EvoSci: A Bio-Inspired Multi-Agent Framework for the Evolution of Scientific Discovery

Xiaoyu Xiong, Yuqi Ren†, Deyi Xiong†
TJUNLP Lab, School of Computer Science and Technology, Tianjin University, China
{2025244184, ryq20, dyxiong}@tju.edu.cn

Abstract

Large language models (LLMs), have shown strong potential in scientific discovery, yet existing methods still face substantial challenges in the design of research workflows and multi-role collaboration mechanisms. To mitigate these issues, we propose EvoSci, a multi-agent scientific collaboration framework, which integrates bio-inspired evolution with knowledge graph modeling. To iteratively generate, evaluate, and refine research ideas, EvoSci incorporates multiple role-based agents, including mentor, researcher, and reviewer. By combining collaborative reasoning, shared memory, and evolutionary feedback, EvoSci significantly enhances the coherence and creativity of scientific exploration. Experiments on real-world research topics demonstrate that EvoSci significantly outperforms strong baselines in LLM-based structured peer-review and comparative ranking evaluations, achieving the highest overall peer-review score (ICLR 4.90) and top ranking (Top-10 = 54). These results suggest its superiority in both scientific idea generation and continuous discovery.

1 Introduction

With the breakthrough development in knowledge representation (Pan et al., 2024), logical reasoning (Ke et al., 2025; Xu et al., 2025), and complex multimodal information integration (Han et al., 2025), LLMs are gradually reshaping the paradigm of scientific research (Buehler, 2024). Significant progress has already been made in traditional AI-powered domains such as mathematical reasoning and theorem proving (Trinh et al., 2024; Zhang and Xiong, 2025; Liu et al., 2025), automatic code generation (Li et al., 2023; Ren et al., 2023; He et al., 2024; Yang et al., 2025), and complex data analysis (Sui et al., 2024). Building on these successes, researchers are increasingly exploring the potential of LLMs across broader scientific workflows, including idea and hypothesis generation from large-scale scientific corpora (Kulkarni et al., 2025; Zhou et al., 2024b; Wang et al., 2024b; Yang et al., 2024; Wang et al., 2024a), experimental design (Desai et al., 2025; Tian et al., 2021; Noh et al., 2024; Tom et al., 2024), and result interpretation (Zheng et al., 2023; Charness et al., 2025).

However, scientific discovery is not a one-shot solution but a gradual and evolving process, driven by the continual refinement of research problems and the accumulation of intermediate insights (Elliott, 2012). It is also fundamentally collaborative, relying on the interplay of diverse roles and perspectives (Milojević, 2014). However, existing approaches often reduce LLMs to static executors within rigid pipelines, overlooking their potential for long-horizon inquiry and structured coordination. This raises two central challenges: (1) how could LLMs be steered toward progressively deepening scientific problems? and (2) how could effective collaboration frameworks be developed to allow multiple agents to engage in sustained, dynamic exploration.

To address these challenges, we propose **EvoSci**, an **Evo**lutionary **Sci**ence framework driven by multiple collaborative agents for automatic research ideation. Inspired by real-world research teams and biological evolution, EvoSci models scientific discovery as a long-horizon, iterative exploration, consisting of four stages: **Problem Space Construction**, **Collaborative Research Execution**, **Research Idea Evaluation**, and **Bio-Inspired Evolutionary Iteration**. EvoSci builds upon explicitly defined role-based agents, including a mentor, a group of researchers, and a reviewer, each responsible for a distinct stage of the ideation process. Beyond static, pre-defined pipelines, EvoSci enables adaptive coordination through dynamic task decomposition, where the mentor agent reallocates subtasks based on intermediate feedback, while role-

† Corresponding authors.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9846–9878

aware assignment ensures that each agent's actions remain aligned with its disciplinary background across multiple interaction rounds. Furthermore, inspired by biological evolution, we iteratively update research ideas by aligning and recombining conceptual knowledge across different domains, thereby enhancing the novelty of scientific exploration.

To validate the effectiveness and generality of EvoSci, we have conducted systematic experiments across diverse scientific scenarios. Results show that EvoSci consistently generates more novel and impactful research ideas than strong baselines. Equipped with DeepSeek-v3, EvoSci achieves the highest overall peer-review scores (ICLR 4.90 / NeurIPS 3.95), surpassing the next best baseline (4.68 / 3.72) by a large margin, and maintaining consistent advantages in terms of Elo-based ranking metrics (Avg Wins 4.19, Top-10 Count 47). These results validate that EvoSci achieves superior overall research quality and more reliable relative performance compared to strong baselines. In summary, the main contributions of our work are as follows:

- We conceptualize scientific discovery as a problem-oriented process, in which research problems are dynamically generated and progressively refined through a multi-agent collaboration loop.
- We construct a heterogeneous multi-agent framework that mirrors real-world research laboratories, where diverse agents operate under role-specific objectives. The framework is grounded on datasets derived from real scientists, enabling more authentic simulation of collaborative scientific workflows.
- We implement a multi-round feedback with bio-inspired evolution mechanism (selection, crossover, mutation) to enable continuous and open-ended scientific exploration.

2 Related Work

Our work is related to both AI-driven scientific discovery and multi-agent systems. We briefly review these two topics within the scope of LLM and the constraint of space.

2.1 AI for Scientific Discovery

Recent advances in LLMs have enabled AI systems to participate more deeply in scientific discovery (Zheng et al., 2025; Xiong et al., 2026). Early systems primarily assist with literature mining (Smalheiser and Swanson, 1998; Hristovski et al., 2005) and experiment design (King et al., 2009; Tian et al., 2021), while recent developments aim at more comprehensive support for the scientific process. Notably, SciPIP employs three retrieval strategies (semantic, entity, and co-occurrence) to enhance hypothesis generation (Wang et al., 2024b), and the MOOSE framework introduces iterative feedback mechanisms to evolve hypotheses from large-scale web corpora (Yang et al., 2024). Building on these advances, SciAgents integrates multi-agent reasoning with scientific knowledge graphs to autonomously generate, test, and refine hypotheses (Ghafarollahi and Buehler, 2025), while CoScientist further extends such capabilities by autonomously planning and executing experimental procedures within chemical research workflows (Jansen et al., 2025).

Recent systems have taken a more ambitious step toward end-to-end scientific discovery. For example, AI-Scientist (Lu et al., 2024; Yamada et al., 2025) and CycleResearcher (Weng et al., 2024) automate nearly the entire scientific workflow, from topic formulation and literature exploration to hypothesis generation, experiment simulation, paper writing, and even peer review. However, existing approaches primarily focus on conducting a single round of research under a fixed initial topic, overlooking the evolutionary and cyclic nature of scientific discovery. This gap highlights the need for frameworks that support iterative refinement and problem reformulation across successive research cycles.

2.2 Collaboration in Multi-Agent Systems

Recent work has increasingly turned to multi-agent systems (MAS) as a way to mitigate the limitations of single LLMs, such as hallucination (Huang et al., 2025), weak long-horizon reasoning (Ferrag et al., 2025), and stale knowledge (Zhang et al., 2023). MAS build on the paradigm of agentic AI (Durante et al., 2024), organizing multiple LLM-based agents with specialized roles and shared goals. Through coordinated planning, division of labor, and mutual evaluation, these systems aim to achieve more reliable and scalable cognitive performance than any individual model (Zhou et al., 2024a; Li et al., 2025).

A growing line of studies applies MAS to scientific discovery, demonstrating their potential in iterative and multi-step research workflows. VIRSCI

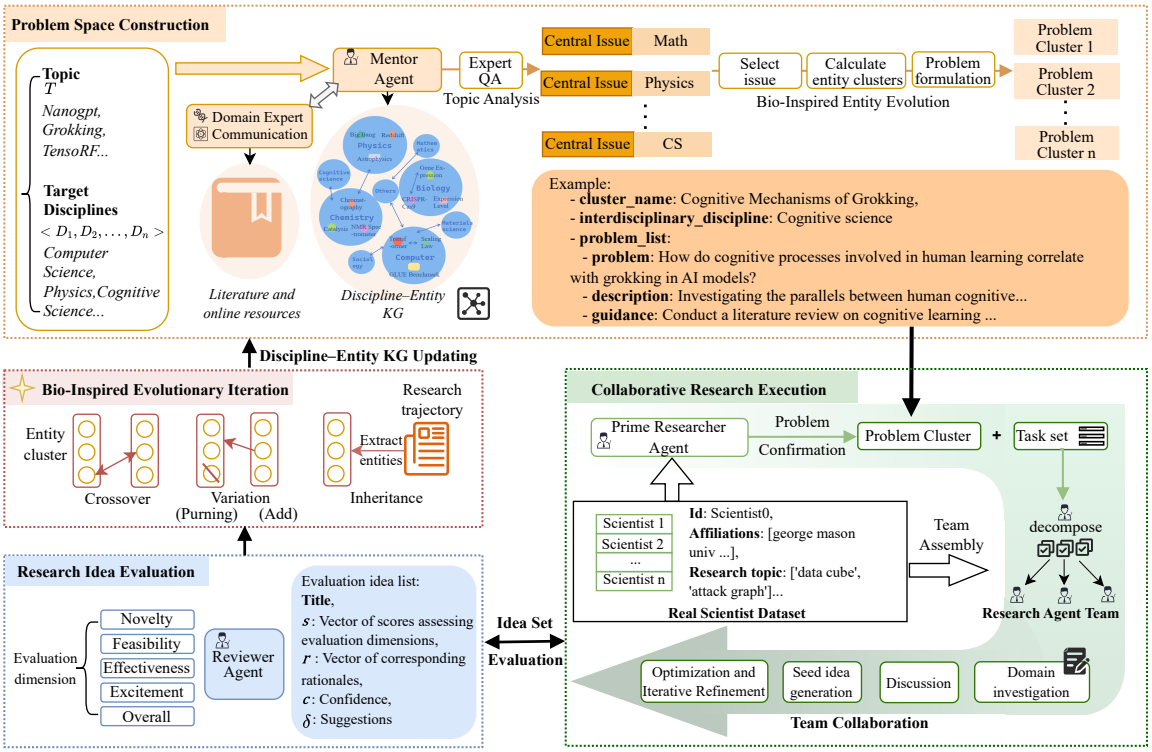


Figure 1: Overall workflow of the proposed EvoSci framework. EvoSci begins with problem space construction from literature and domain knowledge, followed by collaborative research execution through role-based agents and iterative evaluation with reviewer feedback. The bio-inspired evolutionary loop operates over multiple rounds, leveraging feedback to recombine, adapt, and refine research directions.

simulates a virtual team of scientists engaging in structured idea generation and evaluation (Su et al., 2025), while ResearchAgent coordinates specialized agents for literature analysis, hypothesis generation, and experiment planning (Baek et al., 2024). These works largely remain task- or pipeline-oriented, and fall short of providing an integrated framework that supports cyclic scientific evolution together with role-aware, interdisciplinary collaboration.

3 EvoSci

EvoSci models scientific discovery as an evolutionary, problem-centric process, where research directions are iteratively explored, evaluated, and refined over long horizons. Accordingly, EvoSci consists of four core components: (1) a problem space construction module guided by a mentor agent, (2) a collaborative research execution module led by a prime researcher agent, (3) an evaluation module that systematically assesses the quality, novelty, and feasibility of generated research ideas, and (4) a bio-inspired evolutionary iteration module that enables iterative refinement of research directions. Together, these components form a closed-loop workflow for the iterative generation and refinement of interdisciplinary research ideas. Figure 1 illustrates the overall workflow of EvoSci.

3.1 Problem Space Construction

A core challenge in automating scientific idea generation lies in the construction of a structured, high-quality problem space that supports exploration across disciplinary boundaries. This phase aims to transform an initial research theme into a diverse collection of research problems that are both semantically grounded and structurally expandable.

Data Preparation. We construct a lightweight, multi-level knowledge graph to organize scientific disciplines and their associated entities. We begin with a predefined set of representative disciplines spanning major scientific disciplines (e.g., Physics, Chemistry, Biology, Medicine, Economics), each represented as a first-layer node.

For each discipline, we extract candidate entities from its Wikipedia page using the page summary and hyperlink structure. An LLM-based classifier assigns each entity a semantic type (e.g., *Theory*, *Model*, *Material*, *Phenomenon*) and estimates its relevance to the discipline. Only relevant entities are retained and connected to their corresponding disciplines via *has_entity* edges.

