Many Heads Are Better Than One: Improved Scientific Idea Generation by A LLM-Based Multi-Agent System

Haoyang Su[1,*], Renqi Chen[1,*], Shixiang Tang[1,5,†], Zhenfei Yin[3], Xinzhe Zheng[1], Jinzhe Li[1], Biqing Qi[1], Qi Wu[4], Hui Li[4], Wanli Ouyang[1,5], Philip Torr[3], Bowen Zhou[1,6], Nanqing Dong[1,2,†,‡]

[1]Shanghai Artificial Intelligence Laboratory [2]Shanghai Innovation Institute
[3]Department of Engineering Science, University of Oxford
[4]Shanghai Institute for Science of Science
[5]Department of Information Engineering, Chinese University of Hong Kong
[6]Department of Electronic Engineering, Tsinghua University

Abstract

The rapid advancement of scientific progress requires innovative tools that can accelerate knowledge discovery. Although recent AI methods, particularly large language models (LLMs), have shown promise in tasks such as hypothesis generation and experimental design, they fall short of replicating the collaborative nature of real-world scientific practices, where diverse experts work together in teams to tackle complex problems. To address the limitations, we propose an LLM-based multi-agent system, *i.e.*, Virtual Scientists (VIRSCI), designed to mimic the teamwork inherent in scientific research. VIRSCI organizes a team of agents to collaboratively generate, evaluate, and refine research ideas. Through comprehensive experiments, we demonstrate that this multi-agent approach outperforms the state-of-the-art method in producing novel scientific ideas. We further investigate the collaboration mechanisms that contribute to its tendency to produce ideas with higher novelty, offering valuable insights to guide future research and illuminating pathways toward building a robust system for autonomous scientific discovery. The code is available at https://github.com/open-sciencelab/Virtual-Scientists.

1 Introduction

The rapid acceleration of scientific research necessitates innovative tools for exploring new concepts and tackling complex challenges (Park et al., 2023b). Automatic scientific discovery has emerged as a promising solution to accelerate innovation, aligning with the ultimate goal within the scientific community (Langley, 1987). With the development of artificial intelligence (AI), automatic scientific discovery has the potential to revolutionize how research is conducted by automating key steps in the scientific process, ranging from hypothesis generation to experimental design (Raghu and Schmidt, 2020; Spangler et al., 2014).

Recent works like AI Scientist (Lu et al., 2024), ResearchTown (Yu et al., 2024), and HypoGen (Qi et al., 2024) leverage LLMs (OpenAI, 2023; Dubey et al., 2024) as intelligent agents to simulate the scientific idea generation process and advance automatic scientific discovery at various stages, including literature review (Shi et al., 2023; Hsu et al., 2024) and experimental design (Wu et al., 2023; Huang et al., 2024). However, these efforts either rely on a **single-agent** system, overlooking the collaborative nature of real-world research (Kayacik et al., 2019; Gauch, 2003; Linsey et al., 2005), or employ an oversimplified collaboration framework and unrealistic data (e.g. manually crafted personal profiles, synthetic collaboration networks) to model a **multi-agent** system, failing to capture the dynamic relationships that characterize real scientific teams. Consequently, they provide limited insights into multi-agent collaboration, which is essential for advancing autonomous scientific discovery.

To address these limitations, we build a virtual *Scientific Research Ecosystem* to benchmark the potential of multi-agent systems in scientific idea generation. Specifically, *Scientific Research Ecosystem* is a digit twin of research communities at the given time point, *e.g.,* Jan 1st, 2024. It have three components: (1) virtual scientists whose backgrounds and publications are cloned from real scientists at that moment, (2) a past paper database where papers are published before that moment and (3) a contemporary paper database where papers are published after that moment. To evaluate the novelty of the generated ideas, we introduce three metrics from different perspectives: dissimilarity to past papers, alignment with research trends, and the potential influence on contemporary research (Shao

*Equal contribution.
†Corresponding authors: Nanqing Dong (dongnanqing@pjlab.org.cn) and Shixiang Tang (tangshixiang@pjlab.org.cn).
‡Project lead.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 28201–28240
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

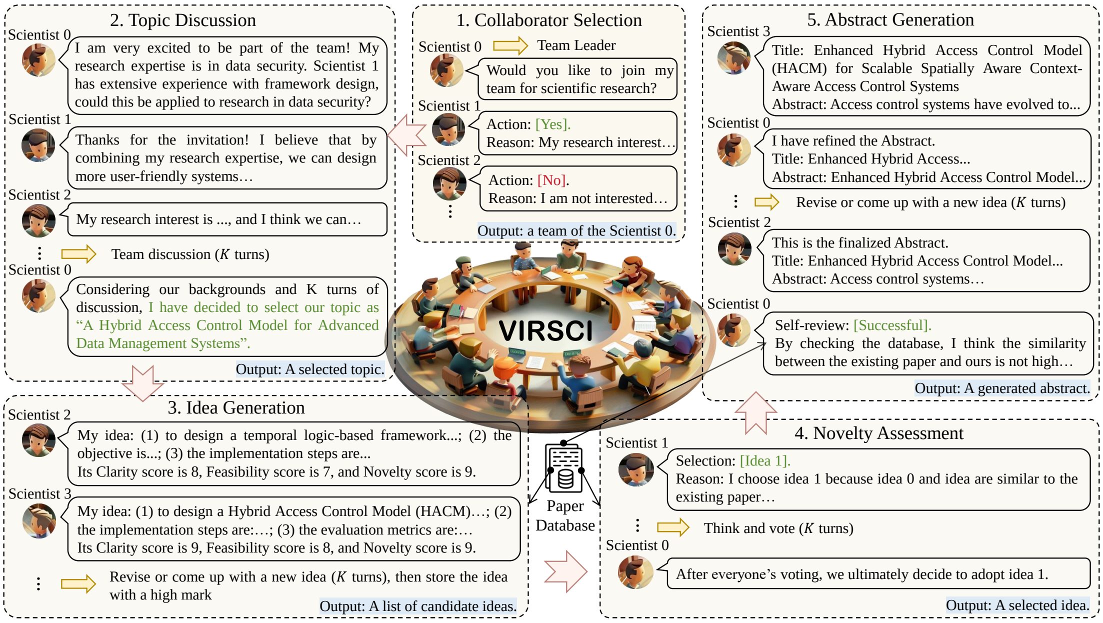


Figure 1: The proposed LLM-based multi-agent system, VIRSCI, includes five key steps: *Collaborator Selection*, where a research team is assembled; *Topic Discussion*, where the research topic is determined; *Idea Generation*, where team members propose and refine ideas; *Novelty Assessment*, where ideas are evaluated and voted on to select the best one; and *Abstract Generation*, where the selected idea is developed into a complete abstract.

et al., 2020; Yang et al., 2022). By comparing the generated abstract against two paper databases, we ensure the generated ideas are innovative and align with emerging scientific directions, validating the effectiveness of our approach. Real examples and their professional analysis are shown in Appx. F.

Based on the virtual *Scientific Research Ecosystems*, we propose an LLM-based multi-agent system, Virtual Scientists (VIRSCI), designed to harness the potential of LLM agents in assisting autonomous scientific idea generation. Leveraging the inherent human-like reasoning capabilities of LLMs (Xie et al., 2024), VIRSCI simulates the collaborative process of generating scientific ideas (Perry-Smith and Mannucci, 2017; Muzzio and Gama, 2024), by the following five steps (see Fig. 1): (1) *Collaborator Selection*, (2) *Topic Selection*, (3) *Idea Generation*, (4) *Idea Novelty Assessment*, and (5) *Abstract Generation*. In *Collaborator Selection* stage, given a randomly selected agent as the team leader, it will exploit historical co-authors based on its collaboration history and academic social networks, while also exploring potential collaborators whose expertise and research interests align with the team's goals (March, 1991; Rzhetsky et al., 2015; Zeng et al., 2019). In *Topic Selection* stage, the scientists will discuss topics of common interest. The discussion will be terminated and restarted if a consensus cannot be achieved among the majority of the team. The scientists who are not interested can choose to leave the discussion at will. Otherwise, the discussion continues until a final topic is determined. In the *Idea Generation* stage, the virtual scientists retrieve relevant papers from the past paper database and engage in both inter- and intra-team discussions, where collaborators participate in iterative dialogues based on their backgrounds. This inter- and intra-team discussion distinguishes our approach from previous group discussion patterns (Zhang et al., 2024; Qi et al., 2024; Qian et al., 2024), enabling agents within the team to proactively seek advice from external agents (inter-team) through an "Invitation Mechanism", while effectively balancing diverse perspectives within the team (intra-team). Once the discussions are over, the *Abstract Generation* stage begins, and the team generates a comprehensive abstract that encapsulates the proposed ideas.

We conduct extensive experiments to verify the effectiveness of VIRSCI in producing novel scientific ideas on both single-discipline and multi-discipline datasets. The findings prove that the multi-agent system improves the single-agent executive by, on average, **+13.8%** and **+44.1%** in alignment with and potential impacts on contemporary research, respectively. Furthermore, our experiments investigate collaboration mechanisms among agents that influence the performance of idea generation. The patterns observed in the experimental results align with findings from prior Science of Science studies (Fortunato et al., 2018; Wu et al., 2019; Zeng et al., 2021; Shi and Evans, 2023) published in Top venues, e.g., *Science* and *Nature*, providing valuable insights to guide future research toward autonomous scientific discovery. Our core contributions are as follows:

1) To the best of our knowledge, we propose the first multi-agent system with a scientific research ecosystem for conducting and benchmarking scientific collaborations, named VIRSCI, where real data is used for role-playing and objective evaluation.

2) To simulate a reliable scientific collaboration process, we propose an end-to-end pipeline that spans team organization to idea generation. A novel inter- and intra-team discussion mechanism is introduced to promote communication topology and enhance the simulation realism.

3) Extensive experiments demonstrate that multi-agent collaboration improves outcome quality, surpassing existing methods. We also conduct systematic experiments to investigate factors influencing the system, with findings aligning with established principles from the Science of Science, such as fresh teams tend to generate more innovative research, underscoring VIRSCI's reliability as a powerful tool for autonomous scientific discovery.

