IRIS: Interactive Research Ideation System for Accelerating Scientific Discovery

Aniketh Garikaparthi[1], Manasi Patwardhan[1], Lovekesh Vig[1], Arman Cohan[2]

[1]TCS Research [2]Yale University
{aniketh.g, manasi.patwardhan, lovekesh.vig}@tcs.com,
arman.cohan@yale.edu

Abstract

The rapid advancement in capabilities of large language models (LLMs) raises a pivotal question: *How can LLMs accelerate scientific discovery?* This work tackles the crucial first stage of research, generating novel hypotheses. While recent work on automated hypothesis generation focuses on multi-agent frameworks and extending test-time compute, none of the approaches effectively incorporate transparency and steerability through a synergistic Human-in-the-loop (HITL) approach. To address this gap, we introduce IRIS: Interactive Research Ideation System, an open-source platform designed for researchers to leverage LLM-assisted scientific ideation. IRIS incorporates innovative features to enhance ideation, including adaptive test-time compute expansion via Monte Carlo Tree Search (MCTS), fine-grained feedback mechanism, and query-based literature synthesis. Designed to empower researchers with greater control and insight throughout the ideation process. We additionally conduct a user study with researchers across diverse disciplines, validating the effectiveness of our system in enhancing ideation. We open-source our code here.

1 Introduction

With the growing capabilities of large language models (LLMs), the automation of scientific discovery has captured a lot of attention (Gridach et al., 2025). Agentic LLM based systems have shown potential of outperforming PhD researchers and postdocs on short-horizon scientific tasks like question answering, summarization and contradiction detection in various domains (Skarlinski et al., 2024; Asai et al., 2024). These advancements have spurred new opportunities of LLMs accelerating scientific discovery, which is essential given the exponential growth of scientific publications (Landhuis, 2016; Fire and Guestrin, 2019).

Current solutions that leverage LLMs in scientific ideation primarily remain hinged on multi-agent frameworks or extending test-time compute (Si et al., 2024; Hu et al., 2024; Gottweis, 2025), and aim to validate the quality of the final ideas through human validation or LLM-as-a-judge evaluations (Wang et al., 2024; Li et al., 2024; Baek et al., 2025). However, these approaches often fail to integrate human supervision during generation in a truly complementary manner, neglecting the nuanced expectations and goals of the user. Consequently, despite investing significant computational resources to develop objectively "novel" ideas, they might not align with the user's *research goals*, inevitably leading to dissatisfaction (Ou et al., 2022; Kim et al., 2024).

Moreover, the importance of meaningful human intervention in the research process cannot be overstated. Notably, AI models have been known to fabricate convincing yet fraudulent scientific information (Májovský et al., 2023). More troubling are cases of deceptive and misaligned AI behaviors (Ryan Greenblatt, 2025; Booth, 2025; Betley et al., 2025; Baker et al., 2025). Recent developments of more capable Agentic LLMs have shown difficulties in transparently delegating sub-tasks, leading to *"reward hacking"* behaviors (Anthropic, 2025). In the context of idea generation, we find signs of similar *"reward hacking"* where LLMs adopt fancy terminology e.g. "Prompt Learning and Optimization Nexus" for building a library of prompts, or often proposing the use of "graphs" without any clear motivation or description behind the design choice. We observe that naive recursive feedback loops (Baek et al., 2025) forcing the LLM to be more novel inevitably lead to gamifying LLM-as-a-judge metrics without adding actual value. Gupta and Pruthi (2025) carefully study the results of AI-Researcher (Si et al., 2024) and advise careful assessment of LLM generated hypotheses due to signs of skillful plagiarism. These examples
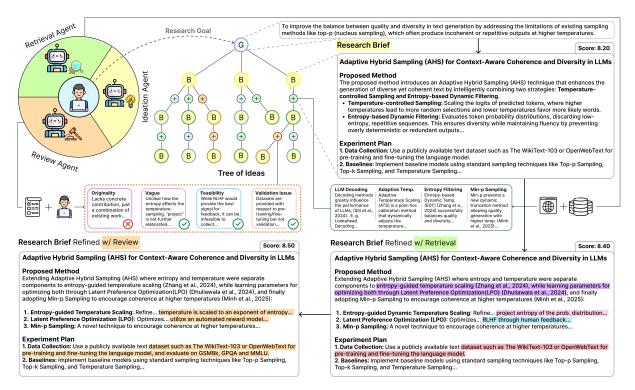
Figure 1: Human-in-the-loop Idea Generation with Monte-Carlo-Tree-Search. $\mathcal{G}$: Research Goal, $\mathcal{B}$: Research Brief

highlight the pitfalls of premature reliance on fully automated systems, underscoring the need for well-designed Human-in-the-Loop (HITL) systems for scientific ideation; ensuring outcomes are accurate and aligned with human goals.

