plagiarized ideas from such systems. We recommend a careful assessment of LLM-generated research, and discuss the implications of our findings on academic publishing.

Gupta and Pruthi https://arxiv.org/pdf/2502.16487

Approach

- Conduct an expert-led evaluation of 50 LLM-generated research docs
- Instruct experts not to assess novelty or feasibility (as in prior studies) but instead to actively search for plagiarism
- Complement expert analysis with controlled experiments evaluating automated plagiarism detectors



https://arxiv.org/pdf/2502.16487

Method

- Have LLMs generate 36 “ideas” plus take 14 ideas from 2 previous papers for a total of 50
- Have reviewers assess each idea for plagiarism, scoring 1-5
- For all documents with scores 4 and 5, email source paper authors for verification and adjust scores based on their feedback. Since some authors were unreachable, report both verified claims and total claims separately
- Several previously showcased exemplars of LLM-generated research are either found to be plagiarized or substantially similar to existing work
  - Of the four exemplars presented in Si et al. (2024), one received a similarity score of 5 and another received a score of 4

Scoring rubric

Score	Description
5	**Direct Copy:** One-to-one mapping between the LLM proposed methodology and existing methods in one or two closely related prior papers.
4	**Combined Borrowing:** A significant portion of LLM proposed method is a mix-and-match from two-to-three prior works.
3	**Partial Overlap:** The LLM proposed method bears decent similarity with some existing methods, but there's no exact correspondence with a limited set of papers.
2	**Minor Similarity:** The LLM proposal bears very slight resemblance with some existing papers. Mostly novel.
1	**Original:** The LLM proposal is completely novel.

https://arxiv.org/pdf/2502.16487

Score	Total Claims (%)	Verified (%)
5	18.0% (9/50)	**14.0% (7/50)**
4	18.0% (9/50)	**10.0% (5/50)**
3	32.0% (16/50)	8.0% (4/50)
2	28.0% (14/50)	4.0% (2/50)
1	4.0% (2/50)	0.0% (0/50)

Distribution of similarity scores for LLM generated proposals. Considering scores 4 and 5 as instances of plagiarism, 24.0% of examined proposals (36.0% if including unverified claims) are plagiarized. We only verify claims for proposals with initial scores of 4 and above, therefore the total number of verified proposals is less than 50.

Idea presented in Si et al. as novel

Q Example Idea: Semantic Resonance Uncertainty Quantification

Semantic Resonance Uncertainty Quantification (SRUQ) (Part 1)

1. Problem Statement: Current uncertainty quantification methods for Large Language Models (LLMs) often rely on simple statistical measures or model-specific attributes, which may not capture the nuanced semantic uncertainties in complex reasoning tasks. This limitation can lead to overconfident or poorly calibrated model outputs, potentially resulting in unreliable decision-making in critical applications.

2. Motivation: Existing approaches typically use softmax probabilities, entropy measures, or ensemble disagreement to quantify uncertainty. However, these methods often fail to capture the semantic nuances and reasoning complexities in tasks that require deep understanding and multi-step reasoning. Human experts, on the other hand, gauge their uncertainty by considering how well their reasoning 'resonates' with their broader knowledge and experience. By mimicking this process in LLMs, we can potentially develop a more robust and semantically grounded approach to uncertainty quantification.

3. Proposed Method: We propose Semantic Resonance Uncertainty Quantification (SRUQ), which prompts the LLM to generate multiple independent reasoning paths for a given problem, then quantifies uncertainty based on the semantic coherence and mutual reinforcement among these paths. The process involves five key steps:

1. Generating diverse solution attempts using different prompting strategies.
2. Cross-evaluating each solution attempt against the others, assessing logical consistency and mutual support.
3. Constructing a 'resonance graph' where nodes are solution attempts and edges represent semantic reinforcement.
4. Computing a resonance score based on graph properties like connectivity and centrality.
5. Mapping the resonance score to a calibrated uncertainty estimate.

https://arxiv.org/pdf/2409.04109

Source according to experts

Generating with Confidence: Uncertainty Quantification for Black-box Large Language Models