2 Related Work

2.1 AI for Scientific Discovery

In recent years, AI has transformed scientific discovery by offering powerful tools to enhance research processes (Xu et al., 2021). Generative AI, in particular, accelerates basic scientific discoveries by tackling complex tasks such as molecular identification (Vignac et al., 2022), protein structure prediction (Abramson et al., 2024), and proteomics research (Ding et al., 2024), significantly reducing experimental iteration times. Besides, with the advent of LLMs, AI methodologies can step further and collaborate in streamlining critical stages of the scientific pipeline, including hypothesis generation, experimental design, data acquisition, and analysis(Zheng et al., 2023; Wang et al., 2023; Miret and Krishnan, 2024; Wysocki et al., 2024; Lu et al., 2024; Si et al., 2024). Nevertheless, these approaches lack the collaborative nature of the scientists intrinsic to real-world research. VIRSCI is the first to harness the power of an LLM-based multi-agent system to facilitate the generation of research ideas in autonomous scientific discovery.

2.2 Collaboration in Multi-Agent Systems

A multi-agent system for team collaboration leverages autonomous agents to coordinate, communicate, and solve tasks within a shared environment, simulating human teamwork dynamics (Dorri et al., 2018). Traditional systems typically rely on semi-autonomous agents using explicit protocols and structured messages to achieve shared goals (Dunin-Keplicz and Verbrugge, 2011; Bakliwal et al., 2018). The emergence of LLMs has reshaped this landscape by enabling agents to communicate in natural language, fostering more intuitive and flexible interactions (Park et al., 2023a). Studies have demonstrated that LLM-based multi-agent systems outperform single-agent setups in tasks such as programming, game playing, and complex reasoning (Du et al., 2024; Wang et al., 2024; Light et al., 2023). However, prior multi-agent frameworks (Qi et al., 2024; Yu et al., 2024) for focused scientific idea generation (1) lack dynamic team composition, restricting team members from accessing insights beyond their initial group, and (2) do not incorporate real scientist data and collaboration relationship, leading to impractical conclusions. To overcome these limitations, we introduce an inter- and intra-team discussion mechanism, enabling flexible collaboration among virtual scientists (supported by real data) and fostering the generation of *de novo* scientific ideas.

3 The VIRSCI

In this paper, we aim to build a multi-agent system using real-world academic datasets to simulate how a scientist assembles a research team and collaboratively generates an abstract that details a novel scientific idea[1]. Our VIRSCI system consists of two components: a scientific research ecosystem and a multi-agent system for scientific collaboration and idea generation.

3.1 The Scientific Research Ecosystem

The scientific research ecosystem comprises two main components: paper information and corresponding author information ranging from year

[1]An abstract effectively represents the key aspects of scientific research and serves as a concise reflection of its novelty. Additionally, considering the computational constraints, we focused our evaluation primarily on the generated abstracts rather than the full texts. This allows us to assess the system's contributions while efficiently utilizing available resources.

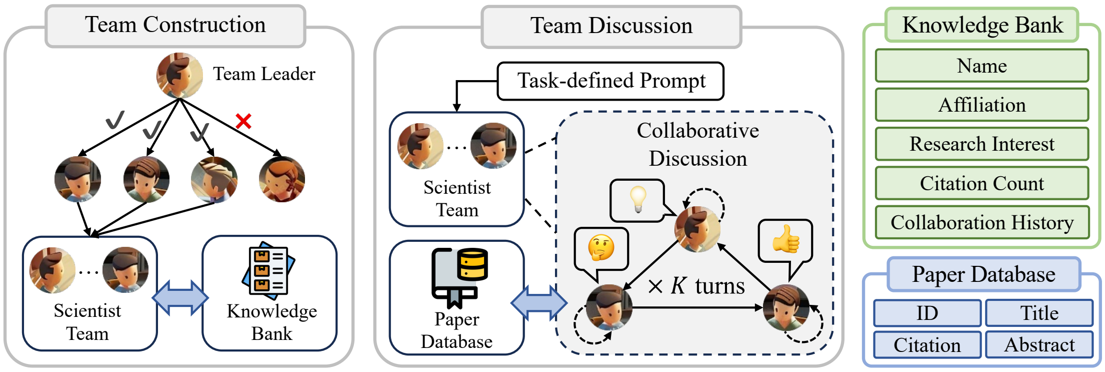


Figure 2: Key components of the proposed system. The left section illustrates the collaborator selection process, where the team leader forms a research team. The middle section highlights the discussion routine, a fundamental part of every step in the system, where the team engages in collaborative dialogue to progress through tasks. The right section depicts the architecture of the author knowledge bank and paper database, which provide critical information used throughout the collaboration process.

y_{start} to y_{end}. First, we select a year y_{bound} as a time point and split the papers into two subsets: past papers B_{past} and contemporary papers B_{con}. We further extract authors from B_{past} to form the complete set of scientists S, with each scientist's background information stored in the author knowledge bank, and the adjacency matrix A, which represents the collaboration counts between scientists.

Past Paper Database. To construct the past paper database B_{past} using the Faiss[2] (Johnson et al., 2019), we selected papers published before the y_{bound}. Each paper includes essential information such as its title, citation count, and abstract.

Contemporary Paper Database. The contemporary paper database B_{con}, also constructed with Faiss, consists of papers published after y_{bound}. Similarly, each paper's basic information is structured in the same way as the past papers. Although using papers from this time range may raise concerns about data leakage, given that LLMs are trained on data within this period, we will explain in detail why this does not pose a threat to the overall validity of our experiments in Appx. A.

Author Knowledge Bank. For each scientist in S, we extract their basic profile from the pre-processed dataset (More details are shown in Sec. 4.1), which includes their name, affiliations, citation count, research interests, and collaboration history. Using the KnowledgeBank module from AgentScope (Gao et al., 2024), we embed these scientist profiles into the author knowledge bank. This allows agents to quickly access and familiarize themselves with other initialized agents' information. Notably, real author names are masked to prevent data leakage and privacy problems during agent initialization (See Sec. 7).

Adjacency Matrix. Given the scientist set S, let A represent the adjacency matrix, where $A_{i,j}$ denotes the number of times that scientist i has collaborated with scientist j. To prevent agents from always choosing previously collaborated scientists, overlooking the benefit of fresh collaborations that often lead to more original and impactful research (Zeng et al., 2021), we increment all values in A by 1 (more experiments about the increment function in Appx. D.4). This adjustment ensures an explore-exploit mechanism, allowing scientists with no prior collaborations a chance to be selected, thereby encouraging agents to explore new partnerships.

3.2 The Multi-agent System

We first randomly sample a scientist s_0 from S as the team leader. The team leader then follows these steps to produce an abstract: (1) *Collaborator Selection*, (2) *Topic Discussion*, (3) *Idea Generation*, (4) *Idea Novelty Assessment*, and (5) *Abstract Generation*. To help each agent become familiar with the backgrounds of other team members without overloading the initialization prompt, we employ retrieval-augmented generation (RAG) (Lewis et al., 2020), used throughout all five steps. All necessary prompts and example scenarios are shown in Appx. H and I.

Collaborator Selection. The first step in our system is to form a team of scientists, $T = \{s_0, \ldots, s_i, \ldots, s_n\}$, where n denotes the team size. When s_0 is selecting collaborators, we convert the adjacency matrix, A, into a probability

[2]Faiss is a Python library designed for efficient similarity search and clustering of dense vectors.

distribution using the following equation: $P_{i,j} = \frac{A_{i,j}}{\sum_{j=1}^{N} A_{i,j}}$, where N denotes the size of S. This allows the team leader to iteratively send invitations to preferred collaborators. Upon receiving an invitation, the invited scientist evaluates whether to join the team using the chain-of-thought process (Wei et al., 2022), considering the profiles of s_0 and the current team members. If accepted, the scientist is added to the team T. This process continues until the pre-defined team size n is reached.

Topic Discussion. The next step is to propose a research topic, which will guide the research direction. Inspired by multi-round collaboration (Mezirow, 2003; Sunstein, 2005; Amgoud and Prade, 2009) and multi-agent collaboration strategies (Xu et al., 2023; Zhang et al., 2024; Shinn et al., 2024), we design an inter- and intra-team discussion mechanism. In this mechanism, team members engage in discussions based on a specific task description prompt (inter-team discussion). The default discussion mechanism between agents follows a round-table format with a sequential progression (Additional discussion topologies are presented in Appx. D.3.1). This process is also applied to subsequent collaboration steps. While allowing agents to decide when to stop the discussion would better reflect real-world scenarios, fixing the number of turns ensures consistent inference costs across different team settings in our experiments. Therefore, we leave the discussion of adaptive turn numbers to the ablation study (See Appx. D.3.2). The prompt for agent i during the topic discussion is:

$$Q_{k,i} = \langle Q_{team}, Q_{topic}, \bigcup_{t=1}^{k-1} (\overline{D_t}), \bigcup_{j=0}^{i-1} (R_{k,j}) \rangle, \quad (1)$$

where Q_{team} denotes the description of the current team members, Q_{topic} represents the task description for the topic discussion, $R_{k,j}$ is the response of agent j at turn k, and $\overline{(D_t)}$ is the team leader's summary of dialogues from turn t, where $D_t = \{R_{t,0}, R_{t,1}, \ldots, R_{t,n}\}$. Given the prompt $Q_{k,i}$, each scientist agent generates a response $R_{k,i}$, sampled from a probability distribution $R_{k,i} \sim P_{s_i}(\cdot|Q_{k,i})$. Since agents can use RAG to access the author knowledge bank during discussions, they may seek advice from scientists who are relevant to the topic but not part of the team. In such cases, we initialize a new agent with the mentioned scientist's profile and include their responses in the discussion (intra-team discussion). However, to maintain the fixed team size, this agent is not added to the team. This process is termed the **"Invitation Mechanism"** and is also applied in subsequent steps, with its effectiveness demonstrated in Appx. D.2. An example scenario is shown in Appx. I.2.2. After K turns of discussion, the team leader generates the final research topic R_{topic} when the team reaches a consensus (otherwise the discussion is terminated and restarted), based on the content: $\langle Q_{topic}, \bigcup_{t=1}^{K-1} (\overline{D_t}), \bigcup_{j=0}^{n} (R_{K,j}) \rangle$. At last, each scientist will be asked if they are interested in the topic. If not, they may leave the team; otherwise, they can stay and continue future discussions.