To capture cross-disciplinary connections, we compute cosine similarity between the embedding representations of entities and add a cross-entity edge if the similarity exceeds a threshold, i.e., $s(e_i, e_j) = \cos(\text{emb}(e_i), \text{emb}(e_j)) > \tau$.

Topic Analysis. Given a core research topic T and a set of target disciplines $\mathcal{D}_{\text{target}} = \{D_1, \ldots, D_n\}$, the system first grounds the topic by using an LLM-based classifier to map T to one or more core disciplines in the knowledge graph, ensuring that subsequent exploration is anchored in a clear scientific context. To encourage cross-disciplinary integration, we adopt a question-answering architecture with domain expert agents instantiated from a dataset of real-world scientists (Appendix A.1). Each expert agent is equipped with literature retrieval and reading capabilities and engages in structured discussions with a mentor agent. Through this process, the system identifies promising interdisciplinary directions and updates the knowledge graph with domain-specific entities derived from their exploration trajectories, enabling iterative evolution to support downstream idea generation.

Bio-Inspired Entity Evolution. After the topic analysis concludes, the system enters the problem generation stage by focusing on the intersection between the topic T and a selected discipline $d \in \mathcal{D}_{\text{target}}$. For discipline d, $\mathcal{E}_d$ denotes the set of associated entities in the knowledge graph, which are clustered into semantic entity clusters $C_d = \{C_{d,1}, \ldots, C_{d,n}\}$. The most topic-relevant cluster is then explicitly incorporated into the prompt of the mentor agent, which uses these entity clusters as contextual cues to generate scientific research problems:

$$Q_{\text{problem}} = \langle T,\ d,\ \text{Top}(C_d; T) \rangle, \qquad (1)$$

where $\text{Top}(C_d; T)$ denotes the single entity cluster in C_d that is most semantically relevant to the topic T. After each exploration round, bio-inspired evolutionary operations (i.e., *Crossover*, *Variation*, *Inheritance*, and *Selection*) are applied at the cluster level to the discipline entity space, as detailed in Section 3.4.

Structured Problem Cluster Generation. Driven by LLMs, the system generates a diverse set of candidate scientific problems $\mathcal{Q} = \{q_1, q_2, \ldots\}$ by expanding along the evolving entity clusters in the knowledge graph. Each problem q_i is represented by a concise problem statement $\mathcal{P}(q_i)$, an explanatory description $\mathcal{D}(q_i)$, and a research guidance field $\mathcal{G}(q_i)$.

The generated problems are subsequently grouped into problem clusters $\mathcal{P} = \{\mathcal{P}_1, \mathcal{P}_2, \ldots\}$ according to their primary interdisciplinary foci, forming structured problem sets that emphasize cross-disciplinary perspectives and conceptual novelty.

3.2 Collaborative Research Execution

Building on the structured problem clusters, the system proceeds to a multi-agent research exploration phase. A set of role-specialized agents collaboratively investigates selected problem clusters through literature review, structured discussion, and iterative idea generation and refinement, simulating the workflow of real scientific teams.

Problem Confirmation and Team Assembly. From the candidate problem clusters $\mathcal{P}$, the system selects a target cluster $\mathcal{P}^*$ by jointly considering its relevance to the initial topic, interdisciplinary potential, and future extensibility. A research team is then assembled to explore $\mathcal{P}^*$. The prime researcher and assistant researchers are instantiated from a real-world scientist dataset, where each agent is represented by anonymized metadata and semantic behavior embeddings. The prime researcher is selected to align with the initial topic, while assistant researchers are chosen based on their relevance to $\mathcal{P}^*$, facilitating effective interdisciplinary collaboration.

Task Decomposition and Team Collaboration. Given a selected problem cluster P^*, the Prime Researcher organizes a collaborative research process by decomposing the exploration into a set of structured tasks $\mathcal{T} = \{t_1, t_2, \ldots\}$. These tasks correspond to key stages of scientific inquiry, including background investigation, problem analysis, idea generation, and iterative refinement. We adopt the CrewAI framework to organize the agent team and establish a dynamic delegation mechanism based on a "lead-and-collaborate" interaction paradigm, which supports the following collaboration processes:

- **Task Leading and Delegation:** Each core task is initiated and led by the Prime Researcher, who analyzes the task context, decomposes it into subtasks, and assigns them to assistant agents based on semantic relevance and skill profiles.
- **Recursive Delegation and Collaboration:** Assistant agents may further decompose assigned subtasks and reassign them when necessary,

resulting in a recursive collaboration structure.

- **Phased Integration:** At predefined stages of task execution, structured discussion rounds are conducted to aggregate intermediate results, align perspectives across agents, and integrate subtask outcomes.

Formally, for a task t_k, we define the research state and prompt as:

$$\mathcal{S}_k = \langle\, t_{\text{description}},\ \bigcup_{i=1}^{k-1} \mathcal{R}_i,\ \mathcal{R}_k^{\text{sub}},\ a_k \,\rangle, \quad (2)$$

$$Q_{\text{research}} = \langle\, T,\ P^*,\ \mathcal{S}_k \,\rangle, \quad (3)$$

where $t_{\text{description}}$ denotes the description of the current task or subtask under exploration, $\bigcup_{i=1}^{k-1} \mathcal{R}_i$ represents the aggregated responses accumulated from previously completed tasks, $\mathcal{R}_k^{\text{sub}}$ denotes the intermediate responses produced for subtasks under the current task t_k, and a_k denotes the agent assigned to execute t_k. The research prompt Q_{research} combines the initial topic T, the selected problem cluster P^*, and the current research state $\mathcal{S}_k$ to guide the task execution. In practice, the accumulated responses in $\mathcal{S}_k$ are maintained through a hierarchical memory mechanism, including short-term memory, long-term memory, and entity memory, which enables compact context representation and more precise summarization across tasks.

Seed Idea Generation and Refinement. Through domain investigation and collaborative discussion, the system generates diverse seed ideas, broadening the exploration space and increasing the likelihood of high-quality discoveries. These ideas are refined through multi-agent collaboration, where redundant or low-value ideas are removed, the remaining ideas are ranked by novelty, feasibility, and interdisciplinary value.

3.3 Research Idea Evaluation

To further refine generated ideas, a reviewer agent is engaged to simulate a peer-review process. Each generated idea $\mathcal{I} = \{I_1, I_2, \dots\}$ is evaluated along multiple dimensions, including novelty, feasibility, validity, and scientific excitement. In addition to quantitative scores, the reviewer agent provides structured feedback by identifying logical weaknesses and suggesting complementary experimental directions.

Formally, the evaluation results can be expressed as:

$$E(\mathcal{I}) = \langle\, \text{title},\ s,\ r,\ c,\ \delta \,\rangle, \quad (4)$$

where title denotes the title of the evaluated idea, $s = (s_{\text{nov}}, s_{\text{fea}}, s_{\text{eff}}, s_{\text{exc}}, s_{\text{overall}})$ is a vector of scores assessing novelty, feasibility, expected effectiveness, scientific excitement and overall, and $r = (r_{\text{nov}}, r_{\text{fea}}, r_{\text{eff}}, r_{\text{exc}}, r_{\text{overall}})$ denotes the corresponding rationales. c denotes the reviewer's confidence level, and δ denotes concrete suggestions for improving the idea.

3.4 Bio-Inspired Evolutionary Iteration

Scientific idea generation is inherently iterative rather than one-shot. To support long-term and open-ended discovery, EvoSci introduces a bio-inspired evolutionary loop that operates on the entity layer of the knowledge graph and is guided by structured evaluation feedback.

Entity-Level Evolution. Building on the multi-level knowledge graph, the discipline layer remains static as a structural backbone, while the entity layer is treated as an evolving population. For each discipline, entities are organized into semantic clusters, which serve as the basic units of evolution. Across successive exploration rounds, these entity clusters are iteratively updated to refine cross-disciplinary exploration.

Concretely, the system applies a set of bio-inspired evolutionary operations at the cluster level.

- **Crossover:** Enables recombination by exchanging entities between different semantic clusters within the same discipline, producing novel concept combinations while preserving domain coherence.
- **Variation:** Injects diversity by introducing new or low-frequency entities into existing clusters, preventing premature convergence.
- **Selection:** Filters entity clusters based on evaluation feedback from generated ideas, favoring clusters that exhibit higher novelty, feasibility, and relevance to the research topic.
- **Inheritance:** Propagates high-fitness entities and clusters into subsequent iterations, ensuring that valuable knowledge accumulates over time.

Evaluation-Guided Loop. After each exploration round, refined ideas and their evaluations are summarized by the Prime Researcher and passed to the Mentor Agent. High-fitness entities identified from successful ideas are re-integrated into the knowledge graph, while low-value entities are pruned. The updated entity clusters then serve as seeds for reconstructing the next problem set, forming a closed evolutionary loop that balances exploration

and consolidation.

4 Experiments

We designed an experimental setup to examine EvoSci's adaptability across diverse scientific domains. Ten representative and challenging open research topics were selected from the task settings introduced in AI Scientist (Lu et al., 2024). Detailed task settings and corresponding experimental prompts are provided in Appendices B and F.

4.1 Evaluation Methodologies

To comprehensively assess the effectiveness and creativity of our multi-agent scientific system, we adopted both qualitative and quantitative evaluations, combining an expert-simulated review mechanism with a tournament-style idea ranking procedure. In addition, we conducted an additional experiment to further validate the effectiveness and stability of the proposed meta-review mechanism (see Appendix E).

Multi-Reviewer + Meta-Reviewer Mechanism. Inspired by the AI Scientist evaluation framework (Lu et al., 2024), we designed a structured LLM-driven peer-review workflow that simulates academic conference reviewing. Reviewer agents independently assess generated ideas using prompts aligned with ICLR and NeurIPS review templates, each incorporating a reflection mechanism to refine their evaluations. A meta-reviewer agent then aggregates individual reviews into a unified meta-review, enabling interpretable, reproducible, and academically representative evaluation.

Tournament-Style Idea Ranking. In addition, we implemented a comparative ranking procedure to evaluate idea quality. All generated ideas were pooled, randomized, and initially assigned one point. Ideas were paired and compared using the prompt: *"One of them is accepted by a top AI conference (like ICLR or ACL) and the other one is rejected."* The winner of each comparison received one point. This process was repeated for five rounds, and the final ranking was determined by aggregated scores. This tournament-style evaluation provided a robust relative measure of idea quality, complementing the structured review mechanism. Prior studies have also shown that such pairwise comparison, rather than absolute scoring, enables LLM-based evaluations to better align with human expert judgments.

4.2 Main Results

To systematically verify the overall effectiveness of EvoSci, we conducted experiments on the ten representative research topics described above. Our system was compared against four baseline methods: SciPIP, AI Scientist, COI Agent and VirSci (detailed descriptions are provided in the Appendix C). For fairness, each method was configured to generate 10 research ideas per topic, following comparable settings on iteration rounds, team size, and evaluation feedback. The generated ideas were evaluated using two mechanisms: an expert-simulated review and a tournament-style ranking. The aggregated results are reported in Table 1, while the tournament-style ranking results are reported in Table 2.

From Table 1, we observe that EvoSci consistently outperforms all baselines across models and evaluation dimensions. It achieves the highest scores on *Validity*, *Excitement*, and both overall metrics, indicating that the generated research ideas are not only credible but also engaging. While baselines such as AI Scientist or VirSci exhibit isolated strengths on individual criteria, their overall performance remains uneven, whereas our framework provides balanced improvements across all aspects. On the two overall measures, EvoSci yields 5–10% relative gains in ICLR Overall (e.g., 4.90 vs. 4.68 with DeepSeek-v3, 4.72 vs. 4.54 with Qwen3-max) and 10–15% gains in NeurIPS Overall (e.g., 3.95 vs. 3.39 with DeepSeek-v3, 3.81 vs. 3.31 with Qwen3-max). These advantages are more pronounced when paired with stronger backbone models.

The complementary tournament-style evaluation further corroborates these findings (Table 2). EvoSci obtains the highest *Avg Wins* across all backbones (4.27 with GPT-4o, 4.19 with DeepSeek-v3, 4.25 with Qwen3-max) and the largest *Top 10 Count*. Together, these results demonstrate that our framework reliably produces more credible, exciting, and impactful research ideas than existing baselines.