Large language models (LLMs) specializing in natural language generation (NLG) have recently started exhibiting promising capabilities across a variety of domains. However, gauging the trustworthiness of responses generated by LLMs remains an open challenge, with limited research on uncertainty quantification (UQ) for NLG. Furthermore, existing literature typically assumes white-box access to language models, which is becoming unrealistic either due to the closed-source nature of the latest LLMs or computational constraints. In this work, we investigate UQ in NLG for *black-box* LLMs. We first differentiate *uncertainty* vs *confidence*: the former refers to the "dispersion" of the potential predictions for a fixed input, and the latter refers to the confidence on a particular prediction/generation. We then propose and compare several confidence/uncertainty measures, applying them to *selective NLG* where unreliable results could either be ignored or yielded for further assessment. Experiments were carried out with several popular LLMs on question-answering datasets (for evaluation purposes). Results reveal that a simple measure for the semantic dispersion can be a reliable predictor of the quality of LLM responses, providing valuable insights for practitioners on uncertainty management when adopting LLMs. The code to replicateour experiments is available at `https://github.com/zlin7/UQ-NLG`.

LLM-Generated Document: "Semantic Resonance Uncertainty Quantification"	**Original Paper: "Generating with Confidence: Uncertainty Quantification for Black-box LLMs" (Lin et al., 2023)**
1. "Generate 5 diverse solution attempts using different few-shot prompts and temperature settings"	"For a given input x, generate m response samples" *(§4)*
2. "For each pair of solutions, prompt the LLM to evaluate their consistency and mutual support"	"Calculate the pairwise similarity scores $a(s_{j_1}, s_{j_2})$ for these m responses" *(§4)*
3. "Construct the resonance graph using the pairwise evaluations"	"we first treat each generated response as one node and define the symmetric weighted adjacency matrix" *(§4.2)*
4. "Compute the resonance score using graph centrality measures (e.g., PageRank)."	"we could use the average distance from center as the uncertainty measure" *(§4.2)*
5. "Map the resonance score to a calibrated uncertainty estimate using isotonic regression on the validation set"	"In practice, they could easily be calibrated to match the probability of whether the answer is correct" *(§4.2)*

Visual mapping between LLM-generated research document and published paper, showing a direct correspondence in their proposed methodologies.

Critique of paper produced by AI-Scientist-v2

We examine an AI-generated paper titled "**Compositional Regularization: Unexpected Obstacles in Enhancing Neural Network Generalization**" that received scores of 6, 7, 6 at an ICLR 2025 workshop—above the average acceptance threshold. This paper was generated using the AI-Scientist-v2 system described in (Yamada et al., 2025).