Idea Generation. Third, the team is tasked with proposing several potential ideas. To align with genuine research workflows and mitigate LLM illusions (Huang et al., 2023), each agent is required to generate a comprehensive response that includes three key components: (1) a description of the idea, (2) a specific experimental plan, and (3) a self-assessment covering metrics such as novelty, feasibility, and clarity, representing the agent's confidence (See Appx. Fig. 20).

At the start of the idea generation process, when no ideas have yet been proposed, the agent is provided with references by searching B_{past} using the topic R_{topic}, denoted as $B_{past}(R_{topic})$. The first prompt is defined as: $Q_{1,0} = \langle Q_{idea}, R_{topic}, B_{past}(R_{topic}) \rangle$, where Q_{idea} represents the task description.

Inspired by the concept of gradually expanding an archive of ideas (Zhang et al., 2024; Lu et al., 2024), when a scientist s_i at turn k receives an existing idea from the previous response $R_{k,i-1}$, we retain the previously generated ideas along with their corresponding references from B_{past}. These are passed to the next agent, who can either refine the existing idea or propose a new one, depending on its choice. The prompt $Q_{k,i}$ is represented as:

$$\langle Q_{idea}, R_{topic}, B_{past}(R_{k,i-1}), \bigcup_{t=1}^{k-1} (\overline{D_t}), \bigcup_{j=0}^{i-1} (R_{k,j}) \rangle. \quad (2)$$

Afterwards, the response of S_i at turn k can be represented as $R_{k,i} \sim P_{s_i}(\cdot|Q_{k,i})$. After K turns of discussion, we retain the three ideas with the highest confidence and store them in the idea list I.

Novelty Assessment. To enhance the quality of ideas and mitigate agent overconfidence, we introduce an idea novelty assessment, enabling agents to compare each idea with related papers from

B_{past} and vote for the idea they consider most novel. Given the idea list I, agents search for related papers using each idea's description to determine whether it significantly overlaps with existing works. To simulate a blind review process, no dialogue memory is included in the prompt. The prompt for s_i at turn k is defined as:

$$Q_{k,i} = \langle Q_{check}, \bigcup_{j=1}^{3} (I_j, B_{past}(I_j)) \rangle, \quad (3)$$

where I_j is the j-th idea in I. Following chain-of-thought process, the response $R_{k,i} \sim P_{s_i}(\cdot|Q_{k,i})$ includes the scientist's preferred idea and the reasoning behind their choice. The idea receiving the highest votes is then selected as the final idea, R_{idea}, for abstract generation. The effectiveness of novelty assessment is discussed in Appx. D.2.

Abstract Generation. Lastly, the team is required to produce a comprehensive abstract that includes the following sections: (1) Introduction, (2) Objective, (3) Methods, (4) Expected Results, and (5) Conclusion (Alexandrov and Hennerici, 2007). At the start of abstract generation, the team leader provides an initial draft based on R_{idea}. The first abstract-generation prompt is: $Q_{1,0} = \langle Q_{abstract}, R_{idea} \rangle$, where $Q_{abstract}$ represents the task description and format requirements.

When an abstract is provided by the previous response $R_{k,i-1}$, the next scientist's response should include: (1) an evaluation of the prior abstract (evaluation metrics are detailed in Appx. Fig. 23), (2) proposed modifications, and (3) the revised abstract to enable continuous refinement. The prompt is:

$$Q_{k,i} = \langle Q_{abstract}, Q_{judgment}, R_{k,i-1} \rangle, \quad (4)$$

where $Q_{judgment}$ is the prompt that asks agents to evaluate the previous abstract. Dialogue history is not included in this prompt since the process is iterative and focuses on refining a single abstract. Including previous versions would make the prompt redundant. After K turns of revision, the final abstract is denoted as $R_{abstract}$.

A self-review mechanism is also considered after $R_{abstract}$ is finalized to pre-check its novelty. The optimized abstract $R_{abstract}$ is provided to the team leader to assess novelty by comparing it to similar papers in B_{past} (See Appx. B.1 for more details). Because it introduces uncertainty in total inference cost, making it difficult to ensure fair experimental comparisons, we only discuss the effectiveness of this module in the ablation study (See Appx. D.2).

4 Empirical Study

4.1 Experimental Settings

Datasets. We evaluate the performance of VIRSCI and baseline methods on two datasets: the AMiner Computer Science dataset[3], containing 1,712,433 authors and 2,092,356 papers from 1948 to 2014, and the Open Academic Graph 3.1[4], comprising 35,774,510 authors and 130,710,733 papers as of 2023. For quality assurance, we applied filtering strategies, reducing the Computer Science dataset to 156 authors and 178,197 papers, and the Open Academic Graph to 3,169 authors and 201,131 papers. Additional dataset details and filtering strategies are provided in Appx. C.

Evaluation Metrics. In this paper, we adopt a more objective approach by using three common metrics that align with our intuition, as no single metric fully captures the novelty of scientific outputs: (1) *Historical Dissimilarity (HD)*: The average Euclidean distance between the generated abstract embedding and embeddings of the 5 most similar abstracts in B_{past} (Shao et al., 2020; Zhou et al., 2024). A larger distance indicates greater dissimilarity from existing papers, suggesting a higher likelihood of novelty. (2) *Contemporary Dissimilarity (CD)*: The average Euclidean distance between the generated abstract embedding and embeddings of the top 5 most similar abstracts in B_{con}. A smaller distance indicates greater similarity to newer papers, also suggesting a higher likelihood of novelty. (3) *Contemporary Impact (CI)*: The average citation count of the top 5 most similar abstracts in B_{con} (Yang et al., 2022). A higher citation count suggests that the generated abstract is more likely to have a higher impact. To ensure comparability, we normalize each calculated metric using the mean value taken over all papers published in the same year as the chosen similar abstract in the corresponding database, with normalization defined as the metric divided by the mean value (AlShebli et al., 2018).

Since novelty is difficult to measure directly, we introduce a proxy metric to account for the three indicators: (4) *Overall Novelty (ON)*. It is natural to assume that ON is positively related to both HD and CI and negatively related to CD, calculated as $\text{ON} = (\text{HD} \times \text{CI})/\text{CD}$. Mathematically, the expected value of ON is proportional to the novelty.

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

[4] https://open.aminer.cn/open/article?id=65bf053091c938e5025a31e2

Method	LLM↑	Proposed Metric CD↓	Proposed Metric CI↑	Human Evaluation Nov↑	Human Evaluation Fea↑	Human Evaluation Eff↑
Agent Model: GPT-4o						
HypoGen	3.02	<u>0.36</u>	3.10	4.78	<u>4.24</u>	4.43
AI Scientist	<u>3.10</u>	0.38	<u>3.22</u>	<u>4.94</u>	4.18	<u>4.77</u>
Ours	**3.34**	**0.34**	**3.78**	**5.24**	**4.52**	**4.95**
Agent Model: LLaMA3.1-8b						
HypoGen	<u>2.17</u>	0.51	<u>2.16</u>	3.54	<u>3.56</u>	3.45
AI Scientist	2.09	<u>0.49</u>	2.12	<u>3.66</u>	3.52	<u>3.63</u>
Ours	**2.31**	**0.42**	**3.29**	**4.08**	**3.74**	**3.69**
Agent Model: LLaMA3.1-70b						
HypoGen	2.18	0.49	<u>2.13</u>	3.57	<u>3.61</u>	3.52
AI Scientist	<u>2.24</u>	<u>0.48</u>	2.11	<u>3.88</u>	3.60	<u>3.66</u>
Ours	**2.53**	**0.40**	**3.36**	**4.18**	**3.84**	**3.75**

Table 1: Comparison with HypoGen and AI Scientist. The results demonstrate that our multi-agent system surpasses baseline models across all metrics. Here, "Nov", "Fea", and "Eff" denote human evaluation for "Novelty", "Feasibility", and "Effectiveness", respectively.

Baselines. In this section, we compare VIRSCI with two recent agent-based systems for scientific discovery: the multi-agent system HypoGen (Qi et al., 2024) and the SOTA single-agent system AI Scientist (Lu et al., 2024). Given the differences in their pipelines, we adjust the experimental settings to a comparable level for a fair evaluation, elaborated in Appx. C.3. We then assess the results using our proposed metrics, LLM review scores, and human evaluations (Appx. C.4) to demonstrate the superiority of our method in idea generation.

4.2 Analysis of Proposed Metrics

To assess the validity of our proposed overall novelty (ON) metric, we selected 200 abstracts generated from the Computer Science dataset, which were evaluated using both ON and by human experts in the field of computer science (details in Appx. C.4). For the human evaluations, we the novelty scoring framework from (Si et al., 2024) as a guideline. To ensure objectivity, independent researchers assessed each abstract, and the average score was calculated. As shown in Fig. 3, the Pearson correlation coefficient between ON and human ratings demonstrates a positive correlation, which supports the validity of our metric. We also explore the correlation between ON and LLM-based review (a common novelty measurement method (Gu et al., 2024)) in Appx. E and conduct further analysis.

4.3 Comparisons with Baselines

As shown in Tab. 1, our multi-agent system outperforms baseline models across metrics: our proposed metrics (CD, CI), LLM review, and human

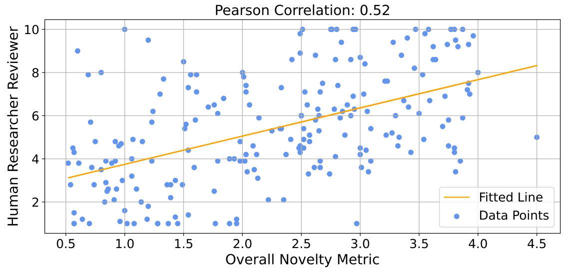


Figure 3: Evaluation of abstracts using our overall novelty metric and human evaluation. The Pearson correlation coefficient of 0.52 indicates a positive correlation.