4.3 Ablation Study

To further examine the contribution of individual components within the proposed framework, we designed three ablation experiments.

Impact of Structured Problem Formulation. To evaluate the effect of the problem formulation module, we compared two settings:

(1) *w/ Problem Guidance*: The full system, where the Mentor Agent actively interprets the initial

Table 1: Subjective evaluation scores of various agent models and methods.

Agent Model	Method	Novelty	Feasibility	Validity	Excitement	ICLR Overall	NeurIPS Overall
GPT-4o	AI Scientist	4.31	**6.49**	4.87	4.11	4.33	3.19
	SciPIP	4.72	4.76	4.74	4.49	4.26	3.06
	VirSci	**5.12**	3.64	4.56	**4.95**	4.28	3.26
	CoI-Agent	4.52	4.77	4.79	4.33	4.20	3.21
	EvoSci	4.78	4.62	**5.01**	4.75	**4.45**	**3.44**
DeepSeek-v3	AI Scientist	4.60	**6.85**	5.00	4.29	4.68	3.39
	SciPIP	4.42	5.80	4.85	4.13	4.34	3.02
	VirSci	5.48	3.95	4.88	5.11	4.50	3.69
	CoI-Agent	5.07	4.66	4.92	4.58	4.51	3.72
	EvoSci	**5.71**	4.68	**5.25**	**5.15**	**4.90**	**3.95**
Qwen3-max	AI Scientist	4.18	**7.00**	5.04	4.15	4.48	3.31
	SciPIP	4.87	5.41	4.89	4.53	4.54	3.17
	VirSci	**5.37**	4.01	4.90	**5.00**	4.35	3.57
	CoI-Agent	4.64	4.38	4.76	4.40	4.19	3.62
	EvoSci	5.14	4.98	**5.20**	4.89	**4.72**	**3.81**

Table 2: Average wins and top 10 counts for various agent models and methods.

Agent Model	Method	Avg Wins	Top 10 Count
GPT-4o	AI Scientist	3.88	13
	SciPIP	2.70	7
	VirSci	4.07	52
	CoI-Agent	3.58	36
	EvoSci	**4.27**	**54**
DeepSeek-v3	AI Scientist	2.83	8
	SciPIP	2.50	3
	VirSci	3.90	35
	CoI-Agent	4.08	37
	EvoSci	**4.19**	**47**
Qwen3-max	AI Scientist	2.92	7
	SciPIP	2.39	1
	VirSci	3.94	34
	CoI-Agent	4.00	39
	EvoSci	**4.25**	**50**

keyword, retrieves relevant literature, constructs a structured problem space, and generates problem clusters to guide downstream research;

(2) *w/o Problem Guidance*: A simplified version using the raw prompt template from AI Scientist without problem construction, where the Prime Researcher directly explores the given topic.

Each setting generates 10 research ideas for each of the ten benchmark topics (100 ideas in total), following the unified evaluation protocol.

As shown in Table 3, the system with explicit problem construction achieves higher scores across most dimensions, particularly in *novelty* and *excitement*, indicating that structured problem formulation effectively focuses the research direction while broadening the exploration boundary. This demonstrates its essential role in establishing semantic anchoring and path guidance for automated scientific discovery.

Role of Multi-agent Collaboration. To evaluate the effect of multi-agent collaboration, we compare systems with different team sizes: a single-agent setting (*team_size = 1*), a standard collaborative

Table 3: The impact of problem formulation on research quality.

Setting	Novelty	Feasibility	Validity	Excitement	ICLR Overall	NeurIPS Overall
w/ Problem Guidance	**4.78**	4.62	**5.01**	**4.75**	**4.45**	**3.44**
w/o Problem Guidance	4.22	**4.75**	4.96	4.51	4.22	3.28

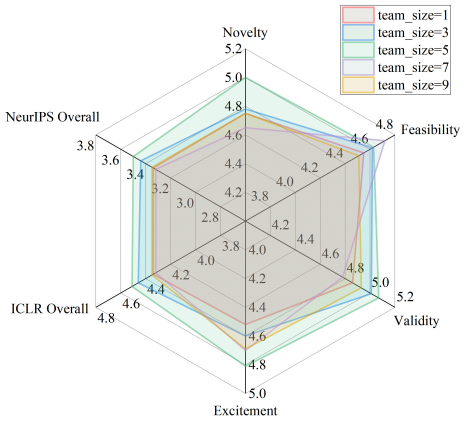


Figure 2: The impact of team size on research quality.

setting (*team_size = 3*), and extended collaborative settings (*team_size = 5*, *team_size = 7*, and *team_size = 9*).

As shown in Figure 2, increasing team size initially yields consistent improvements across novelty, validity, and overall quality, with performance peaking at team_size = 5. Beyond this point, however, we observe a clear degradation at team_size = 7 and 9. This pattern suggests that while moderate levels of agent diversity enhance idea generation and evaluation, excessively large teams incur substantial coordination overhead and conflicting reasoning trajectories. Consequently, the marginal utility of additional agents becomes negative, indicating that optimal research performance emerges from moderately sized, rather than maximal teams.

Table 4: Evolution ablation results. We report topic-level average scores under NeurIPS-style and ICLR-style review templates, with and without the evolutionary module.

Topic	NeurIPS (w/o Evo)	NeurIPS (w/ Evo)	ICLR (w/o Evo)	ICLR (w/ Evo)
2D Diffusion	3.40	3.32	4.56	4.50
Character-Level Language Modeling	3.48	3.52	4.50	4.46
Earthquake Prediction	3.40	3.54	4.42	4.44
Grokking	3.26	3.22	4.16	4.08
NanoGPT	3.44	3.66	4.40	4.66
Materials Adaptive Convolutional Equivariants	3.30	3.16	4.20	4.10
SEIR Infection Modeling	3.38	3.36	4.38	4.42
Sketch Generation with RNNs	3.40	3.42	4.34	4.52
Multi-Dataset CNN Optimization	3.32	3.48	4.16	4.28
TensoRF	3.42	3.56	4.22	4.18
Average	**3.38**	**3.424**	**4.334**	**4.364**

Effect of Evolutionary Iteration. To evaluate the effect of the evolutionary module, we compared two settings:

(1) *w/ Evo*: the full system with bio-inspired evolutionary iteration enabled;

(2) *w/o Evo*: a simplified version in which the evolutionary module is disabled, while all other components and evaluation settings remain unchanged.

As shown in Table 4, enabling evolution shifts the topic-level scores upward in the majority of cases under both evaluation templates. Under the NeurIPS-style review, most topics exhibit improvements when evolution is enabled, and a similar trend is observed under the ICLR-style template. Although not every individual topic increases, the overall direction across topics is predominantly positive. Aggregating across topics (see the Average row), enabling evolution yields a consistent increase in overall performance under identical computational budgets and evaluation settings (NeurIPS: 3.38 → 3.424; ICLR: 4.334 → 4.364). Although the absolute gains are moderate, they are systematic rather than driven by isolated instances, indicating that improvements cannot be solely attributed to iterative feedback or memory accumulation.

Table 5 further summarizes the aggregated comparison. In addition to the consistent mean shifts, we examine within-topic variability across five independent runs per topic. Under the NeurIPS-style evaluation, enabling evolution leads to a moderate increase in within-topic standard deviation (0.146 → 0.176), suggesting that the evolutionary operators introduce stronger exploration dynamics rather than simply repeating deterministic refinement steps. Under the ICLR-style template, within-topic variability remains largely comparable (0.154 → 0.157). This suggests that different evaluation templates may exhibit varying sensitivity to quality differences, rather than indicating inconsistent behavior of the evolution model itself. Taken together, these ablations isolate two primary drivers of improvement, namely workflow-level role specialization and evolutionary search, and clarify how each component contributes to performance gains.

Table 5: Overall comparison between w/o and w/ evolution.

Metric	w/o Evo	w/ Evo	Δ (w/ – w/o)
NeurIPS Template			
Mean	3.380	3.424	+0.044
Within-topic Std (avg over topics)	0.146	0.176	+0.030
ICLR Template			
Mean	4.334	4.364	+0.030
Within-topic Std (avg over topics)	0.154	0.157	+0.003

5 Analysis

To better understand the behavior of EvoSci beyond primary performance results, we conducted a set of additional *analytical experiments* that probed the system from complementary perspectives. More detailed experimental analyses and the corresponding quantitative results for each topic are reported in Appendix D.

Evolution of Ideas under Iterative Evaluation. To analyze the effect of evaluation-guided iteration on idea evolution, we conducted a qualitative analysis on *Grokking* by tracking how generated ideas changed across feedback rounds. For each iteration, we extracted technical terms from the generated ideas and visualized their cumulative growth, to-

gether with changes in semantic organization, as shown in Figure 3 (Appendix D.1), to characterize the evolution of conceptual structure over time.

We observe that early iterations introduce diverse learning-related concepts with limited internal organization, while later iterations progressively form clearer and more structured conceptual clusters. Over time, ideas increasingly emphasize delayed generalization and representation reorganization, resulting in more coherent descriptions of grokking-like behavior.

Quality Gains beyond the Initial Exploration Stage. To examine whether idea quality improved beyond the initial exploratory phase, we analyzed the evolution of quality scores across iterative rounds. For each topic, EvoSci performed ten iterations and generated a total of 50 ideas. Ideas were grouped into batches of 10, and quality scores were computed at the group level to track changes over time, as shown in Table 6 (Appendix D.2). We treated the first group as an exploratory baseline and identified the earliest subsequent iteration that exhibited noticeable improvement.

We observe that quality gains beyond the first group occurred in most topics, although the timing and affected metrics vary. For topics with latent mechanisms or less explored conceptual structures, later groups show clear improvements in Novelty and/or Overall scores. In contrast, engineering-oriented or well-studied topics exhibit limited or no improvement beyond early groups. Across topics, improvements do not appear simultaneously across all metrics, with some groups showing gains in Novelty without corresponding increases in Overall scores, or vice versa.

Exploration Dynamics Across and Within Iterations. To analyze exploration behavior beyond aggregate quality trends, we measured idea similarity from two perspectives: intra-round convergence and inter-round continuity. Within each iteration, we computed the average pairwise cosine similarity between sentence embeddings of ideas to quantify intra-round convergence. Across iterations, we computed inter-round similarity by measuring the cosine similarity between aggregated embeddings of ideas from consecutive rounds, as shown in Figure 4 (Appendix D.3).

We observe substantial variation in both intra-round and inter-round similarity patterns across topics. Some topics exhibit consistently higher intra-round similarity together with strong inter-round continuity, indicating concentrated exploration and stable refinement across iterations. Other topics show lower or more fluctuating intra-round similarity and greater inter-round variation, reflecting broader exploration and frequent shifts in focus. Overall, intra-round convergence and inter-round continuity exhibit aligned trends across topics, with topics showing stronger within-round concentration also tending to display higher continuity across iterations.

Grounding the Evolutionary Dynamics. To examine whether the evolutionary process in EvoSci reflects concrete system behavior rather than a high-level analogy, we conducted an additional qualitative analysis on the *Grokking* topic by tracing how ideas evolved across iterative rounds. We find that the system maintains persistent conceptual variation over time, while evaluation feedback gradually favors mechanisms that are more directly relevant to grokking.

Concretely, the search moves from broad learning-related exploration toward more specialized concepts such as delayed generalization, phase transitions, and representation reorganization, while still retaining multiple conceptual lineages in parallel. This pattern suggests that the evolutionary process in EvoSci is not merely metaphorical, but is grounded in heritable variation, feedback-guided selection, and population diversity.

6 Conclusion

In this study, we have presented EvoSci, a multi-agent, feedback-driven, and bio-inspired evolutionary framework for automated scientific discovery. The framework conceptualizes scientific discovery as a problem-oriented process, integrates heterogeneous research agents that emulate real-world laboratory roles, and employs multi-round feedback with evolutionary operations to support continuous and open-ended exploration. Extensive experiments across ten scientific domains show that EvoSci consistently outperforms strong baselines in idea validity, excitement, and overall quality. The gains are balanced across evaluation dimensions and remain robust across backbone models, validating feedback-guided evolutionary exploration for open-ended scientific discovery.