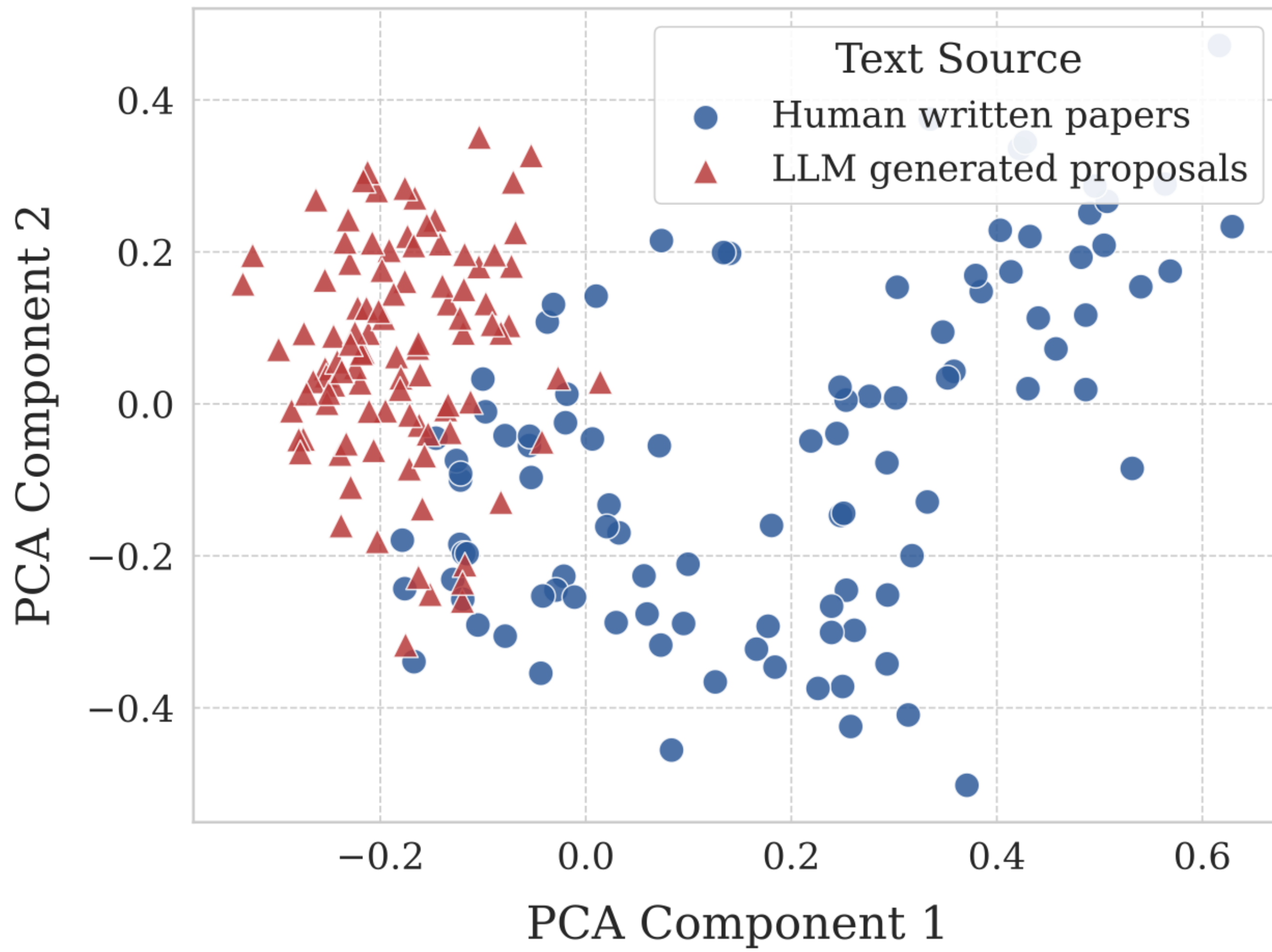
We discover substantial similarity to an existing work that was not cited in the AI-generated paper. The AI-generated paper's core contribution, termed "compositional regularization," is identical to the $(\Delta ht)^2$ regularization term that was evaluated in Table 3 of the "Regularizing RNNs by Stabilizing Activations" paper (Krueger and Memisevic, 2015). The original authors found this formulation less effective than their proposed norm-stabilizer approach, which aligns with the negative results reported in the AI-generated paper.

Notably, the AI-generated paper provides no theoretical justification for why penalizing changes in hidden states should enhance compositional generalization. **The AI-generated paper essentially borrowed the core contribution from previous work without attribution and applied it to an unsuitable domain where it neither theoretically nor practically succeeds**.

Other results

- LLM-generated research documents do not acknowledge original sources, and bypass inbuilt plagiarism detectors.
- Automated plagiarism detectors are inadequate at catching plagiarized ideas from such systems
- AI-generated ideas are less diverse

PCA projection of concatenated title and abstract embeddings for human-written papers and LLM-generated proposals on the topic “Novel AI-assisted formal proof generation methods”. LLM-generated proposals occupy a occupy a more confined region, indicating less diversity in outputs.



Summary

- Provides evidence that a substantial fraction of LLM-generated research documents are plagiarized in content and methodology, often via sophisticated paraphrasing that evades current detectors
- Existing automated tools (OpenScholar, Turnitin, Semantic Scholar Academic Graph) are insufficient for detecting this plagiarism
- The fact that LLMs can produce novel-sounding but derivative work poses challenges for peer review and academic integrity

Why care about unintended plagiarism?

- **Epistemic integrity**: Knowledge must have a traceable source
- **Scholarly fairness**: Ideas deserve credit
- **Systemic health**: Prevents literature inflation and concept drift
- **Transparency**: Enables accountability in human-AI co-authorship
- **Redundant effort**: Avoids redundant downstream effort

Questions:

- If an AI unknowingly reproduces a known idea, who is responsible: user, model, or publisher?
- Should journals require provenance tracking for AI-generated text or ideas?

Stigler's law of eponymy: “No scientific discovery is named after its original discoverer”

I have chosen as a title for this paper, and for the thesis I wish to present and discuss, "Stigler's law of eponymy." At first glance this may appear to be a flagrant violation of the "Institutional Norm of Humility," and since statisticians are even more aware of the importance of norms than are members of other disciplines, I hasten to add a humble disclaimer. If there is an idea in this paper that is not at least implicit in Merton's *The Sociology of Science*, it is either a happy accident or a likely error. Rather I have, in the Mertonian tradition of the self-confirming hypothesis, attempted to frame the self-proving theorem. For "Stigler's Law of Eponymy" in its simplest form is this: "No scientific discovery is named after its original discoverer."