Despite the recent innovations made in LLM-based scientific ideation, several key limitations persist. These include (1) generating hypotheses in a single pass (Si et al., 2024), which overlooks the iterative nature of the ideation process. In contrast, Pu et al. (2024) find that researchers typically seek to refine their hypotheses into concrete *research briefs*. (2) Optimization through feedback on coarse-grained criteria like rigorousness, originality, generalizability etc. (Baek et al., 2025), while often critiquing entire ideas rather than specific components. (3) Simplistic retrieval augmentation such as appending keywords or abstracts of previous papers in context (Wang et al., 2024; Si et al., 2024), whereas effective ideation demands a deeper, more holistic understanding of the domain literature. (4) Unstructured and sub-optimal search of the idea space through either refinement of a generated base-idea (exploitation) (Wang et al., 2024; Baek et al., 2025), or through initial search and plan (exploration) without subsequent refinement of promising ideas (Hu et al., 2024). Finally, there is a lack of open-source implementations that

would encourage broader adoption. In light of these challenges, we propose IRIS, tackling each of these limitations while enabling human intervention at every stage of the ideation process. Specifically, we make the following contributions:

- **HITL Framework:** A user-centered design balancing human control with automation instead of entirely delegating the process of ideation to AI

- **Monte Carlo Tree Search:** A systematic method to iteratively explore the idea space and extend test time compute via alternating phases of exploration and exploitation (§3.2)

- **Fine-grained Review based Refinement:** An exhaustive taxonomy (Table 2) with fine-grained actionable feedback for improving hypotheses (Figure 2) (§3.1)

- **Query-based Retrieval:** Generating targeted queries for retrieving relevant literature, with re-ranking, clustering and summarization to produce comprehensive, technical and cited responses (§3.1)

- **Open Source:** Publicly available platform for AI-Assisted scientific ideation

Finally, we conduct a user study with researchers from diverse disciplines validating the effectiveness of our designed system (§4).
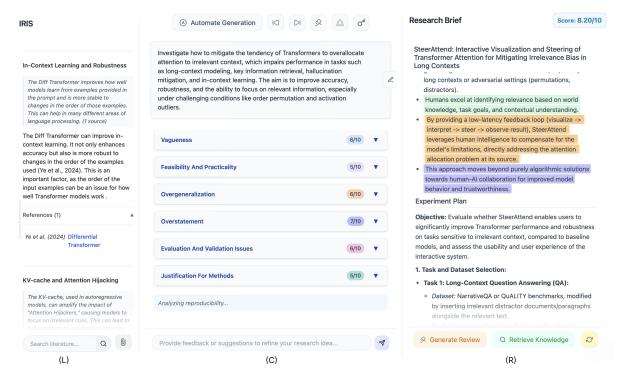
2

Figure 2: IRIS Platform Interface with (L) Retrieval Panel, (C) Chat Overview Panel, (R) Research Brief Panel

2 Related Works

2.1 AI Assisted Research

The integration of (AI) into scientific research has evolved from early concept-linking tools (Swanson, 1986; Sybrandt et al., 2020; Nadkarni et al., 2021) to sophisticated systems that enhance various research stages. In recent years, LLMs have significantly transformed research life-cycles by assisting in literature searches (Zheng et al., 2024; Ajith et al., 2024; Asai et al., 2024), citation recommendations (Pillai and R, 2022; Zhang and Zhu, 2022; Press et al., 2024), review of scientific documents (Zhou et al., 2024), experimental design (Huang et al., 2024; Schmidgall et al., 2025), scientific claim verification (Lu et al., 2024), theorem proving (Song et al., 2025), manuscript writing (Weng et al., 2025), and reading assistants[1].

2.2 Human-AI Co-creation Systems

The emergence of Gen AI has introduced a new dimension to Co-creation systems, setting them apart from previous ones where machines primarily served as supportive tools for human users (Davis et al., 2015; Muller et al., 2020; Weisz et al., 2024). Recent studies, such as those by Kantosalo and Jordanous (2021); Liu et al. (2024), demonstrate the effectiveness of Gen AI tools in creative tasks, particularly through their steerability and explain-ability. This has led to growing emphasis among researchers to develop design guidelines for integrating Gen AI into existing frameworks (Amershi et al., 2019; Shneiderman, 2020). We build IRIS for researcher-in-the-loop ideation while incorporating design principles from prior work, such as minimizing opacity, adopting granular feedback, encouraging AI processing delays (Amershi et al., 2019; Liu et al., 2024), and replacing rigid post-hoc analysis with oversight across planning, generation, and retrospection stages (Shneiderman, 2020).