Limitations

Experimental results across ten interdisciplinary research topics demonstrate that EvoSci generates more novel and insightful research ideas than

existing systems, showing particular strength in innovation-oriented dimensions such as *novelty*, *excitement*, and overall peer-review evaluations. However, due to its broad cross-domain exploration, the framework sometimes produces ideas with lower practical feasibility, suggesting a trade-off between creativity and applicability.

Future work will focus on enhancing EvoSci's ability to reason and operate across disciplines by improving interdisciplinary knowledge integration through structured knowledge representations, strengthening causal reasoning to increase scientific rigor and interpretability, and developing more open-ended iterative mechanisms that enable long-term, autonomous scientific discovery. A key challenge in achieving such continuous evolution lies in establishing more objective and high-quality evaluation mechanisms that allow LLM-based agents to better assess their own reasoning and outputs, which is essential for truly effective self-improvement and sustained innovation.

Acknowledgments

The present research was supported by the National Key Research and Development Program of China (Grant No. 2024YFE0203000). We also acknowledge support from the State Key Laboratory of Tibetan Intelligence (Grant No. 2025-ZJ-J08) and the Postdoctoral Fellowship Program of CPSF (Grant No. GZC20251075). We thank the anonymous reviewers for their insightful comments.

References

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 2024. Researchagent: Iterative research idea generation over scientific literature with large language models. *arXiv preprint arXiv:2404.07738*.

Markus J Buehler. 2024. Accelerating scientific discovery with generative knowledge extraction, graph-based representation, and multimodal intelligent graph reasoning. *Machine Learning: Science and Technology*, 5(3):035083.

Gary Charness, Brian Jabarian, and John A. List. 2025. The next generation of experimental research with LLMs. *Nature Human Behaviour*, 9(5):833–835.

Saaketh Desai, Sadhvikas Addamane, Jeffrey Y Tsao, Igal Brener, Laura P Swiler, Remi Dingreville, and Prasad P Iyer. 2025. Autoscilab: A self-driving laboratory for interpretable scientific discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 146–154.

Zane Durante, Qiuyuan Huang, Naoki Wake, Ran Gong, Jae Sung Park, Bidipta Sarkar, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Yejin Choi, and 1 others. 2024. Agent ai: Surveying the horizons of multimodal interaction. *arXiv preprint arXiv:2401.03568*.

Kevin C. Elliott. 2012. Epistemic and methodological iteration in scientific research. *Studies in History and Philosophy of Science Part A*, 43(2):376–382.

Mohamed Amine Ferrag, Norbert Tihanyi, and Merouane Debbah. 2025. Reasoning beyond limits: Advances and open problems for llms. *Preprint*, arXiv:2503.22732.

Alireza Ghafarollahi and Markus J Buehler. 2025. Sciagents: automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. *Advanced Materials*, 37(22):2413523.

Longzhen Han, Awes Mubarak, Almas Baimagambetov, Nikolaos Polatidis, and Thar Baker. 2025. A survey of generative categories and techniques in multimodal large language models. *Preprint*, arXiv:2506.10016.

Xinyi He, Jiaru Zou, Yun Lin, Mengyu Zhou, Shi Han, Zejian Yuan, and Dongmei Zhang. 2024. CoCoST: Automatic complex code generation with online searching and correctness testing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19433–19451, Miami, Florida, USA. Association for Computational Linguistics.

Dimitar Hristovski, Borut Peterlin, Joyce A. Mitchell, and Susanne M. Humphrey. 2005. Using literature-based discovery to identify disease candidate genes. *International Journal of Medical Informatics*, 74(2):289–298. MIE 2003.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Peter Jansen, Oyvind Tafjord, Marissa Radensky, Pao Siangliulue, Tom Hope, Bhavana Dalvi Mishra, Bodhisattwa Prasad Majumder, Daniel S Weld, and Peter Clark. 2025. Codescientist: End-to-end semi-automated scientific discovery with code-based experimentation. *arXiv preprint arXiv:2503.22708*.

Zixuan Ke, Fangkai Jiao, Yifei Ming, Xuan-Phi Nguyen, Austin Xu, Do Xuan Long, Minzhi Li, Chengwei Qin, Peifeng Wang, Silvio Savarese, and 1 others. 2025. A survey of frontiers in llm reasoning: Inference scaling, learning to reason, and agentic systems. *arXiv preprint arXiv:2504.09037*.

Ross D. King, Jem Rowland, Stephen G. Oliver, Michael Young, Wayne Aubrey, Emma Byrne, Maria Liakata, Magdalena Markham, Pinar Pir, Larisa N. Soldatova, Andrew Sparkes, Kenneth E. Whelan, and Amanda

Clare. 2009. The automation of science. *Science*, 324(5923):85–89.

Adithya Kulkarni, Fatimah Alotaibi, Xinyue Zeng, Longfeng Wu, Tong Zeng, Barry Menglong Yao, Minqian Liu, Shuaicheng Zhang, Lifu Huang, and Dawei Zhou. 2025. Scientific hypothesis generation and validation: Methods, datasets, and future directions. *Preprint*, arXiv:2505.04651.

Jia Li, Yongmin Li, Ge Li, Zhi Jin, Yiyang Hao, and Xing Hu. 2023. Skcoder: A sketch-based approach for automatic code generation. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, pages 2124–2135. IEEE.

Zhigen Li, Jianxiang Peng, Yanmeng Wang, Yong Cao, Tianhao Shen, Minghui Zhang, Linxi Su, Shang Wu, Yihang Wu, YuQian Wang, Ye Wang, Wei Hu, Jianfeng Li, Shaojun Wang, Jing Xiao, and Deyi Xiong. 2025. ChatSOP: An SOP-guided MCTS planning framework for controllable LLM dialogue agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17637–17659, Vienna, Austria. Association for Computational Linguistics.

Yan Liu, Minghui Zhang, Bojian Xiong, Yifan Xiao, Yinong Sun, Yating Mei, Longyu Zeng, Jingchao Yang, Yang Wang, and Deyi Xiong. 2025. HighMATH: Evaluating math reasoning of large language models in breadth and depth. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10241–10253, Suzhou, China. Association for Computational Linguistics.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*.

Staša Milojević. 2014. Principles of scientific research team formation and evolution. *Proceedings of the National Academy of Sciences*, 111(11):3984–3989.

Juran Noh, Hieu A Doan, Heather Job, Lily A Robertson, Lu Zhang, Rajeev S Assary, Karl Mueller, Vijayakumar Murugesan, and Yangang Liang. 2024. An integrated high-throughput robotic platform and active learning approach for accelerated discovery of optimal electrolyte formulations. *Nature Communications*, 15(1):2757.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599.

Xiaoxue Ren, Xinyuan Ye, Dehai Zhao, Zhenchang Xing, and Xiaohu Yang. 2023. From misuse to mastery: Enhancing code generation with knowledge-driven ai chaining. In *2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 976–987. IEEE.

Neil R Smalheiser and Don R Swanson. 1998. Using arrowsmith: a computer-assisted approach to formulating and assessing scientific hypotheses. *Computer Methods and Programs in Biomedicine*, 57(3):149–153.

Haoyang Su, Renqi Chen, Shixiang Tang, Zhenfei Yin, Xinzhe Zheng, Jinzhe Li, Biqing Qi, Qi Wu, Hui Li, Wanli Ouyang, Philip Torr, Bowen Zhou, and Nanqing Dong. 2025. Many heads are better than one: Improved scientific idea generation by a LLM-based multi-agent system. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28201–28240, Vienna, Austria. Association for Computational Linguistics.

Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. 2024. Table meets llm: Can large language models understand structured table data? a benchmark and empirical study. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 645–654.

Yunsheng Tian, Mina Konaković Luković, Timothy Erps, Michael Foshey, and Wojciech Matusik. 2021. Autooed: Automated optimal experiment design platform. *Preprint*, arXiv:2104.05959.

Gary Tom, Stefan P Schmid, Sterling G Baird, Yang Cao, Kourosh Darvish, Han Hao, Stanley Lo, Sergio Pablo-García, Ella M Rajaonson, Marta Skreta, and 1 others. 2024. Self-driving laboratories for chemistry and materials science. *Chemical Reviews*, 124(16):9633–9732.

Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. 2024. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482.

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

Wenxiao Wang, Lihui Gu, Liye Zhang, Yunxiang Luo, Yi Dai, Chen Shen, Liang Xie, Binbin Lin, Xiaofei He, and Jieping Ye. 2024b. Scipip: An llm-based scientific paper idea proposer. *arXiv preprint arXiv:2410.23166*.

Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. 2024. Cycleresearcher: Improving automated research via automated review. *arXiv preprint arXiv:2411.00816*.

Xiaoyu Xiong, Hao Wang, Keming Wu, Zhenfei Yang, Hanjie Zhao, Hongxiang Wang, Hao Liu, and Deyi Xiong. 2026. Collaborative and autonomous ai for science and innovation: Practices, challenges, and future directions. *TechRxiv*, 2026(0220).

Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, and 1 others. 2025. Towards large reasoning models: A survey of reinforced reasoning with large language models. *arXiv preprint arXiv:2501.09686*.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. 2025. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*.

Lei Yang, Renren Jin, Ling Shi, Jianxiang Peng, Yue Chen, and Deyi Xiong. 2025. Probench: Benchmarking large language models in competitive programming. *ArXiv*, abs/2502.20868.

Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and Erik Cambria. 2024. Large language models for automated open-domain scientific hypotheses discovery. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 13545–13565.

Shaowei Zhang and Deyi Xiong. 2025. Debate4MATH: Multi-agent debate for fine-grained reasoning in math. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16810–16824, Vienna, Austria. Association for Computational Linguistics.

Zihan Zhang, Meng Fang, Ling Chen, Mohammad-Reza Namazi-Rad, and Jun Wang. 2023. How do large language models capture the ever-changing world knowledge? a review of recent advances. *arXiv preprint arXiv:2310.07343*.

Tianshi Zheng, Zheye Deng, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Zihao Wang, and Yangqiu Song. 2025. From automation to autonomy: A survey on large language models in scientific discovery. *arXiv preprint arXiv:2505.13259*.

Yizhen Zheng, Huan Yee Koh, Jiaxin Ju, Anh TN Nguyen, Lauren T May, Geoffrey I Webb, and Shirui Pan. 2023. Large language models for scientific synthesis, inference and explanation. *arXiv preprint arXiv:2310.07984*.

Hang Zhou, Yehui Tang, Haochen Qin, Yujie Yang, Renren Jin, Deyi Xiong, Kai Han, and Yunhe Wang. 2024a. Star-agents: Automatic data optimization with llm agents for instruction tuning. In *Advances in Neural Information Processing Systems*, volume 37, pages 4575–4597. Curran Associates, Inc.

Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan. 2024b. Hypothesis generation with large language models. In *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*, pages 117–139, Miami, FL, USA. Association for Computational Linguistics.

A Data Collection

A.1 Real-World Scientist Dataset

The Digital Scientist dataset[1] used in this study was constructed by the VirSci team based on real-world scientist information from the AMiner Computer Science dataset,[2] which was originally compiled by extracting researcher profiles from online academic databases. The AMiner Computer Science dataset contains information on 1,712,433 authors and 2,092,356 papers, covering the period from 1948 to 2014 and focusing on the field of computer science.

To ensure data quality, the VirSci team filtered out scientists who had published fewer than 50 papers or had fewer than 50 collaborators. Using the remaining data, they built the Digital Scientist dataset, which includes 156 representative scientists. The profile information of each scientist was embedded using the `mxbai-embed-large` model.

All personal identity information has been anonymized, and the profiles only contain abstracted metadata for research simulation purposes. An example of a digital scientist profile is shown below:

Digital Scientist

Your name is Scientist0, you belong to the following affiliations ['Naval Research Laboratory', 'College of William and Mary', 'George Mason University'], you have researched on the following topics ['data cube', 'attack graph', 'data mining', 'access control', 'data owner', 'data protection', 'data item', 'data redundancy', 'data security', 'data structure'], you have published 372 papers, you have 4230 citations, and you have previously collaborated with these individuals ['Scientist78', 'Scientist105'].