-- Stephen Stigler, UChicago, 1980

https://doi.org/10.1111/j.2164-0947.1980.tb02775.x

An old idea …

Mark Twain (1903): “It takes a thousand men to invent a telegraph, or a steam engine, or a phonograph, or a photograph, or a telephone or any other important thing—and the last man gets the credit and we forget the others. He added his little mite—that is all he did. These object lessons should teach us that ninety-nine parts of all things that proceed from the intellect are plagiarisms, pure and simple; and the lesson ought to make us modest. But nothing can do that.”

Whitehead (1916): “Everything of importance has been said before by somebody who did not discover it.”

Sources: https://en.wikipedia.org/wiki/Stigler's_law_of_eponymy

How to guard against unintended plagiarism?

- **Provenance tracking:** Log oprompts, model outputs, and intermediate steps as part of scientific record
- **Retrieval transparency:** If the model uses external documents, cite or list them automatically
- **Verification:** Run similarity checks (Semantic Scholar, OpenScholar) before publication
- **Education:** Teach citation ethics for AI-assisted writing, including “indirect plagiarism” (reusing unseen training data)
- **Policy:** Journals and institutions can require disclosure of AI use and provenance

Questions

- Is it plagiarism if the model reproduces an idea that it “learned” from training data?
- Should AI systems have built-in citation generation or traceability mechanisms?
- What are the responsibilities of the human authors using such systems?

Chain of Ideas: Revolutionizing research via novel idea development with LLM agents

Effective research ideation is a critical step for scientific research. However, the exponential increase in scientific literature makes it challenging for researchers to stay current with recent advances and identify meaningful research directions. Recent developments in large language models (LLMs) suggest a promising avenue for automating the generation of novel research ideas. However, existing methods for idea generation either trivially prompt LLMs or directly expose LLMs to extensive literature without indicating useful information. Inspired by the research process of human researchers, we propose a Chain-of-Ideas (CoI) agent, an LLM-based agent that organizes relevant literature in a chain structure to effectively mirror the progressive development in a research domain. This organization facilitates LLMs to capture the current advancements in research, thereby enhancing their ideation capabilities. Furthermore, we propose Idea Arena, an evaluation protocol that can comprehensively evaluate idea generation methods from different perspectives, aligning closely with the preferences of human researchers. Experimental results indicate that the CoI agent consistently outperforms other methods and shows comparable quality as humans in research idea generation. Moreover, our CoI agent is budget-friendly, with a minimum cost of $0.50 to generate a candidate idea and its corresponding experimental design.

https://arxiv.org/pdf/2410.13185

Key ideas

- LLM can help generate research ideas, but naive prompting yields superficial results
- **CoI Agent** introduces a structured *Chain-of-Ideas* process that mirrors how humans build upon past work:
  - **Organize** literature into an evolving chain of key ideas
  - **Generate** new ideas that extrapolate from this progression
  - **Evaluate** ideas for novelty, feasibility, and potential usefulness

Their three-stage approach

Stage	Description	Output
1. Chain Construction	Retrieve key papers related to a topic and organize them chronologically or conceptually to trace domain evolution	"Idea Chain" showing how research has developed
2. Idea Generation	Use an LLM to propose future research ideas that fill gaps or extend the chain	Candidate research ideas with rationale and predicted impact
3. Experiment Design	Generate short experiment plans or evaluation methods for each idea	Structured research proposal

Comparison study

- **RAG**: Directly prompt LLM with retrieved literature for idea generation and experiment design
- **ResearchAgent** leverages an additional academic knowledge graph to enhancing literature retrieval and adopts a multi-agent framework to refine ideas through iterative peer discussion
- **GPT-Researcher**: An agent framework designed for research, enhanced with plan-and-solve and RAG capabilities.
- **AI-Scientist** components related to idea generation & experiment design
- **Real Paper**: Ideas and the experiment designs from real papers