2.3 Automated Hypothesis Generation

Spangler et al. (2014) demonstrate the first proof of principle for automated hypothesis generation through text mining of scientific literature, leveraging techniques such as entity detection and graph-based diffusion of information. Rising capabilities of text completion models has driven significant advancements in this field (Wang et al., 2024; Lu et al., 2024; Li et al., 2024; Hu et al., 2024; Si et al., 2024; Kumar et al., 2024; Baek et al., 2025; Gottweis, 2025). However, current efforts focus on fully automated systems, often overlooking the critical role of human involvement. *Acceleron* demonstrates one of the first human-in-the-loop (HITL) framework assisting researchers in validation of motivation behind a research problem and synthesizing a method for the same (Nigam et al., 2024), followed by Pu et al. (2024) making an attempt to develop

---

[1] JenniAI, SciSpace, ScholarAI

3

an interactive idea generation system. These approaches remain limited, allowing idea exploration only within a predefined framework, restricting flexibility and adaptability. Furthermore, their system lacks sophisticated components like automated fine-grained feedback, literature retrieval targeted to the research goal and scaling test-time compute.

3 IRIS

Broadly, the system expects as input a research goal $\mathcal{G}$ consisting of a research problem and it's motivation, and outputs a research brief $\mathcal{B}$ consisting of a Title, Proposed Methodology and Experiment Plan, while improving it's quality; either in *semi-automatic* manner through directions from the researcher or *autonomously* exploiting Monte Carlo Tree Search (MCTS). We provide detailed overview of our system including the implementation of agents (§3.1) and MCTS adaptation for hypothesis generation (§3.2).

3.1 Agent Architecture

IRIS employs a three-agent architecture consisting of an ideation agent, a review agent, and a retrieval agent. The ideation agent navigates the search space of possible research ideas, while the review and retrieval agents provide feedback and relevant scientific context respectively.

Ideation Agent generates and iteratively improves the research brief. It can toggle between a *semi-automatic* mode, to receive guidance from a researcher to refine research briefs through steering reviews, retrievals or employing custom feedback, and a completely *autonomous* mode to explore and exploit the idea space by leveraging *actions* which support iterative refinement of the research briefs through MCTS.

Review Agent is accountable for two tasks namely providing *reward* and *feedback*. For evaluation of an idea, we have defined a hierarchical taxonomy of aspects grounded in real-world scientific critique (For example, (Ghosal et al., 2022), (Kennard et al., 2022), (Dycke et al., 2023)), detailed in Table 2. Review Agent is auto-triggered after each new generation of the research brief to provide a *reward* averaged over the scores assigned to distinct aspects, based on the evaluation provided for the complete research brief.

As opposed to the parallel works (Wang et al., 2024; Baek et al., 2025) that focus on coarse-level

criteria and provide broad evaluation of the entire generated research brief, usually, a feedback with respect to an aspect is applicable to only specific parts of the research brief. For example, only some component of the brief can be infeasible or some other component requires more clarity. Addressing this need, when explicitly triggered by the researcher, the review agent switches to a fine-grained evaluation, delivering targeted, actionable feedback on each aspect of the taxonomy for distinct segments of the current research-brief (Figures 1 and 2 (R)). This fine-grained feedback is verified by the researcher and omitted if deemed irrelevant. Then the review agent computes reward based on the scores of the verified aspects of the feedback. This adept human intervention coupled with granular feedback, successfully mitigates *"reward hacking"* behavior of LLMs.

Retrieval Agent: For the input research goal, the retrieval agent synthesizes queries targeted to retrieve literature relevant to the research goal. For answering each query, it adopts Ai2 Scholar QA API[2]. The pipeline consists of two-stage retrieval followed by three-stage generation. The Semantic Scholar API's (storing over 200M open access papers) snippet search endpoint (Kinney et al., 2023) extracts relevant passages, which are re-ranked to retain top-k passages and aggregated at the paper level. With the finalized set of passages, the retrieval agent (i) extracts quotes from the passages relevant to the query, (ii) generates a plan to produce an organized report with sections, and clusters the top-k passages accordingly, and (iii) generates cited sections-wise reports along with summaries (Figure 2 (L)). Our motivation for adopting ScholarQA stems from the limitations of naive RAG failing to appropriately answer global questions targeted at a corpus as opposed to a single document (Edge et al., 2025). We also provide the ability for the researcher to upload papers in the form of PDF documents, which they think to be relevant but have been missed out as the part of the retrieval. The retrieval agent parses the PDF through Grobid based doc2json tool[3] and appends the most relevant chunks to the context for the ideation agent to refine the research brief.

3.2 Monte Carlo Tree Search Framework

To systematically explore the vast space of potential research ideas, IRIS employs Monte Carlo Tree

---

[2] https://allenai.org/blog/ai2-scholarqa
[3] https://github.com/allenai/s2orc-doc2json

Search (MCTS) ([Kocsis and Szepesvári, 2006](#)). MCTS allows the system to effectively extend test-time compute similar to recent work in augmenting LLM reasoning ([Qi et al., 2024](#); [Guan et al., 2025](#)). Unlike applications with objective rewards (e.g., mathematics, code generation), scientific ideation quality is subjective. We adapt MCTS by using the LLM-based Review Agent as a proxy judge to estimate the *quality* (reward) of generated hypotheses.