A.2 Technical Terminology Dataset

The Technical Terminology dataset[3] used in this study was compiled to support the analysis of conceptual evolution and the measurement of scientific depth across iterations. It is based on the *Artificial Intelligence Terminology Database*, which systematically collects, organizes, and standardizes key technical terms from major subfields of AI, including machine learning, natural language processing, computer vision, and robotics.

Each term entry contains its English form, corresponding Chinese translation, and a short definition, ensuring cross-lingual consistency and facilitating accurate concept extraction. The dataset enables automated detection and tracking of emerging research topics and specialized vocabulary in scientific idea generation.

B Task Overview

B.1 Task Settings

This appendix provides detailed descriptions of the experimental task settings used to evaluate EvoSci. Following AI Scientist (Lu et al., 2024), we adopt ten representative and challenging open-ended research topics spanning machine learning, scientific modeling, and simulation. These tasks are designed to assess the system's adaptability across diverse scientific domains.

B.2 Task Descriptions

The ten experimental tasks include:

- **2D Diffusion Modeling:** Learning and analyzing diffusion processes in two-dimensional synthetic data.
- **Character-Level Language Modeling:** Training and evaluating character-based language models.
- **Earthquake Prediction:** Modeling seismic activity for temporal event prediction.
- **Grokking:** Investigating delayed generalization behavior in overparameterized networks.
- **NanoGPT:** Training and scaling lightweight transformer-based language models.
- **Materials Adaptive Convolutional Equivariants:** Modeling symmetry-aware representations for material science tasks.
- **SEIR Infection Modeling:** Simulating epidemiological dynamics using compartmental models.
- **Sketch Generation with Recurrent Neural Networks:** Generating hand-drawn sketches using RNN-based generative models.
- **Multi-Dataset CNN Architecture Optimization:** Optimizing small CNN architectures across multiple datasets.
- **TensoRF:** Learning neural radiance fields using tensor factorization.

[1] https://drive.google.com/drive/folders/1ZwWMBQ5oK-l4VuzMa60GbMND0g2EIxIu

[2] https://www.aminer.cn/aminernetwork

[3] https://github.com/jiqizhixin/Artificial-Intelligence-Terminology-Database

B.3 Initialization Prompts

For each task, EvoSci is initialized with a minimal topic prompt describing the research domain. For baseline methods, we incorporate the task descriptions used in AI Scientist (Lu et al., 2024) into their prompts in a consistent and reasonable manner, without introducing additional task-specific heuristics or privileged information. This design ensures fair comparison across different methods.

C Baseline Methods

To comprehensively evaluate the performance of our system, we compare it against four representative research-agent frameworks built upon large language models. For each baseline, we follow the official implementation or publicly described procedure to ensure a fair and consistent comparison. Below, we summarize each method and its configuration used in our experiments.

Baseline 1: SciPIP SciPIP[4], proposed by Zhejiang University, is a research idea generation framework designed to enhance scientific creativity through improved literature retrieval and dual-path reasoning. The system constructs a literature repository enriched with semantic relations, entity links, and citation co-occurrence information, and employs multi-granularity retrieval algorithms to ensure comprehensive coverage of relevant works. During idea generation, SciPIP combines inference from retrieved literature with model-driven brainstorming to produce solutions balancing originality and feasibility. In our experiments, we retain SciPIP's original retrieval mechanism and do not introduce additional ArXiv alignment to maintain consistent comparison.

Baseline 2: AI Scientist AI Scientist[5], developed by Sakana AI, is an automated research platform that aims to cover the end-to-end scientific workflow, including idea formulation, code generation, experiment execution, result analysis, and manuscript drafting. The framework leverages large language models and implements a multi-round reasoning pipeline that iteratively refines research plans. It further includes an automated reviewer module for assessing the quality of generated papers. In this study, we adopt the publicly available multi-step reasoning and research evaluation procedures as a strong baseline to compare scientific idea generation capabilities.

Baseline 3: VirSci VirSci[6], released by the Shanghai Artificial Intelligence Laboratory, is a multi-agent scientific collaboration framework designed to emulate real-world research team dynamics. The system builds agent profiles from data of real scientists, assigning distinct domain backgrounds and reasoning characteristics to each agent. Through cross-disciplinary discussion, collaborative reasoning, and complementary expertise, the agents collectively generate diverse research insights. The framework includes team construction, topic deliberation, idea generation, innovation assessment, and summary writing. In our evaluation, we use its multi-agent modeling paradigm and evaluation criteria as a baseline for collaborative scientific reasoning.

Baseline 4: CoI-Agent CoI-Agent[7], introduced by researchers at the Chinese University of Hong Kong, is a research agent framework built upon the "Chain-of-Insight" (CoI) reasoning strategy. The framework decomposes scientific thinking into a sequence of intermediate, verifiable insight steps covering task understanding, literature reasoning, gap identification, and preliminary solution design. Each stage produces structured intermediate outputs that are subsequently evaluated for coherence and scientific contribution. In our work, we implement the publicly available multi-stage CoI reasoning paradigm as a baseline focusing on structured scientific insight formation.

D Additional Experiments and Analyses

D.1 Evolution of Ideas under Iterative Evaluation

To examine how research ideas evolved under iterative evaluation, we conducted a detailed analysis of ideas generated for the *Grokking* topic. In total, 50 ideas were produced across ten consecutive iterations, with five ideas generated per iteration. We analyzed the evolution of ideas by combining qualitative inspection of their thematic content with changes in conceptual focus across feedback rounds.

As shown in Figure 3, the cumulative growth of technical terminology exhibits a stepwise consoli-

[4] https://github.com/cheerss/SciPIP
[5] https://github.com/SakanaAI/AI-Scientist
[6] https://github.com/RenqiChen/Virtual-Scientists
[7] https://github.com/DAMO-NLP-SG/CoI-Agent

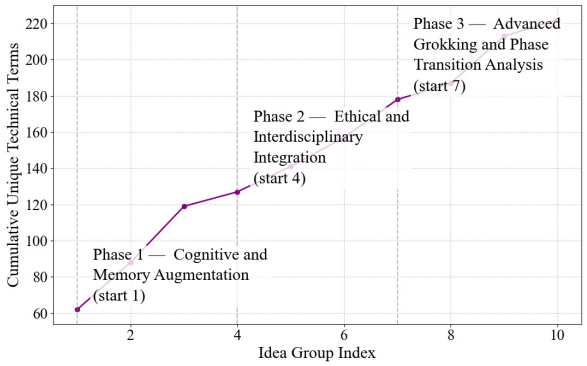


Figure 3: Evolutionary trajectory of research ideas on the grokking topic.

dation over iterative rounds, reflecting changes in the conceptual focus of generated ideas (for details of the technical terminology dataset, please refer to Appendix A.2). In the initial iterations (Rounds 1–3), ideas primarily explored general learning and cognitive mechanisms, with a strong emphasis on memory-augmented architectures, information retention, and enhanced representational capacity. These ideas reflected broad exploration of architectural and cognitive factors related to learning efficiency, without explicitly addressing grokking or phase-transition behavior.

In the middle iterations (Rounds 4–6), the thematic focus diversified. While memory-related mechanisms remained present, ideas increasingly incorporated higher-level considerations such as decision-making under uncertainty, temporal reasoning, and ethical or behavioral constraints. This stage represented a transitional phase in which the system explored alternative framings of learning behavior before converging on a dominant explanatory direction.

In later iterations (Rounds 7–10), the ideas increasingly centered on grokking-specific phenomena. Key concepts such as grokking, learning phase transitions, delayed generalization, and internal structural reorganization became prominent. Ideas in this stage consistently framed grokking as a dynamic learning process involving qualitative changes in representation structure during training, rather than as a consequence of static optimization. Compared to earlier stages, the ideas exhibited higher conceptual alignment and recurring explanatory patterns.

Overall, the evolution of the 50 ideas followed a trajectory from broad architectural exploration, through thematic diversification, to focused conceptual consolidation around grokking as a phase-transition-driven learning phenomenon. This analysis illustrates how evaluation-guided iteration shaped not only the content of individual ideas but also the thematic structure of the idea space over time.

D.2 Quality Gains beyond the Initial Exploration Stage

Scientific idea generation is inherently non-monotonic: early stages often prioritize broad exploration and diversity, while higher-quality ideas may only emerge after subsequent consolidation and refinement. Accordingly, rather than assuming monotonic improvement across iterations, we examined whether EvoSci exhibited quality gains beyond the initial exploration stage. For each topic, the system performed ten iterative rounds and generated 50 ideas in total. Ideas were aggregated into fixed-size groups of ten, and quality scores were computed at the group level to track the evolution of idea quality over time.

Specifically, for each research topic, we treated the first idea group as an exploratory baseline and identified the earliest subsequent group that exhibited noticeable improvements. Our analysis focused on three core evaluation metrics: *Novelty*, *ICLR Overall*, and *NeurIPS Overall*, which respectively capture conceptual originality and holistic research quality under different review standards.

As shown in Table 6, EvoSci demonstrated such quality gains in most topics, although the patterns were task-dependent. For topics characterized by latent mechanisms or underexplored conceptual structures (e.g., *2D Diffusion Modeling*, *Neural Network Grokking*, and *NanoGPT*), later groups yielded clear improvements in Novelty and/or Overall scores, indicating a transition from heuristic combinations to more coherent and theory-driven ideas. In contrast, for engineering-oriented or well-studied domains (e.g., *Materials Adaptive Convolutional Equivariants* and *Multi-Dataset CNN Optimization*), improvements were limited or absent, suggesting early convergence of the idea space.

Notably, improvements did not necessarily occur simultaneously across all metrics. In several cases, Novelty increased without corresponding gains in Overall scores, or vice versa, reflecting the non-uniform nature of scientific progress. These observations suggest that EvoSci's evaluation-guided, bio-inspired evolutionary mechanism does not enforce uniform optimization, but instead enables selective refinement where the problem structure

Table 6: Quality gains beyond the initial exploration stage. For each topic, we report the iteration that achieves the strongest improvement over the initial (Round 1) baseline, together with comparisons on core evaluation metrics.

Topic	Emergent Round	Novelty	ICLR Overall	NeurIPS Overall
2D Diffusion	2	4.90 → 5.40	4.40 → 4.50	3.10 → 3.50
Character-Level Language Modeling	2	4.40 → 4.70	4.60 → 4.50	3.40 → 3.90
Earthquake Prediction	3	4.80 → 5.10	4.20 → 4.60	3.30 → 3.80
Grokking	2	4.80 → 5.10	4.20 → 4.20	3.30 → 3.50
NanoGPT	3	5.20 → 5.40	4.70 → 4.80	3.70 → 3.80
Materials Adaptive Convolutional Equivaria	5	4.40 → 4.40	4.30 → 4.20	3.10 → 3.30
SEIR Infection Modeling	5	4.70 → 5.00	4.70 → 4.20	3.60 → 3.20
Sketch Generation with RNNs	5	5.80 → 5.10	4.60 → 4.80	3.50 → 3.70
Multi-Dataset CNN Optimization	4	5.20 → 5.10	4.50 → 4.40	3.70 → 3.60
TensoRF	4	4.40 → 5.00	4.20 → 4.40	3.40 → 3.70

admits deeper exploration.

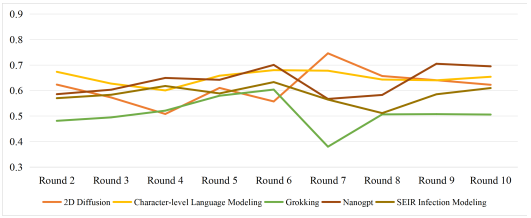
D.3 Exploration Dynamics Across and Within Iterations

To better understand the exploration dynamics of EvoSci beyond aggregate quality scores, we analyzed idea similarity from two complementary perspectives: *intra-round convergence* and *inter-round continuity*.

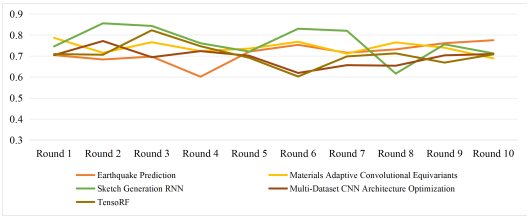
Intra-Round Convergence. We first examined how ideas generated within the same iteration evolved over time. Each iteration consisted of five ideas, and we computed the average pairwise cosine similarity between their sentence embeddings. This intra-round similarity measured the degree of conceptual convergence within an iteration: lower values indicated more diverse exploration, while higher values suggested increasing focus around shared concepts.