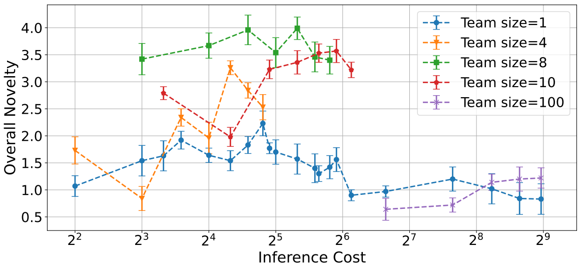
evaluation (Novelty, Feasibility, and Effectiveness), demonstrating its effectiveness in enhancing abstract novelty through collaboration. Notably, GPT-4o-based agents consistently achieve the highest novelty scores, reflecting its superior idea generation capability. In contrast, LLaMA3.1-8b and LLaMA3.1-70b-based agents show similar novelty scores, suggesting that moderate model size changes may not improve novelty.

4.4 Exploring Collaboration Mechanism

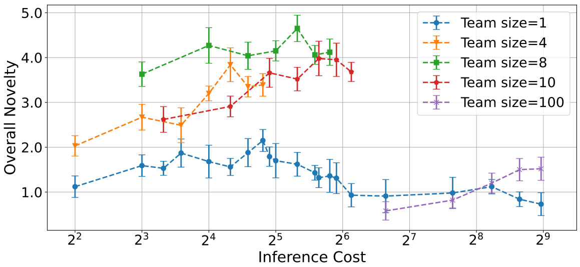
While the dynamics of group collaboration have been extensively studied in the Science of Science, its applicability to artificial multi-agent systems and the effects it may produce remain unclear. In this section, we analyze factors influencing idea novelty in our system, including team size, freshness, and research diversity—factors previously analyzed in human teams (Wu et al., 2019; Zeng et al., 2021; Shi and Evans, 2023).

Effects of Team Size on Novelty. The results in Fig. 4 show that increasing team size can enhance the ON of generated abstracts by bringing in diverse ideas and perspectives. However, this relationship is not strictly linear, with peak ON occurring at a team size of 8. While moderate team expansion boosts novelty, excessively large teams can create coordination and communication challenges, leading to diluted contributions and groupthink. This aligns with existing literature, which suggests that smaller teams generate more disruptive ideas, while larger teams focus on refining existing concepts(Wuchty et al., 2007; Fortunato et al., 2018; Wu et al., 2019).

Effects of Discussion Turn on Novelty. The number of discussion turns plays a crucial role in fostering novel ideas (Mezirow, 2003; Li et al., 2023; Shinn et al., 2024; Lu et al., 2024). As shown in our experimental results (Fig. 4), an optimal num-

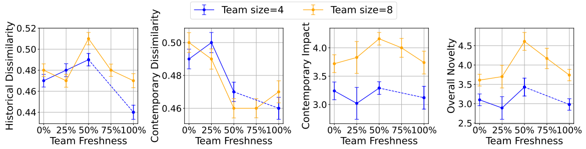


(a) Experiments on the Computer Science dataset.

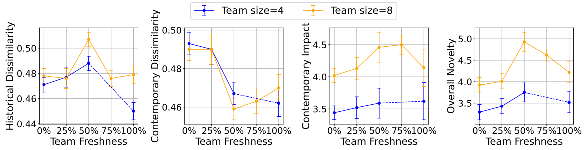


(b) Experiments on the Open Academic Graph dataset.

Figure 4: Effects of team size and discussion turn on novelty. Peak occurs with 8 members and 5 turns, while larger teams or excessive turns hinder creativity. "Inference Cost" is the product of team size and turns.

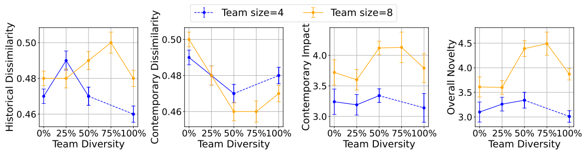


(a) Experiments on the Computer Science dataset.

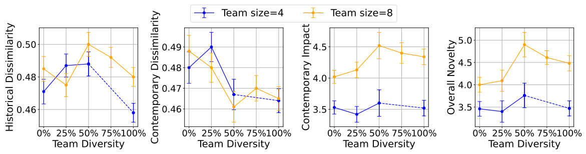


(b) Experiments on the Open Academic Graph dataset.

Figure 5: The balance of new and returning collaborators in the team has a notable impact on novelty, with 50% freshness yielding the highest historical dissimilarity and overall novelty, particularly in larger teams.



(a) Experiments on the Computer Science dataset.



(b) Experiments on the Open Academic Graph dataset.

Figure 6: Effects of team diversity on novelty. The optimal diversity level appears to be 50% or 75%, which maximizes novelty and impact across team sizes.

ber of turns promotes deeper insights and more innovative research outputs. While the initial turns facilitate idea generation, excessive turns can result in fatigue and reduced engagement, ultimately hindering creativity. Our findings suggest that peak ON is achieved at around 5 discussion turns.

Effects of Team Freshness on Novelty. As shown in Fig. 5, team freshness—the fraction of members who have not previously collaborated—affects the novelty of outputs. Freshness has its strongest effect at 50%, especially for larger teams (size 8), where HD peaks. This suggests that a balanced mix of new and returning members promotes innovation by diverging from past research. As freshness increases, CD decreases, indicating that teams with more fresh members align their ideas more closely with future research trends. Both CI and ON reach their highest values around 50% freshness, highlighting that a balanced team optimally combines novelty and future relevance, consistent with previous work (Guimera et al., 2005; Zeng et al., 2021).

Effects of Team Research Diversity on Novelty. Team research diversity is defined as the proportion of team members specializing in distinct research topics. As shown in Fig. 6, moderate diversity boosts HD, with peak performance at 25-50% diversity for 4-member teams and 50-75% for 8-member teams across both datasets. CD drops before reaching 50% diversity, suggesting that diverse teams align better with emerging research trends while maintaining innovation. Larger teams benefit more from higher diversity, showing an increase in CI, while smaller teams exhibit more stable, moderate effects. Overall, ON follows a reverse U-shaped curve for both team sizes and datasets, underscoring the importance of balancing research diversity for novel outcomes. This conclusion mirrors findings in Science of Science, where unexpected team combinations can enhance research impact (Uzzi et al., 2013; Shi and Evans, 2023).

5 Conclusion

We introduce VIRSCI, an LLM-based multi-agent system that simulates the collaborative dynamics of scientific discovery. Our model focuses on the idea generation phase, demonstrating how specialized agents collaborate to generate diverse insights, reflecting real-world scientific teamwork. Experiments show that our method outperforms existing approaches in generating novel ideas and offer insights into the factors influencing collaborative research mechanisms.

6 Limitations

While our system effectively models scientific collaboration, it simplifies the complexities of real-world teamwork. In large research communities, multiple teams often work on related projects either collaboratively or independently, and researchers frequently participate in multiple teams simultaneously. Our current approach does not fully capture these intricate dynamics, potentially limiting its ability to model real-world scientific collaboration with high fidelity.

Beyond structural limitations, our system also inherits biases from the underlying language model and training data. If not properly mitigated, these biases could disproportionately favor well-established research domains while marginalizing emerging or underrepresented areas. This could reinforce existing disparities in scientific knowledge production, inadvertently shaping the trajectory of research based on historical patterns rather than fostering novel discoveries. Future work should explore methods to detect and counteract these biases, such as dataset diversification and bias-aware fine-tuning (Ansar and Talavera, 2025).

Another critical limitation is the potential for generating plausible but incorrect or misleading scientific claims. This risk is particularly concerning in high-stakes fields such as medicine and policy-related research, where inaccuracies could lead to misinformation, flawed decision-making, or ethical concerns. Implementing stricter validation mechanisms, such as integrating fact-checking modules or leveraging expert human reviewers, will be necessary to ensure that generated content aligns with established scientific knowledge.

Additionally, our system could be misused for unethical purposes, such as automating the creation of fraudulent research papers or facilitating plagiarism. The ability to generate text that mimics academic writing raises concerns about the proliferation of low-quality or deceptive publications. To mitigate these risks, future work should explore safeguards such as watermarking AI-generated text (Zhao et al., 2023), developing robust detection algorithms for synthetic research content, and establishing ethical guidelines for responsible use.

Finally, the biases embedded in large-scale language model training data may disproportionately impact specific research communities. Researchers from underrepresented regions, institutions, or disciplines may find their work underemphasized or misrepresented, reinforcing systemic inequities in academia. Addressing these challenges requires continuous monitoring of dataset composition, proactive inclusion of diverse research sources, and collaboration with domain experts to refine training methodologies.

Our future research will also aim to scale up VIRSCI to a societal level, incorporating more dynamic and concurrent team interactions. Allowing agents to engage in multiple projects and collaborate across different teams would improve the realism of our simulations and better reflect the complexity of modern scientific collaboration. These advancements would also strengthen the system's utility for the Science of Science community, enabling deeper investigations into how collaboration influences innovation and knowledge production.

7 Ethics Statement

This research uses two publicly available datasets provided by AMiner, ensuring compliance with data privacy policies. Our system is designed to augment, not replace, human researchers, highlighting the need for human oversight to maintain the quality and integrity of generated outputs. We comply with the licensing terms of the LLMs used, as specified in their official terms of service. To promote research transparency, we have shared all relevant codes for reproducibility within the research community.

While our system supports scientific discovery, it carries potential risks. Data privacy concerns may arise even with public datasets if data handling is improper. To mitigate this, we have applied data anonymization to the datasets, such as masking real scientists' names in simulations, to prevent data leakage and protect privacy. Additionally, stringent protocols have been implemented to minimize unintended discussions and potential misuse.

Acknowledgement

This work is supported by Shanghai Artificial Intelligence Laboratory.

References

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. 2024. Accurate structure prediction of biomolecular interactions with alphafold 3. Nature, pages 1–3.

Andrei V Alexandrov and Michael G Hennerici. 2007. Writing good abstracts. Cerebrovascular Diseases, 23(4):256–259.