Formally, given a research goal $\mathcal{G}$, our system constructs a search tree $\mathcal{T}$ rooted with $\mathcal{G}$. A state s encapsulates the current {research brief b, reward estimate r, latest review feedback f (if applicable, else ϕ), and retrieved knowledge k (if applicable, else ϕ)}. Edges represent actions a taken by the Ideation Agent to transition between states. We define a comprehensive action space $\mathcal{A} = \{a_1$: generate, a_2: refine w/ retrieval, a_3: refine w/ review, a_4: refine w/ user feedback$\}$. The MCTS process iteratively builds the tree over N iterations, guided by the Upper Confidence Bound for Trees (UCT) algorithm ([Coquelin and Munos, 2007](#)). UCT of a node n is defined by:

$$\text{UCT}(n) = \frac{Q(n)}{N(n)} + c\sqrt{\frac{\ln N(n_p)}{N(n)}} \qquad (1)$$

where $Q(n)$ is the total reward at child node n accumulated from its children, $N(n)$ is its visit count, $N(n_p)$ is the visit count of the parent node of n, and c is the exploration constant. Algorithm 1 outlines the MCTS process. Each node n stores its state s_n as defined above, $Q(n)$ and $N(n)$.

Algorithm 1 MCTS for Research Idea Generation

Require: Research goal $\mathcal{G}$, iterations N, max depth $d_{\max}$, actions $\mathcal{A}$, constant c
1: Initialize tree $\mathcal{T}$ with root n_0 (state $s_0 = \mathcal{G}$, $Q(n_0) = 0$, $N(n_0) = 0$).
2: **for** $i = 1$ to N **do**
3:     $n_{\text{leaf}} \leftarrow \text{SELECT}(n_0, c)$
4:     $r \leftarrow \text{EVALUATE}(n_{\text{leaf}})$
5:     **if** depth $< d_{\max}$ **then**
6:         $\text{EXPAND}(n_{\text{leaf}}, \mathcal{A})$
7:     **end if**
8:     $\text{BACKPROPAGATE}(n_{\text{leaf}}, r)$
9: **end for**
10: **return** $\text{BESTCHILD}(n_0)$

Each iteration involves four phases:

SELECT(n_{root}, c): Traverse the tree from the root n_0 to select a leaf node n_{leaf}. At each node n during traversal, if n has any unvisited children ($Q(n) = 0$), one such child is randomly selected. If all children of n have been visited, the next node is chosen by: $\arg\max_{n' \in \text{children}(n)}(\text{UCT}(n'))$.

EVALUATE(n_{leaf}): Obtain reward r for the state s_{leaf} of n_{leaf} via the Review Agent.

EXPAND$(n_{\text{leaf}}, \mathcal{A})$: If n_{leaf} is non-terminal and below $d_{\max}$, create child nodes n' for each applicable action $a \in \mathcal{A}$, with $Q(n') = 0, N(n') = 0$.

BACKPROPAGATE(n_{leaf}, r): Update Q and N values for n_{leaf} and its ancestors with reward r.

BESTCHILD(n_0): After N iterations, select the child of n_0 with the highest average reward Q/N.

Memory: Agents maintain trajectory-level memory. For instance, the Ideation Agent recalls generated briefs, the Retrieval Agent remembers past queries, and the Review Agent tracks prior feedback. This helps steer the generation towards non-redundant refinements.

Cost: MCTS can be computationally intensive. IRIS incorporates budget controls, allowing users to set limits. For tighter budgets, the system prioritizes exploitation by lowering the exploration constant c, ensuring delivery of few refined outputs rather than numerous low-quality ones.

4 Evaluation

To assess the effectiveness and usability of IRIS, we conduct automated evaluations and user studies.

4.1 Experiment Setup

System Implementation: IRIS's user interface is developed using HTML, CSS, JavaScript. The core LLM functionalities are powered by Gemini-2.0-Flash ([DeepMind, 2024](#)) accessed via LiteLLM[4], which allows users to substitute other LLMs of their choice. We utilize Gemini's built-in safety filters to mitigate harmful or inappropriate queries.

Metrics: We employ LLM-as-a-judge, popularly adopted in parallel literature ([Baek et al., 2025](#); [Gottweis, 2025](#)). We use two methods guided by our pre-defined criteria (Table [2](#)). *absolute score:* each generated hypothesis (1-10), and *relative score:* aggregating head-to-head comparisons and preferences to compute ELO ratings.

To contextualize the alignment of LLM-as-a-judge with human preferences in the context of scientific ideation, we prompt baselines Gemini-2.0-Flash, ChatGPT, ChatGPT w/ search and Claude

---

[4]https://docs.litellm.ai/docs/

3.5 Haiku to generate novel research briefs. Then ask users and LLMs to rate the generations in the order of their preference.
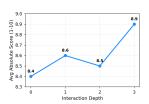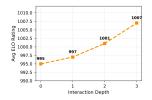
4.2 User Study

We conducted a user study with 8 researchers (N=8) from diverse fields (AI/NLP, Chem, Physics, HCI) and experience levels. Two users voluntarily participated twice (10 total case studies). Each ∼60 min session involved: 1) Defining a research goal, 2) Blindly ranking initial set of hypotheses, 3) Interacting with IRIS, 4) Completing a post-task survey.