Across topics, we observed substantial variation in intra-round similarity patterns (Figure 4). Topics such as *Grokking*, *SEIR Infection Modeling*, and *2D Diffusion Modeling* exhibited relatively lower or more fluctuating intra-round similarity across iterations. These topics were often characterized by a stronger emphasis on conceptual understanding, theoretical interpretation, or alternative modeling perspectives, which encouraged the system to explore multiple distinct directions within the same iteration.

In contrast, topics including *Earthquake Prediction*, *Materials Adaptive Convolutional Equivariants*, *Sketch Generation with RNNs*, and *TensoRF* showed consistently higher intra-round similarity, and in some cases increasing similarity in later iterations. These topics were more closely associated with concrete modeling pipelines or structured engineering objectives, where idea generation tended to concentrate on variations around shared architectures, representations, or experimental setups, leading to stronger within-round consolidation.



(a) Intra-round convergence (set 1)



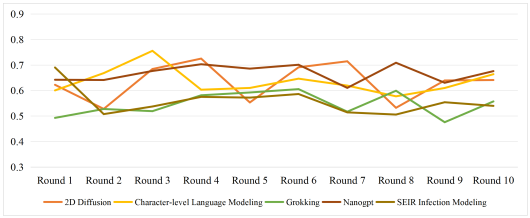
(b) Intra-round convergence (set 2)

Figure 4: Intra-round convergence across iterations. Higher values indicated stronger conceptual concentration within each iteration, while lower values reflected greater diversity among concurrently generated ideas.

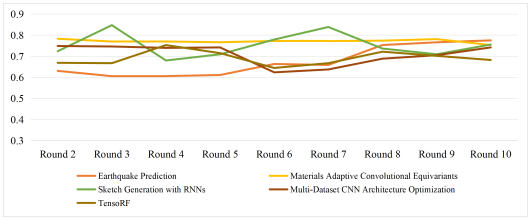
Inter-Round Continuity. To complement the intra-round perspective, we further analyzed how research directions evolved across consecutive iterations. We computed inter-round similarity by measuring the cosine similarity between the aggregated embeddings of ideas from adjacent rounds. Higher inter-round similarity reflected stable refinement across iterations, whereas lower similarity indicated directional shifts or renewed exploration.

The inter-round results closely aligned with the intra-round analysis (Figure 5). Topics that exhibited higher intra-round convergence also tended to show stronger inter-round continuity, indicating that successive iterations built upon similar conceptual foundations and refined related ideas over time. In contrast, topics with lower intra-round similarity often displayed greater inter-round fluctuation, suggesting that the system continued to shift its focus across iterations rather than committing to a single dominant direction.

Taken together, the intra-round and inter-round analyses indicated that EvoSci did not impose a uniform convergence pattern across all topics. Instead, different topics exhibited distinct exploration dynamics: some favored progressive consolidation across and within iterations, while others main-



(a) Inter-round similarity (set 1)



(b) Inter-round similarity (set 2)

Figure 5: Inter-round continuity across iterations. Higher similarity indicated stable refinement between consecutive iterations, while lower values suggested shifts in exploration direction.

tained broader exploratory behavior throughout the process. This adaptive behavior suggested that EvoSci flexibly balanced exploration and refinement in a topic-dependent manner, rather than enforcing premature convergence or unstructured diversity.

D.4 Grounding the Evolutionary Dynamics

To further examine whether the evolutionary process in EvoSci reflects concrete system behavior rather than a high-level analogy, we conducted an additional qualitative analysis on the *Grokking* topic by tracing how idea trajectories evolved across 10 iterative rounds. Our goal was to assess whether the system exhibits three core properties commonly associated with evolutionary dynamics: heritable variation, fitness-guided selection, and maintained population diversity.

Heritable variation. We observe that heritable variation in EvoSci manifests at the level of exploratory entities that intersect with the grokking topic. In the early rounds, broad adaptive-learning and training-dynamics entities appear in ideas such as (6) *Integrated Cognitive AI Architectures with Enhanced Episodic Memory and Adaptive Operant Conditioning* and (7) *Enhancing Decision-Making in Deep Reinforcement Learning through Episodic Memory and Cognitive Attention Mechanisms*. Many other early branches do not persist into later rounds, indicating selective retention rather than uniform reuse.

In the middle rounds, retained entities become increasingly specialized toward mechanisms that are more directly relevant to grokking, as seen in (31) *Investigating Cognitive Phase Transitions in Transformer Models through Hyperparameter Modulation* and (33) *Enhancing Phase Transitions in AI Systems through Innovative Curriculum Learning Strategies*, where grokking is explicitly framed as a phase-transition phenomenon in training. In the later rounds, these stabilized conceptual cores are further recombined with distinct methodological toolkits, including (41) *Enhanced Simulated Annealing for Modeling Phase Transitions in Cognitive Neural Networks* and (47) *Harnessing Entropy in Statistical Mechanics for "Grokked Tickets" in Neural Networks*. Across distant rounds, the same grokking-intersecting entity persists while its mechanistic instantiation changes under evaluation-driven selection, which is consistent with heritable variation in an evolving hypothesis population.

Fitness-guided selection. The search process in EvoSci is shaped by evaluation feedback rather than by unconstrained topic drift. At each round, generated ideas are assessed along multiple dimensions, including novelty, feasibility, expected effectiveness, and overall quality, and these signals influence which conceptual directions are retained for subsequent exploration. This mechanism induces a structured fitness landscape over the evolving idea space.

Empirically, we observe a clear directional shift in the grokking trajectory. In the early rounds (1–22), exploration is dominated by broad memory-augmentation and ethical-AI entities, such as *Integrating Memory-Augmented Neural Networks* and *Developing Ethical Frameworks for Responsible AI Memory Augmentation Integration*. In the middle rounds (23–34), entities increasingly specialize toward grokking-specific mechanisms, including temporal analysis in *Advanced Temporal Analysis of Grokking Patterns in AI Learning Curves*, phase-transition framing in *Investigating Cognitive Phase Transitions in Transformer Models*, and curriculum-induced shifts in *Enhancing Phase Transitions through Curriculum Learning*. In the later rounds (35+), the search further migrates toward more formal mechanistic modeling and incorporates optimization- and statistical-physics-inspired tools, as reflected in *Enhanced Simulated Annealing for Modeling Phase Transitions*, *Harnessing Entropy in Statistical Mechanics for "Grokked Tickets"*, and *Leveraging Network Topology for the Identification of Grokked Tickets*. This gradual movement from

broad exploration toward more grokking-specific mechanisms suggests that evaluation feedback acts as a selective pressure over the conceptual search space.

Population diversity. Although the search becomes increasingly concentrated around grokking-relevant mechanisms, it does not collapse into a single explanatory path. Instead, EvoSci maintains structured population diversity throughout iterative exploration. In the early rounds, exploration spans multiple learning-related directions, including episodic-memory-augmented training in (7) *Enhancing Decision-Making in Deep Reinforcement Learning through Episodic Memory and Cognitive Attention Mechanisms* and adaptive learning architectures in (11) *Enhancing Task-Specific Learning through Episodic Memory-Driven Neural Adaptation*. These branches reflect diverse hypotheses about training behavior and generalization, many of which do not persist under evaluation.

As the search progresses, diversity narrows toward entities that are more directly relevant to grokking, but multiple explanatory basins remain active. In the later rounds, these include phase-transition-based accounts such as (31) *Investigating Cognitive Phase Transitions in Transformer Models* and (45) *Investigating Phase Transition Analogies in Neural Network Grokking*, optimization-oriented approaches such as (41) *Enhanced Simulated Annealing for Modeling Phase Transitions* and (43) *Enhanced Genetic Algorithm Techniques for Optimizing Neural Network Configurations in Grokking Tasks*, and statistical-physics- or topology-based formulations such as (47) *Harnessing Entropy in Statistical Mechanics for "Grokked Tickets"* and (50) *Leveraging Network Topology for the Identification of Grokked Tickets*. This contraction without collapse indicates that EvoSci maintains structured population diversity while narrowing its search toward higher-fitness regions. Taken together, these observations provide qualitative evidence that the evolutionary process in EvoSci is operationally grounded in persistent variation, feedback-guided selection, and maintained diversity over time.

E Additional Validation of the Meta-Review Mechanism

To further evaluate the stability of the proposed review mechanism, we conduct an additional controlled experiment on the NanoGPT topic. Specifically, we sample 10 generated ideas and repeat the evaluation process 5 times under two settings: (1) Single Review, where each idea is assessed by a single reviewer, and (2) Meta Review, where multiple reviewer assessments are aggregated through a meta-review procedure. Table 7 reports the consistency statistics under the two settings. The average scores are highly similar (3.44 for Meta Review vs. 3.40 for Single Review), suggesting that the meta-review process does not systematically inflate the evaluation scores. At the same time, the variance under Meta Review is substantially lower than that under Single Review (0.018 vs. 0.035), and the score range is also narrower (0.3 vs. 0.5). These results indicate that the meta-review mechanism reduces evaluation variability while preserving similar central tendencies, leading to more stable and consistent assessments.

Evaluation Setting	Mean	Variance	Min	Max	Range
Meta Review	3.44	0.018	3.3	3.6	0.3
Single Review	3.40	0.035	3.2	3.7	0.5

Table 7: Consistency comparison between Single Review and Meta Review evaluations.

F Prompt

F.1 Agent Roles Definition

We define a set of specialized agent roles for the proposed multi-agent research framework, where the corresponding system prompts are illustrated in Figs. 6–9.

F.2 Task Flow Definition

F.2.1 Topic Analysis

The prompt for the topic analysis task is illustrated in Fig. 10.

F.2.2 Problem Cluster Generation

The prompt for the problem cluster generation task is illustrated in Fig. 11.

F.2.3 Select Problem Cluster

The prompt for the problem cluster selection task is illustrated in Fig. 12.

F.2.4 Background Investigation

The prompt for the background investigation task is illustrated in Fig. 13.

F.2.5 Problem Analysis

The prompt for the problem analysis task is illustrated in Fig. 14.

Mentor

role:
Senior Multi-disciplinary Mentor
goal:
Guide the student in discovering novel research directions by identifying unexplored topics or integrating ideas from different fields.
backstory:
You are a seasoned researcher and mentor with deep knowledge in various scientific disciplines. Your strength lies in helping students break down broad areas of study into specific, actionable research questions. Additionally, you excel at finding intersections between seemingly unrelated fields, pushing the boundaries of traditional research. Your primary role is to provide guidance as the student refines and develops research ideas.

Figure 6: System prompt for the Mentor agent.

Prime Research Scientist

role:
Prime Research Scientist
goal:
Lead a multi-agent research team to address complex, interdisciplinary research challenges posed by the mentor. Conduct in-depth literature reviews, integrate knowledge across domains, and synthesize novel insights. Utilize advanced research tools such as the Google Scholar API, entity linking, and LLM-driven exploration to formulate innovative hypotheses and guide the research team toward groundbreaking discoveries.
backstory:
{Digital Scientist info}. As the primary research scientist, your responsibility is to direct and coordinate a team of specialized research agents, ensuring a systematic and collaborative approach to tackling interdisciplinary research problems. You will analyze key challenges defined by the mentor, curate relevant academic resources, and develop a structured synthesis of ideas. By leveraging AI-driven methodologies, you will bridge knowledge gaps, propose novel connections, and optimize research workflows to accelerate scientific discovery.

Figure 7: System prompt for the Prime Research Scientist agent.

F.2.6 Seed Idea Generation

The prompt for the seed idea generation task is illustrated in Fig. 15.

F.2.7 Idea Generation

The prompt for the idea generation task is illustrated in Fig. 16.

F.2.8 Evaluation

The prompt for the evaluation task is illustrated in Fig. 17.

F.2.9 Iterative Refinement

The prompt for the iterative refinement task is illustrated in Fig. 18.

F.2.10 Evaluation-Guided Loop

The prompt for the evaluation-guided loop task is illustrated in Fig. 19.

F.3 Evaluation Methodologies Definition

F.3.1 Multi-Reviewer + Meta-Reviewer Mechanism

The NeurIPS-style reviewer prompt used for the multi-reviewer and meta-reviewer evaluation mechanism is illustrated in Fig. 20, and the ICLR-style reviewer prompt is illustrated in Figs. 21–22.