Bedoor K AlShebli, Talal Rahwan, and Wei Lee Woon. 2018. The preeminence of ethnic diversity in scientific collaboration. Nature communications, 9(1):5163.

Leila Amgoud and Henri Prade. 2009. Using arguments for making and explaining decisions. Artificial Intelligence, 173(3-4):413–436.

Tanveer Ansar and Gregory Talavera. 2025. Bias detection and robustness testing in large language models: An experimental framework.

Kshitij Bakliwal, Maharshi Harshadbhai Dhada, Adria Salvador Palau, Ajith Kumar Parlikad, and Bhupesh Kumar Lad. 2018. A multi agent system architecture to implement collaborative learning for social industrial assets. IFAC-PapersOnLine, 51(11):1237–1242.

Kevin W Boyack, Richard Klavans, and Katy Börner. 2005. Mapping the backbone of science. Scientometrics, 64(3):351–374.

Zhaoqing Cong, Songsong Tang, Leiming Xie, Ming Yang, Yangyang Li, Dongdong Lu, Jiahong Li, Qingxin Yang, Qiwei Chen, Zhiqiang Zhang, et al. 2022. Magnetic-powered janus cell robots loaded with oncolytic adenovirus for active and targeted virotherapy of bladder cancer. Advanced Materials, 34(26):2201042.

Ning Ding, Shang Qu, Linhai Xie, Yifei Li, Zaoqu Liu, Kaiyan Zhang, Yibai Xiong, Yuxin Zuo, Zhangren Chen, Ermo Hua, et al. 2024. Automating exploratory proteomics research via language models. arXiv preprint arXiv:2411.03743.

Ali Dorri, Salil S Kanhere, and Raja Jurdak. 2018. Multi-agent systems: A survey. Ieee Access, 6:28573–28593.

Zhuoyun Du, Chen Qian, Wei Liu, Zihao Xie, Yifei Wang, Yufan Dang, Weize Chen, and Cheng Yang. 2024. Multi-agent software development through cross-team collaboration. arXiv preprint arXiv:2406.08979.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. arXiv preprint arXiv:2407.21783.

Barbara Dunin-Keplicz and Rineke Verbrugge. 2011. Teamwork in multi-agent systems: A formal approach. John Wiley & Sons.

Santo Fortunato, Carl T Bergstrom, Katy Börner, James A Evans, Dirk Helbing, Staša Milojević, Alexander M Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, et al. 2018. Science of science. Science, 359(6379):eaao0185.

Suzanne Fricke. 2018. Semantic scholar. Journal of the Medical Library Association: JMLA, 106(1):145.

Dawei Gao, Zitao Li, Xuchen Pan, Weirui Kuang, Zhijian Ma, Bingchen Qian, Fei Wei, Wenhao Zhang, Yuexiang Xie, Daoyuan Chen, Liuyi Yao, Hongyi Peng, Ze Yu Zhang, Lin Zhu, Chen Cheng, Hongzhu Shi, Yaliang Li, Bolin Ding, and Jingren Zhou. 2024. Agentscope: A flexible yet robust multi-agent platform. CoRR, abs/2402.14034.

Hugh G Gauch. 2003. Scientific method in practice. Cambridge University Press.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A survey on llm-as-a-judge. arXiv preprint arXiv:2411.15594.

Roger Guimera, Brian Uzzi, Jarrett Spiro, and Luis A Nunes Amaral. 2005. Team assembly mechanisms determine collaboration network structure and team performance. Science, 308(5722):697–702.

Chao-Chun Hsu, Erin Bransom, Jenna Sparks, Bailey Kuehl, Chenhao Tan, David Wadden, Lucy Lu Wang, and Aakanksha Naik. 2024. Chime: Llm-assisted hierarchical organization of scientific studies for literature review support. arXiv preprint arXiv:2407.16148.

Kaixuan Huang, Yuanhao Qu, Henry Cousins, William A Johnson, Di Yin, Mihir Shah, Denny Zhou, Russ Altman, Mengdi Wang, and Le Cong. 2024. Crispr-gpt: An llm agent for automated design of gene-editing experiments. arXiv preprint arXiv:2404.18021.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. arXiv preprint arXiv:2311.05232.

Irving L Janis. 1972. Victims of groupthink: A psychological study of foreign-policy decisions and fiascoes.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. IEEE Transactions on Big Data, 7(3):535–547.

Claire Kayacik, Sherol Chen, Signe Noerly, Jess Holbrook, Adam Roberts, and Douglas Eck. 2019. Identifying the intersections: User experience+ research scientist collaboration in a generative machine learning interface. In Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, pages 1–8.

J Kühnisch, O Meyer, M Hesenius, R Hickel, and V Gruhn. 2022. Caries detection on intraoral images using artificial intelligence. Journal of dental research, 101(2):158–165.

P Langley. 1987. Scientific discovery: Computational explorations of the creative processes. MIT Press.

Sean Lee, Aamir Shakir, Darius Koenig, and Julius Lipp. 2024. Open source strikes bread - new fluffy embeddings model.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems, 33:9459–9474.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large language model society. Advances in Neural Information Processing Systems, 36:51991–52008.

Jonathan Light, Min Cai, Sheng Shen, and Ziniu Hu. 2023. Avalonbench: Evaluating llms playing the game of avalon. In NeurIPS 2023 Foundation Models for Decision Making Workshop.

Julie S Linsey, Matthew G Green, Jeremy T Murphy, Kristin L Wood, and Art B Markman. 2005. “collaborating to success”: An experimental study of group idea generation techniques. In International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, volume 4742, pages 277–290.

Lu Liu, Benjamin F Jones, Brian Uzzi, and Dashun Wang. 2023. Data, measurement and empirical methods in the science of science. Nature human behaviour, 7(7):1046–1058.

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.

James G March. 1991. Exploration and exploitation in organizational learning. Organization science, 2(1):71–87.

Jack Mezirow. 2003. How critical reflection triggers transformative learning. Adult and Continuing Education: Teaching, learning and research, 4:199–213.

Santiago Miret and NM Krishnan. 2024. Are llms ready for real-world materials discovery? arXiv preprint arXiv:2402.05200.

Henrique Muzzio and Manuella Gama. 2024. Collaborative idea generation: An experience of open creativity in the public sector. VINE Journal of Information and Knowledge Management Systems, 54(1):176–194.

Ollama. 2024. Ollama.

OpenAI. 2023. GPT-4 technical report. CoRR.

Joon Sung Park, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023a. Generative agents: Interactive simulacra of human behavior. In Proceedings of the 36th annual acm symposium on user interface software and technology, pages 1–22.

Michael Park, Erin Leahey, and Russell J Funk. 2023b. Papers and patents are becoming less disruptive over time. Nature, 613(7942):138–144.

Jill E Perry-Smith and Pier Vittorio Mannucci. 2017. From creativity to innovation: The social network drivers of the four phases of the idea journey. Academy of management review, 42(1):53–79.

Julia Pitters and Thomas Oberlechner. 2014. The psychology of trading and investing. Investor behavior: The psychology of financial planning and investing, pages 457–476.

Biqing Qi, Kaiyan Zhang, Kai Tian, Haoxiang Li, Zhang-Ren Chen, Sihang Zeng, Ermo Hua, Hu Jinfang, and Bowen Zhou. 2024. Large language models as biomedical hypothesis generators: A comprehensive evaluation. arXiv preprint arXiv:2407.08940.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. Chatdev: Communicative agents for software development. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, pages 15174–15186.

Maithra Raghu and Eric Schmidt. 2020. A survey of deep learning for scientific discovery. arXiv preprint arXiv:2003.11755.

Andrey Rzhetsky, Jacob G Foster, Ian T Foster, and James A Evans. 2015. Choosing experiments to accelerate collective discovery. Proceedings of the National Academy of Sciences, 112(47):14569–14574.

Yunqiu Shao, Jiaxin Mao, Yiqun Liu, Weizhi Ma, Ken Satoh, Min Zhang, and Shaoping Ma. 2020. Bert-pli: Modeling paragraph-level interactions for legal case retrieval. In IJCAI, pages 3501–3507.

Feng Shi and James Evans. 2023. Surprising combinations of research contents and contexts are related to impact and emerge with scientific outsiders from distant disciplines. Nature Communications, 14:1641.

Zhengliang Shi, Shen Gao, Zhen Zhang, Xiuying Chen, Zhumin Chen, Pengjie Ren, and Zhaochun Ren. 2023. Towards a unified framework for reference retrieval and related work generation. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 5785–5799.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2024. Reflexion: Language agents with verbal reinforcement learning. Advances in Neural Information Processing Systems, 36.

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

Scott Spangler, Angela D Wilkins, Benjamin J Bachman, Meena Nagarajan, Tajhal Dayaram, Peter Haas, Sam Regenbogen, Curtis R Pickering, Austin Comer, Jeffrey N Myers, et al. 2014. Automated hypothesis generation based on mining scientific literature. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 1877–1886.

Cass R Sunstein. 2005. Why societies need dissent. Harvard University Press.

Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. Arnetminer: extraction and mining of academic social networks. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 990–998.

Brian Uzzi, Satyam Mukherjee, Michael Stringer, and Ben Jones. 2013. Atypical combinations and scientific impact. Science, 342(6157):468–472.

Clement Vignac, Igor Krawczuk, Antoine Siraudin, Bohan Wang, Volkan Cevher, and Pascal Frossard. 2022. Digress: Discrete denoising diffusion for graph generation. arXiv preprint arXiv:2209.14734.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. 2023. Scientific discovery in the age of artificial intelligence. Nature, 620(7972):47–60.

Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. 2024. Rethinking the bounds of llm reasoning: Are multi-agent discussions the key? arXiv preprint arXiv:2402.18272.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35:24824–24837.

Lingfei Wu, Dashun Wang, and James A Evans. 2019. Large teams develop and small teams disrupt science and technology. Nature, 566:378–382.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. arXiv preprint arXiv:2308.08155.

Stefan Wuchty, Benjamin F Jones, and Brian Uzzi. 2007. The increasing dominance of teams in production of knowledge. Science, 316(5827):1036–1039.