4.3 Results and Analysis

Metric Validation: Human baseline rankings correlated moderately with LLM based ELO scores (Pearson's r=0.60) but weakly with LLM based absolute scores (r=0.45). With this learning we plan to replace the LLM-as-the-judge scores, displayed to showcase the quality of the idea, with the ELO ratings.

Automated Evaluation: LLM-as-a-judge evaluations (Figure 3) showed that user interaction within IRIS consistently improved hypothesis quality, increasing average absolute scores by 0.5 points and ELO ratings by 12 points for a tree depth of 3.

(a) Absolute Score Improvement.

(b) ELO Rating Improvement.

Figure 3: Iterative improvement in hypothesis quality within IRIS over interaction depth (up to depth 3). Interaction enhances both absolute scores and ELO ratings.

User Study Feedback: Quantitative ratings (Table 1) show users found the fine-grained feedback highly insightful and unpromptedly mentioned better usability and control over other reading assistant interfaces mentioned in §2.

Feature / Aspect	Mean Rating (± Std Dev)
Usefulness of Fine-grained Feedback	4.3 ± 0.7
MCTS Tree Interface (Steerability)	4.2 ± 0.6
Quality of Lit. Summaries	3.7 ± 0.8
Usability and control	4.5 ± 0.7
Overall Satisfaction (Final Research Brief)	3.9 ± 0.7

Table 1: User ratings (1-5 Likert scale) for key IRIS features and overall satisfaction (N=10).

Additionally, through qualitative feedback we arrived at the following insights:

- **Steerability:** All users valued the MCTS tree for control and transparency over ideation.

- **Feedback:** Critiques often reflected user's own concerns (87.5% users) and sometimes sparked novel insights (50% cases).

- **Retrieval:** Found to be facilitating grounding of ideas, but quality varied with domains such as chemistry and physics research, matching the lower rating (3.7/5). We attribute this to reduced availability of relevant literature in the semantic scholar corpus.

- **Relevance:** hypotheses often shared similarities with or extended users' ongoing work (62.5% users).

Overall Improvement: Post-interaction, 25% (2/8) found the hypothesis substantially better, 50% (4/8) marginally better, and 25% (2/8) similar quality. Crucially, all users reported enhanced understanding of the proposed methodology, and considered it to be promising.

5 Conclusion

We introduce IRIS, an Interactive Research Ideation System, to augment automated scientific hypothesis generation with human expertise. We apply MCTS to iteratively explore the idea space, refine ideas with fine-grained segment level reviews and targeted query based multi-document retrieval; offering a steerable environment for researchers during LLM-driven scientific ideation. Our user study validates the usability and effectiveness of our system, demonstrating consistent improvement in hypothesis quality increasing average absolute scores by 0.5 points and ELO ratings by 12 points for a tree depth of 3. Crucially, users frequently considered the generated hypotheses plausible and worthy of further investigation. We position that the potential of LLMs, particularly within human-AI collaborative frameworks, for developing novel scientific hypothesis remains a heavily underexplored avenue. We present IRIS as a concrete step towards realizing this untapped potential.

Limitations

Currently the system relies on the researcher as the judge to verify the quality of the emerging

idea at each iteration, augmented by LLM-as-the-judge. This reliance is based on the assumption of sufficient domain expertise of the researcher. As opposed to this in future we aim for a true Human AI Co-creation System, where more foundational LLMs with scientific expertise, questions researchers for the choices he or she has made leading to a two way socratic review and refinement communication, simulating a more realistic scenario of brain-storming between colleagues or a mentor and a mentee.

Due to budget constraints, we have not explored frontier LLMs such as Claude 3.7 Sonnet, Grok-3 or reasoning models like Gemini-2.5-Pro, o1 etc. The quality of produced hypothesis in terms of novelty and effectiveness would likely benefit from stronger base models.

References

Anirudh Ajith, Mengzhou Xia, Alexis Chevalier, Tanya Goyal, Danqi Chen, and Tianyu Gao. 2024. Litsearch: A retrieval benchmark for scientific literature search. Preprint, arXiv:2407.18940.

Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for human-ai interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19, page 1–13, New York, NY, USA. Association for Computing Machinery.

Anthropic. 2025. Claude 3.7 sonnet system card. Accessed: 2025-03-10.

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'arcy, David Wadden, Matt Latzke, Minyang Tian, Pan Ji, Shengyan Liu, Hao Tong, Bohao Wu, Yanyu Xiong, Luke Zettlemoyer, Graham Neubig, Dan Weld, Doug Downey, Wen tau Yih, Pang Wei Koh, and Hannaneh Hajishirzi. 2024. Openscholar: Synthesizing scientific literature with retrieval-augmented lms. Preprint, arXiv:2411.14199.

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 2025. Researchagent: Iterative research idea generation over scientific literature with large language models. Preprint, arXiv:2404.07738.

Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y. Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. 2025. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. Accessed: 2025-03-11.

Jan Betley, Daniel Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. 2025. Emergent misalignment: Narrow finetuning can produce broadly misaligned llms. Preprint, arXiv:2502.17424.

Harry Booth. 2025. Ai and chess cheating: Palisade research raises concerns. Time. Accessed: 2025-02-25.

Pierre-Arnaud Coquelin and Rémi Munos. 2007. Bandit algorithms for tree search. Preprint, arXiv:cs/0703062.

Nicholas Davis, Chih-Pin Hsiao, Yanna Popova, and Brian Magerko. 2015. An Enactive Model of Creativity for Computational Collaboration and Co-creation, pages 109–133. Springer London, London.

Google DeepMind. 2024. Google gemini ai update: December 2024. https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/. Accessed: 2025-03-24.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. 2023. NLPeer: A unified resource for the computational study of peer review. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5049–5073, Toronto, Canada. Association for Computational Linguistics.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2025. From local to global: A graph rag approach to query-focused summarization. Preprint, arXiv:2404.16130.

Michael Fire and Carlos Guestrin. 2019. Overoptimization of academic publishing metrics: observing goodhart's law in action. GigaScience, 8(6):giz053.

Tirthankar Ghosal, Sandeep Kumar, Prabhat Kumar Bharti, and Asif Ekbal. 2022. Peer review analyze: A novel benchmark resource for computational analysis of peer reviews. PLOS ONE, 17(1):1–29.

Juraj Gottweis. 2025. Towards an ai co-scientist.

Mourad Gridach, Jay Nanavati, Khaldoun Zine El Abidine, Lenon Mendes, and Christina Mack. 2025. Agentic ai for scientific discovery: A survey of progress, challenges, and future directions. Preprint, arXiv:2503.08979.

Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. 2025. rstar-math: Small llms can master math reasoning with self-evolved deep thinking. Preprint, arXiv:2501.04519.

Tarun Gupta and Danish Pruthi. 2025. All that glitters is not novel: Plagiarism in ai generated research. Preprint, arXiv:2502.16487.

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. *Preprint*, arXiv:2410.14255.

Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. 2024. Mlagentbench: Evaluating language agents on machine learning experimentation. *Preprint*, arXiv:2310.03302.

Anna Kantosalo and Anna Jordanous. 2021. Role-based perceptions of computer participants in human-computer co-creativity. In *7th Computational Creativity Symposium at AISB 2021*, pages 20–26, London, UK. AISB.

Neha Nayak Kennard, Tim O'Gorman, Rajarshi Das, Akshay Sharma, Chhandak Bagchi, Matthew Clinton, Pranay Kumar Yelugam, Hamed Zamani, and Andrew McCallum. 2022. DISAPERE: A dataset for discourse structure in peer review discussions. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1234–1249, Seattle, United States. Association for Computational Linguistics.

Yoonsu Kim, Jueon Lee, Seoyoung Kim, Jaehyuk Park, and Juho Kim. 2024. Understanding users' dissatisfaction with chatgpt responses: Types, resolving tactics, and the effect of knowledge level. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, IUI '24, page 385–404, New York, NY, USA. Association for Computing Machinery.

Rodney Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, Miles Crawford, Doug Downey, Jason Dunkelberger, Oren Etzioni, Rob Evans, Sergey Feldman, Joseph Gorney, David Graham, Fangzhou Hu, Regan Huff, Daniel King, Sebastian Kohlmeier, Bailey Kuehl, Michael Langan, Daniel Lin, Haokun Liu, Kyle Lo, Jaron Lochner, Kelsey MacMillan, Tyler Murray, Chris Newell, Smita Rao, Shaurya Rohatgi, Paul Sayre, Zejiang Shen, Amanpreet Singh, Luca Soldaini, Shivashankar Subramanian, Amber Tanaka, Alex D. Wade, Linda Wagner, Lucy Lu Wang, Chris Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Madeleine Van Zuylen, and Daniel S. Weld. 2023. The semantic scholar open data platform. *Preprint*, arXiv:2301.10140.

Levente Kocsis and Csaba Szepesvári. 2006. Bandit based monte-carlo planning. In *Machine Learning: ECML 2006*, pages 282–293, Berlin, Heidelberg. Springer Berlin Heidelberg.

Sandeep Kumar, Tirthankar Ghosal, Vinayak Goyal, and Asif Ekbal. 2024. Can large language models unlock novel scientific research ideas? *Preprint*, arXiv:2409.06185.

Esther Landhuis. 2016. Scientific literature: Information overload. *Nature*, 535:457 – 458.

Long Li, Weiwen Xu, Jiayan Guo, Ruochen Zhao, Xingxuan Li, Yuqian Yuan, Boqiang Zhang, Yuming Jiang, Yifei Xin, Ronghao Dang, Deli Zhao, Yu Rong, Tian Feng, and Lidong Bing. 2024. Chain of ideas: Revolutionizing research via novel idea development with llm agents. *Preprint*, arXiv:2410.13185.