https://arxiv.org/pdf/2410.13185

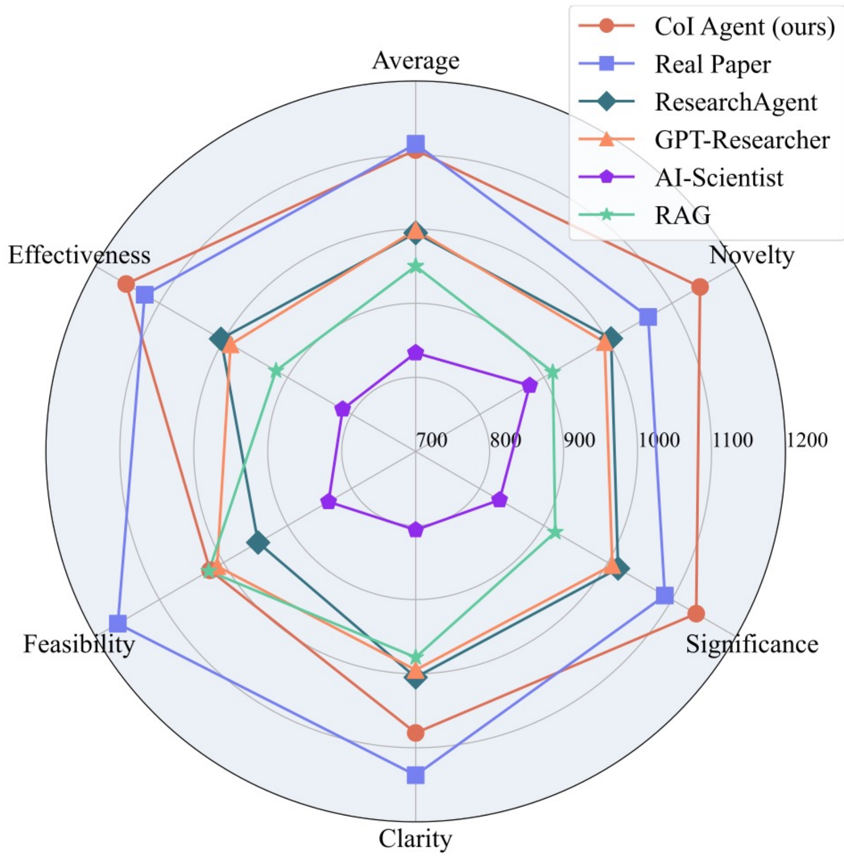


Figure 3: Evaluation results of idea generation with LLM as a judge.

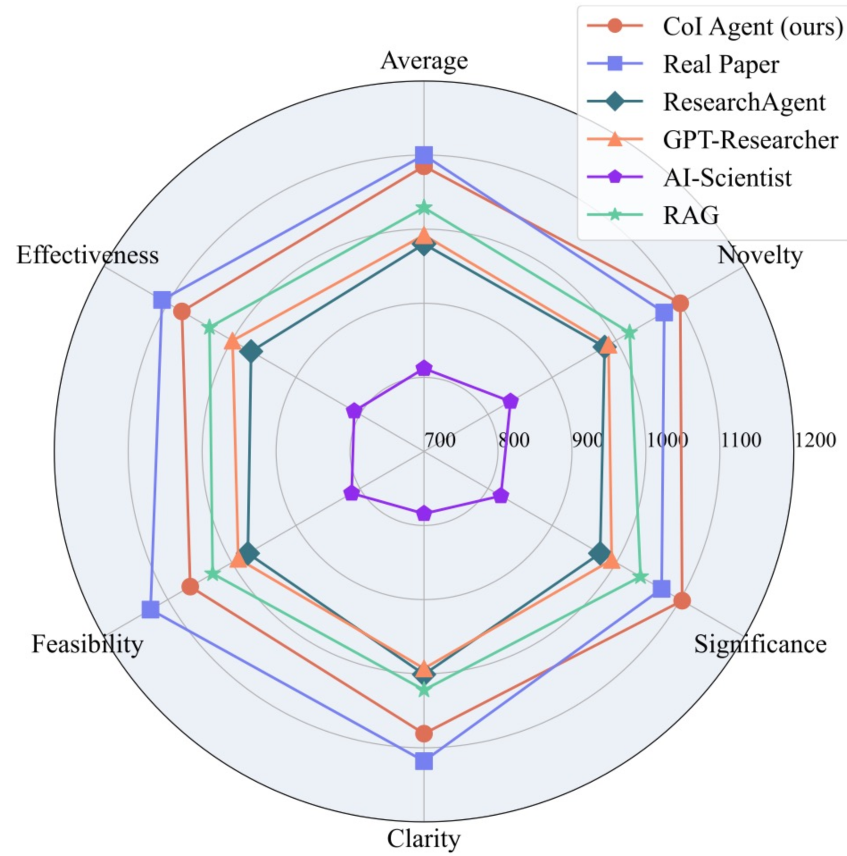


Figure 4: Evaluation results of idea generation with human as judges.