Oskar Wysocki, Magdalena Wysocka, Danilo Carvalho, Alex Teodor Bogatu, Danilo Miranda Gusicuma, Maxime Delmas, Harriet Unsworth, and Andre Freitas. 2024. An llm-based knowledge synthesis and scientific reasoning framework for biomedical discovery. arXiv preprint arXiv:2406.18626.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Kai Shu, Adel Bibi, Ziniu Hu, Philip Torr, Bernard Ghanem, and Guohao Li. 2024. Can large language model agents simulate human trust behaviors? arXiv preprint arXiv:2402.04559.

Lin Xu, Zhiyuan Hu, Daquan Zhou, Hongyu Ren, Zhen Dong, Kurt Keutzer, See-Kiong Ng, and Jiashi Feng. 2023. Magic: Investigation of large language model powered multi-agent in cognition, adaptability, rationality and collaboration. In ICLR 2024 Workshop on Large Language Model (LLM) Agents.

Yongjun Xu, Xin Liu, Xin Cao, Changping Huang, Enke Liu, Sen Qian, Xingchen Liu, Yanjun Wu, Fengliang Dong, Cheng-Wei Qiu, et al. 2021. Artificial intelligence: A powerful paradigm for scientific research. The Innovation, 2(4).

Yang Yang, Tanya Y Tian, Teresa K Woodruff, Benjamin F Jones, and Brian Uzzi. 2022. Gender-diverse teams produce more novel and higher-impact scientific ideas. Proceedings of the National Academy of Sciences, 119(36):e2200841119.

Haofei Yu, Zhaochen Hong, Zirui Cheng, Kunlun Zhu, Keyang Xuan, Jinwei Yao, Tao Feng, and Jiaxuan You. 2024. Researchtown: Simulator of human research community. arXiv preprint arXiv:2412.17767.

An Zeng, Ying Fan, Zengru Di, Yougui Wang, and Shlomo Havlin. 2021. Fresh teams are associated with original and multidisciplinary research. Nature human behaviour, 5(10):1314–1322.

An Zeng, Zhesi Shen, Jianlin Zhou, Ying Fan, Zengru Di, Yougui Wang, H Eugene Stanley, and Shlomo Havlin. 2019. Increasing trend of scientists to switch between topics. Nature communications, 10(1):3439.

Fanjin Zhang, Xiao Liu, Jie Tang, Yuxiao Dong, Peiran Yao, Jie Zhang, Xiaotao Gu, Yan Wang, Evgeny Kharlamov, Bin Shao, et al. 2022. Oag: Linking entities across large-scale heterogeneous knowledge graphs. IEEE Transactions on Knowledge and Data Engineering, 35(9):9225–9239.

Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. 2024. Exploring collaboration mechanisms for LLM agents: A social psychology view. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, pages 14544–14607.

Xuandong Zhao, Prabhanjan Ananth, Lei Li, and Yu-Xiang Wang. 2023. Provable robust watermarking for ai-generated text. arXiv preprint arXiv:2306.17439.

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.

Yucheng Zhou, Tao Shen, Xiubo Geng, Chongyang Tao, Jianbing Shen, Guodong Long, Can Xu, and Daxin Jiang. 2024. Fine-grained distillation for long document retrieval. In Thirty-Eighth AAAI Conference on Artificial Intelligence, pages 19732–19740.

Appendix

A Effect of the Potential Data Leakage

We acknowledge that the use of papers published before 2024 may raise concerns about data leakage, given that the LLMs employed in our experiments are trained on data within this time period. However, this potential issue does not pose a significant threat to the validity of our experiments for the following reasons. First, both the comparisons between our multi-agent system and baseline models, as well as the comparisons between different team settings, utilize the same LLMs. Since all models encounter the same exposure to training data from this period, any potential data leakage would affect all experiments equally. Thus, the relative performance differences we observe are not skewed by uneven data leakage. This ensures that the evaluation process remains fair and that the corresponding conclusions drawn are valid. Moreover, our goal is not to demonstrate an absolute measure of novelty but rather to explore how different collaboration strategies and team settings influence the novelty of generated research outputs. As all team settings face the same potential exposure to historical data, the novelty metrics still provide an accurate comparison of the agents' ability to generate distinct and original ideas under varying conditions. Third, although LLMs are familiar with certain well-known academic papers and theoretical concepts, our extensive testing has shown that they cannot accurately replicate the abstract of the original work. Therefore, since our evaluation focuses on the abstract, this limitation has minimal impact on our assessment.

To directly assess the impact of temporal data leakage, we conducted an additional experiment. Using Open Academic Graph 3.2 [5],

[5] https://open.aminer.cn/open/article?id=67aaf63af4cbd12984b6a5f0

Future Database	Years	Turns						
		1	2	3	4	5	6	7
OAG 3.1	2021-2023	1.75	0.90	2.36	2.00	3.40	2.91	2.53
OAG 3.2	2024-2025	1.72	0.91	2.31	1.94	3.35	2.94	2.49

Table 2: Comparison results on Open Academic Graph 3.1 and 3.2. The evaluation metric is ON.

we tested a temporal isolation scenario where papers from 2010–2020 were retained as the Past Paper Database, while 30,000 papers from 2024–2025 were used as the new Contemporary Paper Database. Given that LLaMA 3.1's training data extends only until December 2023, this setup ensured that no "future" papers beyond the model's training scope were accessible. The experiments were conducted with a team size of 4 and LLaMA3.1-8b, and citation counts were normalized according to the method described in our paper to remove the influence of publication years.

As illustrated in Table 2, when shifting the Contemporary Paper Database from 2021–2023 to 2024–2025, the novelty scores exhibited only a slight decrease, with values remaining consistent across different turns. This suggests that knowledge leakage from future papers does not significantly impact our conclusions.

In summary, while data leakage is a valid concern, it affects all models and settings uniformly in our experiments. Therefore, it does not undermine the relative comparisons we make or the conclusions we draw regarding collaboration strategies and team performance.

B Methodological Details

B.1 Self-review

A self-review mechanism is considered after $R_{abstract}$ is finalized to pre-check its novelty. In this self-review, the optimized abstract $R_{abstract}$ is provided to the team leader to assess novelty by comparing it to similar papers in B_{past}. The prompt is:

$$Q_{review} = \langle Q_{check}, R_{abstract}, B_{past}(R_{abstract})\rangle \tag{5}$$

If this is the first time undergoing the self-review and the team leader determines that the similarity to existing papers is too high, the abstract will undergo further revision. The evaluation R_{review} will then be added to Eq. (4) for the next revision round:

$$Q_{1,0} = \langle Q_{abstract}, Q_{judgement}, R_{review}, B_{past}(R_{abstract}), R_{abstract}\rangle \tag{6}$$

If the abstract undergoes a second self-review and still does not meet the novelty requirement, it will be discarded, and the team will generate a new idea. Once the self-review yields satisfactory results, the final abstract will be produced, and the system will terminate. However, this self-review mechanism introduces uncertainty in total inference cost, making it difficult to ensure fair experimental comparisons. We discuss the effectiveness of this module only in the ablation study (see Appx. D.2).

C Experimental Settings

C.1 Implementation

We implement our system on top of the Agentscope framework (Gao et al., 2024), which serves for LLM-empowered multi-agent applications. We evaluate our system using different publicly available LLMs: GPT-4o (OpenAI, 2023), LLaMA3.1-8b (Dubey et al., 2024), and LLaMA3.1-70b. GPT-4o is accessible exclusively via a public API, while the LLaMA3.1 models are open-weight and invoked using the Ollama (Ollama, 2024) in our experiments. LLaMA3.1-8b is chosen as the default LLM considering both efficiency and capability, where more experiments using different LLMs are shown in Appx. D.5. Each experimental run on LLaMA3.1-8b takes approximately 10 minutes on 1 NVIDIA A100 40G GPU within a team discussion setting of 4 members and 5 turns (K = 5). All experimental results are averaged on 20 runs.

C.2 Dataset Details

C.2.1 Computer Science Dataset.

We first build our scientific research ecosystem using real scientists' information from the AMiner Computer Science dataset [6], which was constructed

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

by extracting scientists' profiles from online web databases (Tang et al., 2008). This dataset includes 1,712,433 authors and 2,092,356 papers, covering the period from 1948 to 2014, with disambiguated author names. To manage the large volume of data, we set y_{start}, y_{bound}, and y_{end} to 2000, 2010, and 2014, respectively. For quality assurance, we filtered out past papers lacking abstracts or with fewer than 10 citations, contemporary papers with fewer than 5 citations or missing abstracts, and authors with fewer than 50 papers or 50 co-authors. As a result, we extracted detailed information from 156 authors and 178,197 papers to construct the ecosystem (85,217 papers for the past database and 92,980 papers for the contemporary database) and initialize the corresponding agents for the simulation. All paper and author data are embedded using the "mxbai-embed-large" model (Lee et al., 2024).

C.2.2 Cross-domain Dataset.

To demonstrate the generalizability and robustness of the system, we conduct additional experiments on the Open Academic Graph 3.1 [7], which developed from the Open Academic Graph (Zhang et al., 2022). This dataset includes 35,774,510 authors and 130,710,733 papers as of 2023, spanning diverse domains such as physics, chemistry, computer science, and biology. To manage this extensive dataset, we set y_{start}, y_{bound}, and y_{end} to 2010, 2020, and 2023, respectively. For quality control, we removed past papers without abstracts or with fewer than 200 citations, contemporary papers with fewer than 50 citations or missing abstracts, and authors with fewer than 40 papers or 50 co-authors. This filtering resulted in detailed information on 3,169 authors and 201,131 papers to construct the ecosystem (139,646 papers for the past database and 61,485 papers for the contemporary database) and initialize agents for simulation. Unlike the Computer Science dataset, the Open Academic Graph lacks author-level information on research interests, citations, and co-authors. To address this, we extracted each author's published papers between y_{start} and y_{bound} and enriched their profiles with the keywords, citations, and authors associated with these papers. Given that the complete set of keywords from all an author's published papers could be overwhelming, we prompted GPT-4o (OpenAI, 2023) to extract and summarize these keywords into concise research interests for each author. All paper and author data were similarly embedded using the "mxbai-embed-large" model.