Yiren Liu, Si Chen, Haocong Cheng, Mengxia Yu, Xiao Ran, Andrew Mo, Yiliu Tang, and Yun Huang. 2024. How ai processing delays foster creativity: Exploring research question co-creation with an llm-based agent. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

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. *Preprint*, arXiv:2408.06292.

Martin Májovský, Martin Černý, Matěj Kasal, Martin Komarc, and David Netuka. 2023. Artificial intelligence can generate fraudulent but authentic-looking scientific medical articles: Pandora's box has been opened. *J Med Internet Res*, 25:e46924.

Michael Muller, Justin D Weisz, and Werner Geyer. 2020. Mixed initiative generative ai interfaces: An analytic framework for generative ai applications. In *Proceedings of the Workshop The Future of Co-Creative Systems-A Workshop on Human-Computer Co-Creativity of the 11th International Conference on Computational Creativity (ICCC 2020)*.

Rahul Nadkarni, David Wadden, Iz Beltagy, Noah A. Smith, Hannaneh Hajishirzi, and Tom Hope. 2021. Scientific language models for biomedical knowledge base completion: An empirical study. *Preprint*, arXiv:2106.09700.

Harshit Nigam, Manasi Patwardhan, Lovekesh Vig, and Gautam Shroff. 2024. An interactive co-pilot for accelerated research ideation. In *Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 60–73, Mexico City, Mexico. Association for Computational Linguistics.

Changkun Ou, Daniel Buschek, Sven Mayer, and Andreas Butz. 2022. The human in the infinite loop: A case study on revealing and explaining human-ai interaction loop failures. In *Proceedings of Mensch Und Computer 2022*, MuC '22, page 158–168, New York, NY, USA. Association for Computing Machinery.

Reshma S Pillai and Deepthi L R. 2022. A survey on citation recommendation system. In *2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT)*, pages 423–429.
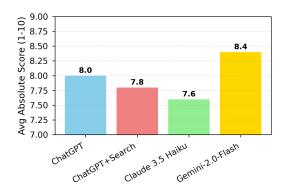
Ori Press, Andreas Hochlehnert, Ameya Prabhu, Vishaal Udandarao, Ofir Press, and Matthias Bethge. 2024. Citeme: Can language models accurately cite scientific claims? *Preprint*, arXiv:2407.12861.

Kevin Pu, K. J. Kevin Feng, Tovi Grossman, Tom Hope, Bhavana Dalvi Mishra, Matt Latzke, Jonathan Bragg, Joseph Chee Chang, and Pao Siangliulue. 2024. Ideasynth: Iterative research idea development through evolving and composing idea facets with literature-grounded feedback. *Preprint*, arXiv:2410.04025.

Zhenting Qi, Mingyuan Ma, Jiahang Xu, Li Lyna Zhang, Fan Yang, and Mao Yang. 2024. Mutual reasoning makes smaller llms stronger problem-solvers. *Preprint*, arXiv:2408.06195.

et.al Ryan Greenblatt. 2025. Alignment faking in large language models. Accessed: 2025-02-25.

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. *Preprint*, arXiv:2501.04227.

Ben Shneiderman. 2020. Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered ai systems. *ACM Trans. Interact. Intell. Syst.*, 10(4).

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2024. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *Preprint*, arXiv:2409.04109.

Michael D. Skarlinski, Sam Cox, Jon M. Laurent, James D. Braza, Michaela Hinks, Michael J. Hammerling, Manvitha Ponnapati, Samuel G. Rodriques, and Andrew D. White. 2024. Language agents achieve superhuman synthesis of scientific knowledge. *Preprint*, arXiv:2409.13740.

Peiyang Song, Kaiyu Yang, and Anima Anandkumar. 2025. Lean copilot: Large language models as copilots for theorem proving in lean. *Preprint*, arXiv:2404.12534.

Scott Spangler, Angela D. Wilkins, Benjamin J. Bachman, Meena Nagarajan, Tajhal Dayaram, Peter Haas, Sam Regenbogen, Curtis R. Pickering, Austin Comer, Jeffrey N. Myers, Ioana Stanoi, Linda Kato, Ana Lelescu, Jacques J. Labrie, Neha Parikh, Andreas Martin Lisewski, Lawrence Donehower, Ying Chen, and Olivier Lichtarge. 2014. Automated hypothesis generation based on mining scientific literature. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, page 1877–1886, New York, NY, USA. Association for Computing Machinery.

Don R. Swanson. 1986. Undiscovered public knowledge. *The Library Quarterly: Information, Community, Policy*, 56(2):103–118.
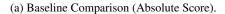
Justin Sybrandt, Ilya Tyagin, Michael Shtutman, and Ilya Safro. 2020. Agatha: Automatic graph mining and transformer based hypothesis generation approach. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 2757–2764, New York, NY, USA. Association for Computing Machinery.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2024. SciMON: Scientific inspiration machines optimized for novelty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 279–299, Bangkok, Thailand. Association for Computational Linguistics.