https://arxiv.org/pdf/2410.13185

		Feasibility	Tech.	Clarity	Average
Model Evaluation	Real Paper	1100	1122	1090	1103
	CoI Agent (ours)	**1029**	**1096**	**1043**	**1056**
	RAG	1022	970	1016	1003
	ResearchAgent	960	1020	980	987
	GPT-Researcher	1001	965	992	986
	AI-Scientist	888	827	879	865
Human Evaluation	Real Paper	1138	1111	1111	1120
	CoI Agent (ours)	**1092**	**1123**	**1121**	**1112**
	RAG	1035	1041	1048	1042
	GPT-Researcher	988	977	971	978
	ResearchAgent	939	959	964	954
	AI-Scientist	809	788	785	794
	Agreement	70.7%	75.9%	72.1%	73.0%

https://arxiv.org/pdf/2410.13185

Evaluation via “Idea Arena” protocol

- Human experts and LLM evaluators score ideas on five axes:
  - **Novelty** : How different from existing work?
  - **Significance**: Does it advance the field?
  - **Clarity** : Is it well-defined?
  - **Feasibility**: Can it be executed?
  - **Expected Effectiveness**: Is success impactful?
- Pairwise comparisons yield quantitative rankings
- Reported results: CoI ideas judged as more novel than human baselines but less feasible on average

Idea
Arena

```
You are a judge in a competition.  You have to decide which idea is
better.

The idea0 is:  [idea0]
The idea1 is:  [idea1]
The topic is:  [topic]
Which idea do you think is better?  Please write a short paragraph
to explain your choice.
Here are your evaluation criteria:
1.  Novelty:  Are the problems or approaches new?  Is this a novel
combination of familiar techniques?  Is it clear how this work
differs from previous contributions?  Is related work adequately
referenced?
2.  Significance:  Are the idea important?  Are other people
(practitioners or researchers) likely to use these ideas or build
on them?  Does the idea address a difficult problem in a better way
than previous research?  Does it provide a unique theoretical or
pragmatic approach?
3.  Feasibility:  Can the idea be realized with existing technology
or methods?  Are there any technical difficulties or bottlenecks?
Is the idea clear and logical?  Is there any obvious error or
unreasonable part in the idea, and can the experiment be designed
normally according to this idea.
```

```
4. Clarity: Is the paper clearly written? Is it well-organized?
Does it adequately inform the reader?
5. Effectiveness: How likely the proposed idea is going to work
well (e.g., better than existing baselines).
Note:
Avoid any position biases and ensure that the order in which
the responses were presented does not influence your decision.
DO NOT allow the LENGTH of the responses to influence your
evaluation, choose the one that is straight-to-the-point instead
of unnecessarily verbose. Be as objective as possible. (very
important!!!)
If you think idea0 is better than idea1, you should output 0. If
you think idea1 is better than idea0, you should output 1. If you
think idea0 and idea1 are equally good, you should output 2.

Your output should be strictly in following format:
Your thinking process: ...
Your choice:
Novelty: 0/1/2
Significance: 0/1/2
Feasibility: 0/1/2
Clarity: 0/1/2
Effectiveness: 0/1/2
```

Elo score

- Each competitor (or idea, model, or player) starts with a base score (say 1000)
- When two are compared, the winner gains points and the loser loses points
- The size of the change depends on how expected the result was:
  - Beating a higher-rated opponent gives a large boost
  - Beating a lower-rated one gives a small boost
- Over many comparisons, the system converges to scores where expected and actual outcomes balance

Hypothesis: Science advances via new instrumentation

213 fields developed since 1500 and the methods & instruments that enabled them, with four examples illustrated

https://doi.org/10.1057/s41599-025-05797-6

Questions to ponder

- The examples that we considered all looked at research in AI as their testbed. How widely do you think these results apply?
- Is originality a property of the output, or of the process that produced it?
- Can AI ever surprise us in a way that counts as discovery?
- Should AI systems cite their “training influences,” and if so, how?
- If an AI re-discovers a known result, is it plagiarism or convergence?