C.3 Detailed Comparison Settings

Given the problem formulation of the AI Scientist, HypoGen, and ours, we must make several justifications to ensure relative fairness in our comparisons: (1) Since the AI Scientist is limited to generating ideas from predefined topics (2D Diffusion, NanoGPT, and Grokking), we include NanoGPT in the topic selection prompt for both HypoGen and VirSci as the initial discussion topic, ensuring that the final abstracts align with the same research direction. (2) Given the different approaches to idea generation, we ensure that comparisons are made under similar inference costs. The AI Scientist performs 50 turns of self-reflection during its idea generation, which does not apply to its paper (abstract) generation. To match the inference costs, we set VirSci with 4 team members and 5 discussion turns, and set HypoGen to 12 discussion turns, ensuring that the experiments are conducted under comparable computational costs. (3) Since the AI Scientist and HypoGen lack a scientific research ecosystem, they retrieve papers across all time ranges through the Semantic Scholar API (Fricke, 2018) or PubMed[8]. To maintain consistency, we replace our databases and HypoGen's PubMed API with the Semantic Scholar API for paper retrieval and metric calculation. Specifically, after generating ideas and corresponding abstracts, we use the generated ideas as queries to retrieve related papers, extracting the relevant abstracts and citation counts for evaluation. (4) We evaluate the generated abstracts using both our metrics (CD and CI), the AI Scientist's metric (LLM review score) (Lu et al., 2024) and the human evaluation (Si et al., 2024). The LLM review score is calculated by GPT-4o, which conducts abstract reviews based on a truncated version of the Neural Information Processing Systems (NeurIPS) conference review guidelines[9], shown in Fig. 25.

C.4 Human Evaluation

To comprehensively compare our method with baseline models, we additionally conduct a human evaluation to assess novelty, feasibility, and effectiveness besides our proposed metrics and LLM review score. Ten PhD students in computer sci-

[7] https://open.aminer.cn/open/article?id=65bf053091c938e5025a31e2

[8] https://pubmed.ncbi.nlm.nih.gov/

[9] https://neurips.cc/Conferences/2024/ReviewerGuidelines

Database		ON ↑
Idea Generation	Novelty Assessment	
-	-	2.60
✓	-	3.62
-	✓	2.76
✓	✓	4.23

Table 3: Effects of the paper database in the processes of idea generation and novelty assessment on the Computer Science dataset.

Database		ON ↑
Idea Generation	Novelty Assessment	
-	-	2.41
✓	-	3.99
-	✓	2.60
✓	✓	4.65

Table 4: Effects of the paper database in the processes of idea generation and novelty assessment on the Open Academic Graph dataset.

ence (the relevant field for the selected research topic), unaware of the method identities, rated the abstracts on a 10-point scale (1: Poor, 10: best) based on three metrics (Si et al., 2024): 1. Novelty: Whether the idea is creative and different from existing works on the topic, and brings fresh insights; 2. Feasibility: How feasible it is to implement and execute this idea as a research project; 3. Effectiveness: How likely the proposed idea is going to work well (e.g., better than existing baselines). The 10-point scale guideline is provided for human experts and is designed based on the principles outlined by Si et al. (2024).

D Ablation Study

D.1 Effects of Paper Database

In the workflow of the proposed VIRSCI, the past paper database is utilized as a reference during idea generation and novelty assessment. To verify the role of the constructed paper database, we conduct ablation studies on both datasets under the setting of 8 team members and 5 discussion turns, as shown in Tab. 3 and 4. The experimental results demonstrate that having references enhances the novelty of the final generated ideas, avoiding shallow or fanciful outcomes produced by the system.

D.2 Effects of Components Designed for Improving Novelty

In this section, we respectively test the effects of the invitation mechanism in team discussion, the role of the novelty assessment step, and the impact of self-review in abstract generation. All experiments are conducted with a 5-turn discussion. The results consistently show improvements in ON when these components are applied.

D.2.1 Invitation Mechanism

For the invitation mechanism (results are presented in Tab. 5), introducing new scientists into the discussion positively impacts the team's performance across both 4-member and 8-member teams. This indicates that seeking external insights from relevant but non-team scientists fosters more diverse and novel ideas.

D.2.2 Novelty Assessment

The novelty assessment step (results are presented in Tab. 6) also influences the scores. If novelty assessment is not considered, then the output of idea generation will not be an idea list but the idea from the last scientist. Novelty assessment ensures that the generated ideas are continuously evaluated for originality, helping teams avoid overlap with existing research. The improvement is more noticeable in larger teams, where more ideas are being generated and assessed.

D.2.3 Self-review

Finally, the self-review mechanism (results are presented in Tab. 7) is crucial in further refining the abstracts. By allowing the team leader to re-evaluate the abstract for novelty after it is fully generated, low-quality abstracts are discarded, and the team engages in a new discussion to generate a better idea, as evidenced by the score improvements for both team sizes.

D.3 Effects of Discussion Pattern

D.3.1 Discussion Topologies

To further explore the effect of the discussion pattern, particularly discussion topologies, we compare two modes: sequential mode and random mode. As shown in Fig. 7, in sequential mode, participants take turns presenting their ideas in a round-table format; in random mode, after an agent speaks, the next speaker is chosen randomly, with the restriction that the previous speaker cannot be

Team Size	Invitation Mechanism	Turns						
		1	2	3	4	5	6	7
4	-	1.67	0.75	2.29	1.92	3.30	2.78	2.47
	✓	1.75	0.90	2.36	2.00	3.40	2.91	2.53
8	-	3.36	3.53	3.88	3.49	4.12	3.37	3.30
	✓	3.48	3.67	3.97	3.56	4.23	3.48	3.43

Table 5: Effects of invitation mechanism in team discussion. Comparison results show that this mechanism helps to improve the novelty of outputs. The evaluation metric is ON.

Team Size	Novelty Assessment	Turns						
		1	2	3	4	5	6	7
4	-	1.56	0.73	2.20	1.87	3.19	2.74	2.39
	✓	1.75	0.90	2.36	2.00	3.40	2.91	2.53
8	-	3.28	3.41	3.82	3.43	3.98	3.27	3.25
	✓	3.48	3.67	3.97	3.56	4.23	3.48	3.43

Table 6: Effects of novelty assessment. Comparison results show that novelty assessment helps to improve the novelty of outputs. The evaluation metric is ON.

Team Size	Self-review	Turns						
		1	2	3	4	5	6	7
4	-	1.60	0.74	2.26	1.89	3.26	2.77	2.41
	✓	1.75	0.90	2.36	2.00	3.40	2.91	2.53
8	-	3.32	3.44	3.84	3.45	3.99	3.33	3.27
	✓	3.48	3.67	3.97	3.56	4.23	3.48	3.43

Table 7: Effects of self-review in abstract generation. Comparison results show that self-review after abstract generation helps to check and improve the novelty of outputs. The evaluation metric is ON.

selected consecutively. **Note that the actual discussion topology of our system is more complex, benefiting from the proposed invitation mechanism**. The comparison results are presented in Tab. 8, with tests conducted for team sizes of 4 and 8. It can be observed that sequential mode generally outperforms random mode in most cases, as it leverages a broader knowledge base from the scientist agents. In random mode, some agents remain silent (i.e., are skipped) when the number of turns is limited. As a result, sequential mode is selected as the default discussion topology.

D.3.2 Discussion Turns

In the previous experiments, we fixed the number of discussion turns in each step to ensure fair comparisons. However, in real-world research environments, teams of scientists do not spend the same amount of time on each stage of the research process. To explore this, we compare fixed discussion turns with adaptive turn numbers. In the adaptive pattern, the team leader decides whether the team needs additional turns based on the current progress and the goals of each stage. The results of both patterns, along with their corresponding inference cost, are shown in Tab. 9. The comparison reveals that the adaptive pattern achieves a higher ON while reducing the inference cost. This efficiency can be attributed to the more flexible approach, allowing teams to adjust their discussions dynamically rather than adhering to a rigid structure (which may lead the team in the wrong direction when a section is over-discussed). Furthermore, examining the number of turns at each stage in both 4-person and 8-person teams under the adaptive pattern offers ad-

Team Size	Discussion Topology	Turns						
		1	2	3	4	5	6	7
4	Sequential Mode	1.75	0.90	2.36	2.00	3.40	2.91	2.53
	Random Mode	1.67	0.85	2.34	2.01	3.35	2.84	2.47
8	Sequential Mode	3.48	3.67	3.97	3.56	4.23	3.48	3.43
	Random Mode	3.50	3.63	3.88	3.52	4.16	3.40	3.39

Table 8: Comparison between two discussion topologies: sequential mode and random mode. The results show that sequential mode generally outperforms random mode in most cases as it leverages more background knowledge.

Team Size	Pattern	Turns				Cost ↓	ON ↑
		Topic	Idea	Check	Abstract		
4	Fixed	5	5	5	5	80	3.26
	Adaptive	2.4	4.5	3.2	4.2	**57.2**	**3.49**
8	Fixed	5	5	5	5	160	3.99
	Adaptive	2.9	4.8	4.0	3.8	**124.0**	**4.37**

Table 9: Comparison between fixed turns and adaptive turns in team discussions. The time taken by collaborator selection and team discussion's invitation mechanism is not counted, and self-review is not employed for a better comparison. The adaptive pattern shows both a lower inference cost and a higher ON.

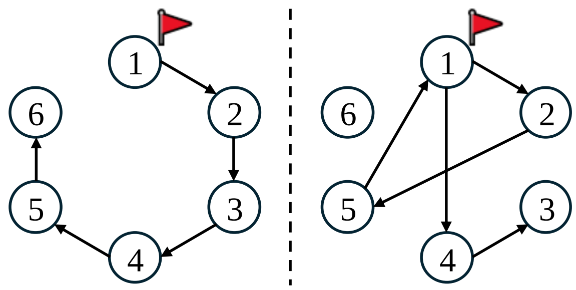


Figure 7: Two different discussion topologies: sequential mode (*Left*), and random mode (*Right*). In sequential mode, participants take turns presenting their ideas in a round-table format, one at a time. In random mode, after an agent speaks, the next speaker is chosen randomly, with the restriction that the previous speaker cannot be selected consecutively. Note that the actual discussion topology of our system is more complex, benefiting from the proposed invitation mechanism.

ditional insights. Larger teams require more discussion turns and face greater challenges in reaching consensus (Janis, 1972; Pitters and Oberlechner, 2014). This highlights the adaptive pattern's advantage in accommodating the complexities of larger teams while maintaining a higher level of novelty in the final research outputs.