Justin D. Weisz, Jessica He, Michael Muller, Gabriela Hoefer, Rachel Miles, and Werner Geyer. 2024. Design principles for generative ai applications. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. 2025. Cycleresearcher: Improving automated research via automated review. *Preprint*, arXiv:2411.00816.

Jinzhu Zhang and Lipeng Zhu. 2022. Citation recommendation using semantic representation of cited papers' relations and content. *Expert Systems with Applications*, 187:115826.

Yuxiang Zheng, Shichao Sun, Lin Qiu, Dongyu Ru, Cheng Jiayang, Xuefeng Li, Jifan Lin, Binjie Wang, Yun Luo, Renjie Pan, Yang Xu, Qingkai Min, Zizhao Zhang, Yiwen Wang, Wenjie Li, and Pengfei Liu. 2024. Openresearcher: Unleashing ai for accelerated scientific research. *Preprint*, arXiv:2408.06941.

Ruiyang Zhou, Lu Chen, and Kai Yu. 2024. Is LLM a reliable reviewer? a comprehensive evaluation of LLM on automatic paper reviewing tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9340–9351, Torino, Italia. ELRA and ICCL.
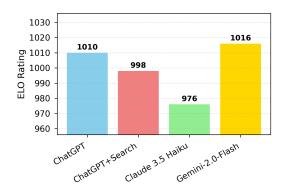
A Review Taxonomy

Aspect	Sub-aspect	Definition
Originality	Lack of Novelty	The idea does not introduce a significant or meaningful advancement over existing work, lacking originality or innovation.
	Assumptions	The idea relies on untested or unrealistic assumptions that may weaken its validity or applicability.
Clarity	Vagueness	The idea is presented in an unclear or ambiguous manner, making it difficult to understand its core components or contributions.
	Contradictory Statements	The idea contains internal inconsistencies or conflicts in its assumptions, methods, or conclusions.
	Alignment	The idea is not aligned with the problem statement and its objectives.
Feasibility	Feasibility and Practicality	The idea is not practical or achievable given current technological, theoretical, or resource constraints.
	Justification for Methods	The idea does not provide sufficient reasoning or evidence to explain why specific methods, techniques, or approaches were chosen.
Effectiveness	Evaluation and Validation Issues	The idea lacks rigorous evaluation methods, such as insufficient benchmarks, inadequate baselines, or poorly defined success metrics.
	Reproducibility and Robustness	The idea does not provide sufficient detail or transparency to allow others to replicate or verify its findings, and is not resilient to variations in input data, assumptions, or environmental conditions. The degree to which the solution consistently produces accurate and dependable results is low, making it less reliable.
Impact	Overgeneralization and Overstatement	The idea extends its conclusions or applicability beyond the scope of the context provided or exaggerates its claims, significance, or potential impact beyond what is supported by evidence or reasoning.
	Impact	The idea is not impactful or significant. It does not solve a real problem. It does not create value by solving a significant problem or fulfilling a need for individuals, organizations, or society.
	Ethical and Social Considerations	The idea does not adhere to ethical standards and is harmful to individuals, communities, or the environment.

Table 2: Hierarchical Review Taxonomy

(a) Baseline Comparison (Absolute Score).

(b) Baseline Comparison (ELO Rating).

IRIS User Feedback Survey

Thank you for using IRIS! Please take a moment to share your feedback on your recent experience. Your input is valuable for improving the system.

Please rate the following aspects of IRIS on a scale of 1 to 5, where:

1 = Very Poor / Not at all Useful

2 = Poor / Slightly Useful

3 = Neutral / Moderately Useful

4 = Good / Useful

5 = Very Good / Very Useful

1. **Fine-grained Feedback:** How useful did you find the fine-grained feedback mechanism for refining the research idea?
   - ( ) 1 ( ) 2 ( ) 3 ( ) 4 ( ) 5

2. **Tree Interface (Steerability):** How effective was the tree interface (MCTS exploration) for exploring different idea paths and steering the generation?
   - ( ) 1 ( ) 2 ( ) 3 ( ) 4 ( ) 5

3. **Literature Summaries:** How would you rate the quality and relevance of the literature summaries provided by the system?
   - ( ) 1 ( ) 2 ( ) 3 ( ) 4 ( ) 5

4. **Usability and Control:** How would you rate the overall usability (ease of use) and your sense of control while interacting with the IRIS interface?
   - ( ) 1 ( ) 2 ( ) 3 ( ) 4 ( ) 5

5. **Overall Satisfaction:** Overall, how satisfied were you with the final research hypothesis generated/refined using IRIS during this session?
   - ( ) 1 ( ) 2 ( ) 3 ( ) 4 ( ) 5

Figure 4: Top: Comparison of hypothesis quality generated by baseline methods (ChatGPT, ChatGPT+Search, Claude 3.5 Haiku, Gemini-2.0-Flash) using LLM-as-a-judge absolute scores and ELO ratings. Bottom: User Survey Feedback Form Questions.