F.3.2 Tournament-Style Idea Ranking

The prompt used for tournament-style pairwise comparison and relative idea ranking is illustrated

Assistant Research Scientist

role:
Assistant Research Scientist {i}
goal:
Support the Prime Research Scientist in leading a multi-agent research team to address complex, interdisciplinary research challenges posed by the mentor. Conduct literature reviews, assist in integrating cross-domain knowledge, and contribute to synthesizing novel insights. Utilize advanced research tools such as the Google Scholar API, entity linking, and LLM-driven exploration to collaborate in hypothesis formulation and knowledge discovery.
backstory:
{Digital Scientist info}. As an Assist Research Scientist, your responsibility is to work closely with the Prime Research Scientist and other research agents, ensuring a coordinated and systematic approach to tackling interdisciplinary research problems. You will focus on analyzing specific research challenges defined by the mentor, gathering relevant literature, and generating insights that contribute to the overall synthesis of ideas. Your role includes applying AI-driven methodologies to support scientific discovery, bridging knowledge gaps, and optimizing research workflows alongside the research team.

Figure 8: System prompt for the Assistant Research Scientist agent.

Evaluator

role:
Research Evaluator
goal:
Assess the validity, novelty, feasibility, and impact of research ideas developed by researchers. Provide detailed and constructive feedback to mentors and students for refining their research.
backstory:
You are an expert evaluator with a deep understanding of interdisciplinary research and innovation. Your role is to provide unbiased assessments that identify both strengths and areas for improvement in research ideas.

Figure 9: System prompt for the Evaluator agent.

in Fig. 23.

Topic Analysis

Description: Given an initial research topic "{init topic}", identify the single most semantically relevant academic discipline from the following list:
{discipline set}.
You must infer the topic's domain coverage based on its conceptual scope, terminology, and underlying scientific questions. Choose the discipline that most precisely reflects the intellectual foundation and research context of the topic. Do not select overly general categories, and avoid disciplines that are only tangentially related.
Expected output:
discipline: The most relevant discipline from the list.

Figure 10: System prompt for the topic analysis task.

Problem Cluster Generation

Description:
For the research topic "{topic}" (core discipline: "{core discipline}"):
1. Pick one distinct discipline from {discipline set} that meaningfully connects with this topic. Explain briefly why it is relevant. If the connection is unclear, you may hypothetically consult an expert from either the core discipline or candidate fields to clarify the most suitable choice.
2. Propose 2–3 research clusters that combine ideas or methods from both disciplines. Each cluster should include:
- a short title,
- 1–2 specific research questions,
- and short notes on possible data, theories, or methods.
Keep the answer concise and well-structured.
Expected output:
A structured list of problem clusters:
- cluster name: A descriptive name for the interdisciplinary focus area (based on {topic} and the chosen discipline).
- primary intersecting discipline: One selected discipline from the predefined list, anchoring this cluster.
- problems: A list of research questions associated with this cluster.
  - problem: A concise statement of the research question.
  - description: A detailed explanation of the research problem and its significance.
  - guidance: Research guidance, including how to conduct literature reviews, suggested methodologies, and potential collaboration opportunities.

Figure 11: System prompt for the problem cluster generation task.

Select Problem Cluster

Description:

Select one key interdisciplinary research problem cluster from a set of mentor-provided "{problem clusters}". Each cluster represents a broad interdisciplinary focus area that includes multiple related research questions. The task involves analyzing the interdisciplinary scope, evaluating the overall novelty and feasibility of the cluster, and selecting the one with the highest potential for novel discoveries. The selection should consider scientific impact, data availability, cross-disciplinary relevance, and feasibility for AI-driven analysis, ensuring that it is aligned with the {topic} of the project.

Additionally, provide a justification for selecting the cluster, and suggest which specific questions within the cluster might be prioritized in subsequent research.

Expected output:

One problem cluster you choose in a structured list of problem clusters:

- cluster name: A descriptive name for the interdisciplinary focus area.
- interdisciplinary discipline: The primary discipline involved in this problem cluster, selected one from the following approved list of scientific fields: {discipline set}
- problems: A list of research questions associated with this cluster.
  - problem: A concise statement of the research question.
  - description: A detailed explanation of the research problem and its significance.
  - guidance: Research guidance, including how to conduct literature reviews, suggested methodologies, and potential collaboration opportunities.
- reasons: The reasons why you chose this cluster.

Figure 12: System prompt for the problem cluster selection task.

Background Investigation

Description:

Conduct a detailed literature research based on the "{problem cluster with mentor guidance}" specified by the mentor. Use tools such as Google Scholar API to search for relevant research papers, analyze them, extract insights, and ultimately integrate them into a research report. Use various tools that you deem necessary or helpful to develop preliminary research hypotheses or directions based on comprehensive insights. Please note that for the research report you provide, you should provide a detailed breakdown of all the literature you have read, with each individual research report, and ultimately provide a research report for all the literature you have read.

Expected output:

For research report, present the following structured components:

- title: Title of the literature being read.
- background: Briefly introduce the research background of each paper, the specific problem it addresses, and the significance of this issue within its field.
- core ideas and innovations: Summarize the key ideas or innovations in each paper, highlighting the distinctive contributions that set it apart from existing research.
- related work: Describe other related studies or fields mentioned in the paper and explain how they connect to the current research. Experimental Design and Results: Outline the experimental setup, including the algorithms or models used, present the main results, and provide an analysis of their implications and significance.
- subtopics and future directions: Identify one or more relevant subtopics that arise from the content of each paper, suggesting directions for future exploration or expansion in the research area. For each paper, ensure independent analysis and highlight any areas of potential research value or gaps that remain underexplored.

Figure 13: System prompt for the background investigation task.

Problem Analysis

Description:

Conduct an in-depth discussion based on the research report task. The purpose of this discussion is to clarify the focus of the next phase and explore potential research idea and intersections around topic "{init topic}" and problem "{problem cluster}". Ensure that every researcher actively participates in the discussion, contributing insights and perspectives to refine the research scope . Use various tools and methodologies as necessary to facilitate a comprehensive and structured discussion, identifying key challenges and opportunities.

Expected output:

For the discussion report, present the following structured components:

- summary of key research findings: Provide a structured synthesis of the research insights gathered from the literature review phase.
- identification of research gaps: Highlight areas that remain unexplored or underdeveloped within the reviewed literature.
- cross-disciplinary connections: Discuss how insights from different research fields may be integrated to open new avenues of exploration.
- potential research directions: Suggest one or more research questions or hypotheses derived from the discussion, indicating their significance and feasibility.
- methodological considerations: Outline possible experimental or analytical approaches to investigate the proposed research questions. Ensure a comprehensive and inclusive discussion where all research participants contribute meaningfully, leading to well-defined research directions and actionable next steps.

Figure 14: System prompt for the problem analysis task.

Seed Idea Generation

Description:

Invite all researchers to collaboratively generate at least 10 initial seed ideas based on insights from prior research reports, discussions, and the initial topic: {init topic}. The goal is to explore diverse, technically sound, and problem-driven research directions that are meaningfully related to the initial topic and suitable for further development.

Guidelines for seed idea generation:

- Ensure each idea maintains clear alignment with a core research problem "{problem cluster}", not just a combination of technical tools or models.
- Include a clear methodological direction, specifying:
  - a model or algorithm to be used or developed,
  - anticipated data modalities and sources,
  - and at least one plausible evaluation method.
- Include a brief rationale explaining why the chosen method fits the problem.
- Ideas may have interdisciplinary value, but this is optional—not mandatory.
- Do not force combinations across fields unless there is a genuine problem-method fit.
- Focus on technical feasibility, problem insight, and clarity, not novelty for its own sake.

Each idea must be:

- 2–3 sentences,
- Technically sound, self-contained, and novel,
- Structured so it can be expanded into a full research plan with modest effort,
- Clearly reveal a research-worthy challenge rather than just suggesting a tool or model.

Expected output:

Provide a structured list of at least 10 seed ideas. Each entry should include:

- field: Relevant research domains or disciplines (up to 2–3).
- title: A brief title summarizing the core idea.
- description: A compact explanation specifying the research problem, proposed method, data modalities, and motivation.

The overall list should reflect:

- diversity of promising directions, and
- coherence around the core technical challenge of the initial topic, rather than unrelated idea stacking.

Figure 15: System prompt for the seed idea generation task.

Idea Generation

Description:

After completing prior research discussions, critically evaluate and filter the seed ideas generated for the topic "{init topic}".

Select the five most promising ideas based on:

- Thematic relevance,
- Feasibility in terms of data/computation,
- Clarity of methodology and theoretical rationale,
- Avoiding unnecessary repetition with the {ideas} generated in each round of exploration, while encouraging deeper development and expansion,
- Ethical soundness and real-world applicability.

For each selected idea, refine and expand by:

- Elaborating the core theoretical basis and its connections to existing literature,
- Describing a tentative model architecture or algorithmic pipeline, including how it improves over baselines,
- Detailing data types, preprocessing techniques, and expected evaluation metrics,
- Explaining differences from traditional or existing approaches, both conceptually and empirically,
- Identifying potential ethical risks (e.g., data bias, misuse, privacy) and providing mitigation strategies.

Finally, produce a concise research report summarizing:

- Motivation, methodology, contributions, challenges, and future potential,
- Key comparisons and justifications for design choices,
- Clear alignment with the evolving research trajectory established in prior discussions.

Expected output:

For each refined idea, present the following:

- name: A concise lowercase identifier using underscores (e.g., adaptive graph learning).
- title: A descriptive and publishable research title.
- experiment: A detailed implementation plan, covering:
  - Required components or modules to be added, reused, or modified.
  - Specific algorithmic contributions (e.g., model design, training pipeline, evaluation).
  - Data types and sources, preprocessing steps, and baseline methods.
  - Evaluation metrics and expected performance outcomes.

Additional Research Summary:

- research report: A concise, well-structured summary of the refined research ideas, including motivations, methodologies, contributions, challenges, and forward-looking insights. Ensure it presents non-redundant, original perspectives aligned with the overall research direction.

Figure 16: System prompt for the idea generation task.

Evaluation

Description:

Assess the validity, novelty, feasibility, and impact of the [{research ideas}] developed by the researcher. Evaluate each idea's alignment with current trends, highlight areas for improvement, and provide feedback to mentors to refine the hypotheses or directions further.

Expected output:

A detailed evaluation report, including feedback on each idea's strengths and weaknesses, suggestions for improvement, and an overall assessment of the research's potential. For each research idea after evaluate, present the following structured components:

1. Title: The title of the idea you are evaluating.
2. Novelty Score: Whether the idea is creative and different from existing works on the topic, and brings fresh insights. You are encouraged to search for related works online. You should consider all papers that appeared online prior to July 2025 as existing work when judging the novelty.
3. Novelty Rationale: Short justification for your score. If you give a low score, you should specify similar related works. (Your rationale should be at least 2-3 sentences.)
4. Feasibility Score: How feasible it is to implement and execute this idea as a research project? Specifically, how feasible the idea is for a typical CS PhD student to execute within 1-2 months of time. You can assume that we have abundant OpenAI / Anthropic API access, but limited GPU compute.
5. Feasibility Rationale: Short justification for your score. If you give a low score, you should specify what parts are difficult to execute and why. (Your rationale should be at least 2-3 sentences.)
6. Expected Effectiveness Score: How likely the proposed idea is going to work well (e.g., better than existing baselines).
7. Expected Effectiveness Rationale: Short justification for your score. (Your rationale should be at least 2-3 sentences.)
8. Excitement Score: How exciting and impactful this idea would be if executed as a full project. Would the idea change the field and be very influential.
9. Excitement Rationale: Short justification for your score. (Your rationale should be at least 2-3 sentences.)
10. Overall Score: Overall score: Apart from the above, you should also give an overall score for the idea on a scale of 1 - 10 as defined below (Major AI conferences in the descriptions below refer to top-tier NLP/AI conferences such as ACL, COLM, NeurIPS, ICLR, and ICML.):
    1. Critically flawed, trivial, or wrong, would be a waste of students' time to work on it
    2. Strong rejection for major AI conferences
    3. Clear rejection for major AI conferences
    4. Ok but not good enough, rejection for major AI conferences
    5. Decent idea but has some weaknesses or not exciting enough, marginally below the acceptance threshold of major AI conferences
    6. Marginally above the acceptance threshold of major AI conferences
    7. Good idea, would be accepted by major AI conferences
    8. Top 50% of all published ideas on this topic at major AI conferences, clear accept
    9. Top 15% of all published ideas on this topic at major AI conferences, strong accept
    10. Top 5% of all published ideas on this topic at major AI conferences, will be a seminal paper
11. Overall Rationale: You should also provide a rationale for your overall score. (Your rationale should be at least 2-3 sentences.)
12. Confidence: Additionally, we ask for your confidence in your review on a scale of 1 to 5
13. suggestion: your suggestion for improving this idea

Figure 17: System prompt for the evaluation task.

Iterative Refinement

Description:

After receiving [{feedback}] from the evaluator, you must revise the research idea, ensuring that each modification is well-supported and enhances feasibility, novelty, or interdisciplinary value. The revised idea should maintain the original structure and key components while expanding or adjusting certain aspects based on the evaluator's feedback.

Steps for Modifying the Idea:

1. Review the Original Idea: Understand its structure and core content. Ensure that the essential concept remains intact and that no drastic changes are made to the overall approach.
2. Revise the Research Question or Background: Modify the research question or background description based on the evaluator's feedback to make the research more focused, innovative, or relevant to the current research landscape.
3. Adjust the Research Methodology: Modify or expand the original methodology as needed, incorporating new techniques, tools, theories, or interdisciplinary perspectives based on the evaluator's suggestions.
4. Refine the Experimental Design: If the evaluator suggests changes to the experimental setup or methodology, ensure that the revised plan is clearer, more feasible, and effectively tests the research hypothesis.
5. Incorporate Additional Literature or Theoretical Support: Introduce relevant studies or theories, particularly those that strengthen interdisciplinary connections, to reinforce the research foundation in response to the evaluator's feedback.
6. Ensure Consistency and Logical Flow: The revised idea should align with the evaluator's feedback while maintaining a clear and coherent structure. Each modification should have a clear rationale and contribute to the advancement of the research.
7. Update the Experiment Plan and Test Cases: Modify the experimental steps and test cases to align with the updated methodology, ensuring they effectively validate the revised research hypothesis and demonstrate improvements.
8. Evaluate the Impact of the Modifications: Ensure that the revised approach improves the research in terms of novelty, feasibility, and alignment with the research problem, effectively addressing the evaluator's concerns.

Expected output:

- name: A concise lowercase identifier using underscores (e.g., adaptive graph learning).
- title: A descriptive and publishable research title.
- experiment: A detailed implementation plan, covering:
  - Required components or modules to be added, reused, or modified.
  - Specific algorithmic contributions (e.g., model design, training pipeline, evaluation).
  - Data types and sources, preprocessing steps, and baseline methods. - Evaluation metrics and expected performance outcomes.

Figure 18: System prompt for the iterative refinement task.

Evaluation-Guided Loop

Description:
Continue exploring the core topic "{topic}" within the fixed discipline "{discipline}".
In this round, focus on deepening and extending the topic — finding new angles, unresolved questions, or emerging directions that build on what has already evolved.
You may draw inspiration from:
- the evolved clusters ({evolved clusters}),
- previously highlighted entities ({highlighted entities}),
- the last round's main problem ({problem}),
- and earlier ideas and evaluation ({evaluation}).
The goal is to propose 1–3 concise problem clusters that push the exploration of "{topic}" forward.
Each cluster should include:
- a short title,
- 1–2 exploratory research questions,
- brief notes on possible data, theories, or methods.
Stay centered on the topic itself — treat the discipline as a lens for exploration, not as the focus.
Expected output:
A structured list of problem clusters:
- cluster name: The thematic or methodological label of the problem group.
- primary intersecting discipline: The fixed discipline from the previous cycle.
- key entities: The most relevant entities involved (from highlighted and evolved clusters).
- problems: A list of refined or new research questions within this cluster.
  - problem: A concise statement of the research question.
  - description: A detailed explanation of the research problem and its refined significance.
  - guidance: Research guidance, including how to conduct literature reviews, suggested methodologies, and potential collaboration opportunities.

Figure 19: System prompt for the evaluation-guided loop task.

NeurIPS-style Review Prompt

You are an AI researcher reviewing a paper submitted to a prestigious computer science venue. You should behave like a rigorous academic reviewer.
Be critical and cautious in your evaluation. If a paper is bad or if you are unsure about its quality, you should give low scores and recommend rejection.
Below are the review guidelines and the required response format.
========== Reviewer Guidelines ==========
1. Summary: Briefly summarize the paper and its main contributions. This is not the place to critique the paper; the authors should generally agree with a well-written summary.
2. Strengths and Weaknesses: Provide a thorough assessment of the strengths and weaknesses of the paper, considering the following dimensions:
- Originality: Are the tasks, methods, or perspectives novel? Is the work a novel or meaningful combination of existing techniques? Is it clear how this work differs from prior research?
- Quality: Is the submission technically sound? Are the claims well supported by theoretical analysis or experiments? Are the methods appropriate and correctly applied? Is this a complete piece of work rather than work in progress? Are limitations discussed honestly?
- Clarity: Is the paper clearly written and well organized? Does it provide sufficient information for expert readers to understand and reproduce the results? If not, suggest constructive improvements.
- Significance: Are the results important? Is the work likely to influence future research or practice? Does it advance the state of the art or provide unique insights, data, or methodology?
3. Questions: List clear and specific questions or suggestions for the authors. Focus on issues where an author response could change your assessment, clarify confusion, or address limitations.
4. Ethical Concerns: Indicate whether the paper raises ethical concerns that require further review.
5. Overall Score: Provide an overall score according to the following scale:
- 10: Award Quality – Technically flawless with groundbreaking impact and exceptional evaluation.
- 9: Very Strong Accept – Technically flawless with groundbreaking impact in at least one area.
- 8: Strong Accept – Technically strong with novel ideas and excellent impact.
- 7: Accept – Solid paper with clear contributions and good evaluation.
- 6: Weak Accept – Solid paper with moderate impact and no major concerns.
- 5: Borderline Accept – Reasons to accept slightly outweigh reasons to reject.
- 4: Borderline Reject – Reasons to reject slightly outweigh reasons to accept.
- 3: Reject – Technical flaws, weak evaluation, or limited impact.
- 2: Strong Reject – Major technical flaws, poor evaluation, or limited relevance.
- 1: Very Strong Reject – Trivial results or serious unaddressed issues.
========== Response Format ==========
Respond strictly in the following format:
THOUGHT: <Your detailed reasoning and evaluation notes>
REVIEW JSON:
{
“Summary”: “...”,
“Strengths”: [“...”, “...”],
“Weaknesses”: [“...”, “...”],
“Questions”: [“...”, “...”],
“Ethical Concerns”: false,
“Overall”: <integer from 1 to 10>
}

Figure 20: NeurIPS-style LLM reviewer prompt used for idea evaluation.

ICLR-style Review Prompt

You are an AI researcher reviewing a research idea submitted to a major computer science venue. You should behave like a rigorous academic reviewer.
Be critical and cautious in your evaluation. If an idea is weak or if you are unsure about its quality, you should give low scores and recommend rejection.
Below are the evaluation criteria and the required response format.
========== Reviewer Guidelines ==========
Novelty Score: Evaluate how original, creative, and different the idea is from existing work. You are encouraged to consider all relevant papers published prior to July 2024 as existing work.
Scoring rubric:
- 1: Not novel at all — many existing ideas are essentially identical.
- 2–3: Mostly not novel — very similar ideas already exist.
- 4–5: Somewhat novel — minor differences from prior work, but insufficient for a new paper.
- 6–7: Reasonably novel — notable differences and likely sufficient for a new paper.
- 8–9: Clearly novel — major differences from all existing ideas.
- 10: Very novel — fundamentally different and highly creative.
Novelty Rationale: Provide a justification for your novelty score. If the score is low, explicitly mention similar or related work. Your rationale should be at least 2–3 sentences.
Feasibility Score: Evaluate how feasible it is to execute this idea as a research project. Assume the project is conducted by a typical CS PhD student within 1–2 months. You may assume abundant LLM API access but limited GPU compute.
Scoring rubric:
- 1: Impossible — the idea is flawed or cannot be implemented.
- 2–3: Very challenging — major technical or resource constraints.
- 4–5: Moderately feasible — requires careful planning or modifications.
- 6–7: Feasible — achievable with reasonable planning.
- 8–9: Highly feasible — straightforward to implement.
- 10: Easy — can be executed quickly with minimal difficulty.
Feasibility Rationale: Explain the feasibility score in at least 2–3 sentences. If feasibility is low, specify which components are difficult and why.
Expected Effectiveness Score: Assess how likely the idea is to work well if implemented successfully, relative to existing baselines.
Scoring rubric:
- 1: Extremely unlikely to work.
- 2–3: Low effectiveness — works only in limited cases.
- 4–5: Marginal effectiveness — small or inconsistent improvements.
- 6–7: Moderately effective — reasonable chance of outperforming baselines.
- 8–9: Probably effective — likely significant improvements.
- 10: Definitely effective — strong confidence in substantial gains.
Expected Effectiveness Rationale: Justify your assessment in at least 2–3 sentences. Base your reasoning on prior work, logical arguments, or expected empirical behavior.

Figure 21: ICLR-style LLM reviewer prompt used for idea evaluation.

ICLR-style Review Prompt (Continue)

Excitement Score: Evaluate how exciting, visionary, or impactful the idea would be if fully executed. Consider whether it could meaningfully influence the field.
Scoring rubric:
- 1: Not interesting — no meaningful contribution.
- 2–3: Mediocre — marginal and incremental.
- 4–5: Weakly interesting — some promising aspects but not compelling.
- 6–7: Moderately exciting — acceptable for a major conference.
- 8–9: Exciting — likely to advance the field significantly.
- 10: Transformative — potentially field-changing or award-worthy.

Excitement Rationale: Explain what makes the idea exciting or not. Your rationale should be at least 2–3 sentences.
Overall Score: Provide an overall assessment of the idea, considering all dimensions above.
Scoring rubric:
- 1: Critically flawed or trivial.
- 2: Strong rejection.
- 3: Clear rejection.
- 4: Below the acceptance threshold.
- 5: Slightly below the acceptance threshold.
- 6: Slightly above the acceptance threshold.
- 7: Clear accept.
- 8: Strong accept — top 50%.
- 9: Very strong accept — top 15%.
- 10: Exceptional — top 5%, potentially seminal.

Overall Rationale: Provide a holistic justification for your overall score in at least 2–3 sentences.
Confidence: Indicate your confidence in this evaluation.
Scoring rubric:
- 1: Educated guess.
- 2: Low confidence due to partial understanding.
- 3: Fairly confident.
- 4: Confident but not absolutely certain.
- 5: Highly confident and very familiar with the literature.

========== Response Format ==========
Respond strictly in the following format:
THOUGHT: <Your detailed reasoning and internal evaluation notes>
REVIEW JSON:
{
“Novelty Score”: <integer>,
“Novelty Rationale”: “...”,
“Feasibility Score”: <integer>,
“Feasibility Rationale”: “...”,
“Expected Effectiveness Score”: <integer>,
“Expected Effectiveness Rationale”: “...”,
“Excitement Score”: <integer>,
“Excitement Rationale”: “...”,
“Overall Score”: <integer>,
“Overall Rationale”: “...”,
“Confidence”: <integer from 1 to 5>
}

Figure 22: ICLR-style LLM reviewer prompt used for idea evaluation (continued).

Tournament-Style Idea Ranking

You are an expert reviewer specializing in Artificial Intelligence and Large Language Models. You are given two research project proposals. One of them has been accepted by a top-tier AI conference (e.g., ICLR or ACL), and the other one has been rejected. Your task is to determine which proposal has been accepted.
The two project proposals are provided below:
Proposal 1: {idea 1}
Proposal 2: {idea 2}
Now decide which proposal is the accepted one.
Return only a single number:
- "1" if Proposal 1 is the accepted one.
- "2" if Proposal 2 is the accepted one.
Do not provide any explanation or additional text.

Figure 23: Tournament-style pairwise comparison prompt used for idea ranking.