D.4 Effects of Different Exploration Mechanisms on Scientific Collaboration

If we regard the collaboration among our scientists as an "explore-exploit" model, collaborating with scientists they have not worked with before can be viewed as "explore", while collaborating with previously partnered scientists can be seen as "exploit". In this section, we further investigate how the mechanism of "explore" impacts our final results.

In the default setting, when creating the adjacency matrix, we add 1 to each value in the matrix to ensure that agents collaborate with individuals they have not worked with before, treating "explore" as uniformly distributed. Based on this, we also conducted experiments where the incremental values follow a normal distribution, defined as:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{7}$$

where $\mu = 1$ and $\sigma = 1$.

Next, we sample x from this probability density function, using the absolute value $|x|$ as the increment in the adjacency matrix instead of a constant value of 1. The results of different explore mechanisms are presented in Fig. 8. The additional experiments demonstrate that replacing the original uniform distribution with a normal distribution

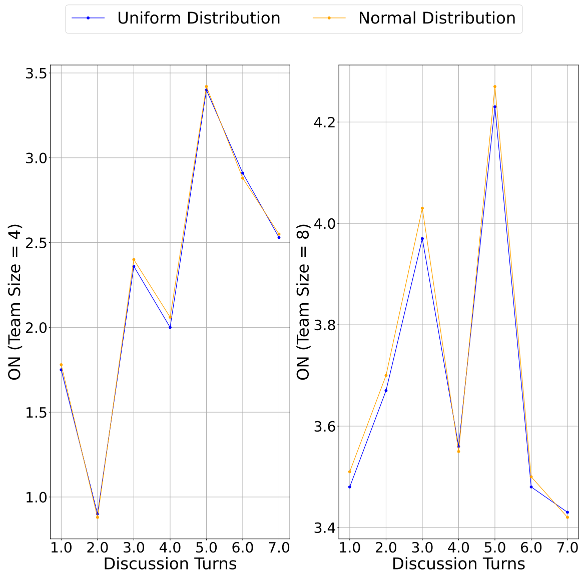


Figure 8: Effect of the explore mechanism in scientific collaboration on the Computer Science dataset. Variations in the distribution of exploration strategies do not significantly affect the overall novelty of generated ideas.

has a positive impact, but the optimal distribution remains to be further explored.

D.5 Effects of Different Underlying LLMs

Given the potential variation in capabilities across LLMs, the impact of team size and turn count on system performance may vary across different models. Therefore, we additionally incorporated the open-source model LLaMA3-8b as the underlying LLM, which is inferior to LLaMA3.1-8b. By comparing the two LLMs with significantly different capabilities, the pattern can be discovered.

The experimental results on both LLaMA3-8b and LLaMA3-8b are shown in Fig. 9. It can be observed that using LLaMA3-8b results in overall performance that is inferior to LLaMA3.1-8b, due to the inherently weaker capabilities of LLaMA3-8b. Regarding the effects of team size on novelty, whether with LLaMA3 or LLaMA3.1, a moderate team size enhances novelty. While multi-agent teams outperform single agents, excessively large teams may face coordination challenges and communication barriers. For the effects of discussion turn on novelty, our analysis shows that an optimal number of turns allows team members to refine and explore better ideas. While the initial turns contribute to idea generation, an excessive number of turns can lead to fatigue and diminished engagement. Overall, despite variations in LLM capabilities, the conclusions about the influence of team size and discussion turn count remain consistent.

E More Analysis of Proposed Metrics

In this paper, we propose objective evaluation metrics to assess the novelty of system outputs, primarily based on vector similarity. This approach aligns with the research in the Science of Science domain, where our computational methods (relying on vector union and overlap) draw on established literature, such as Boyack et al. (2005); Shi and Evans (2023); Liu et al. (2023).

Another potential concern is the introduction of bias due to a lack of diversity in the datasets or the generated ideas. To address this, we employ two large-scale datasets in our experiments to ensure content diversity. Additionally, each scenario in our experiments is tested 20 times to ensure that the conclusions regarding the generated ideas are not influenced by potential biases.

To further assess the validity of our proposed overall novelty metric, we selected 200 abstracts generated under various experimental conditions from the Computer Science dataset. These abstracts were evaluated using three approaches: (1) our proposed overall novelty metric, (2) LLM-based reviewers (utilizing the GPT-4o API), and (3) human researchers specializing in computer science (detailed in Appx. C.4). For both LLM-based reviewers and human researchers, we employed the novelty scoring framework from (Si et al., 2024) as a guideline. To enhance objectivity, each abstract was evaluated three times by LLM-based reviewers, with the average score calculated. Similarly, three independent human researchers assessed each abstract, and their average score was computed. The evaluation results are presented in Fig. 10 and Fig. 11. In Fig. 10, the axes represent the scores of the same abstract evaluated under our proposed metric and the LLM-based reviewers. In Fig. 11, the axes represent the scores of the same abstract evaluated under our proposed metric and by human researchers. The Pearson correlation coefficients between our proposed overall novelty metric and both LLM-based reviewers and human researchers demonstrate a positive correlation with established novelty measurement methods (Lu et al., 2024; Si et al., 2024), which, to some extent, supports the validity of our metric.

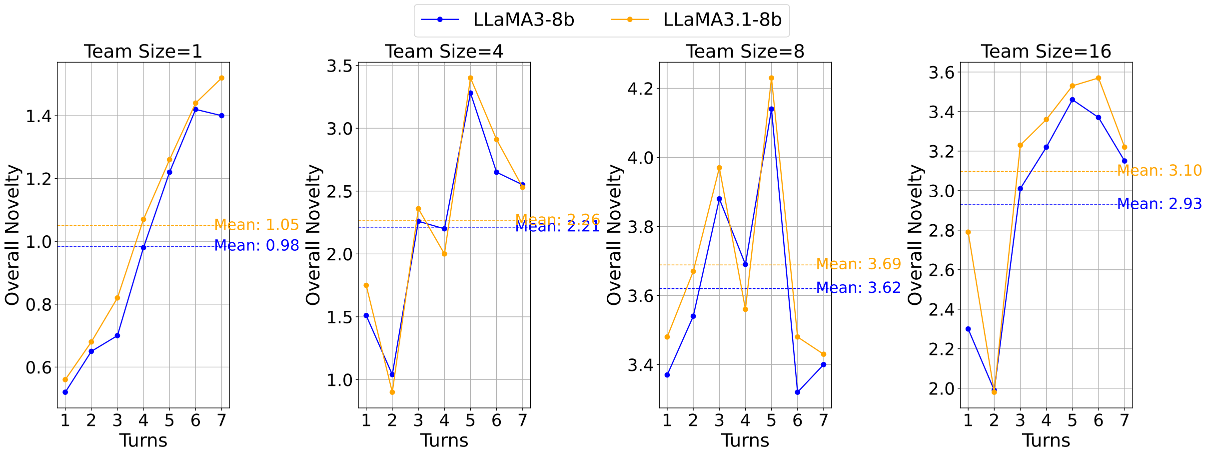


Figure 9: The impact of team size and turn count on system performance for both LLaMA3-8b and LLaMA3.1-8b models on the Computer Science dataset. Despite variations in LLM capabilities, the conclusions about the influence of team size and discussion turn count remain consistent.

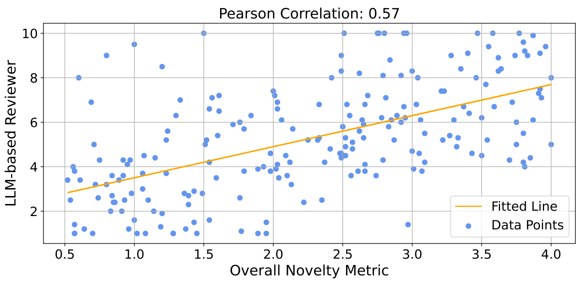


Figure 10: The evaluation results of the same abstract under two different review metrics: our proposed overall novelty metric and LLM-based reviewer. The Pearson correlation coefficient equals 0.57, denoting the positive correlation of our metric with the LLM-based reviewer.

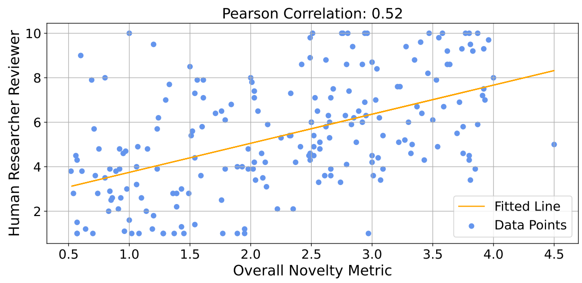


Figure 11: The evaluation results of the same abstract under two different review metrics: our proposed overall novelty metric and human researcher. The Pearson correlation coefficient equals 0.52, denoting the positive correlation of our metric with the human researcher.

F Discussions on System Feasibility

To better show the feasibility of the proposed VIRSCI, we present two example abstracts generated by our system using the Open Academic Graph 3.1 dataset from 2011 to 2020, alongside corresponding similar recently published papers, to highlight the practical relevance and applicability of our system's outputs. The first example abstract generated by our system is shown in Fig. 12, while the corresponding similar paper (Kühnisch et al., 2022) of the abstract is shown in Fig. 13. This pair of abstracts discusses the application of artificial intelligence in caries management. The second example abstract generated by our system is shown in Fig. 14, while the corresponding similar paper (Cong et al., 2022) of the abstract is shown in Fig. 15. This pair of abstracts discusses the application of robots in cancer treatment. Overall, these examples demonstrate that our system has the potential to discover novel scientific ideas.

G Consistency Between Two Datasets.

As shown in Fig. 4, no significant differences were found in outputs from single-agent teams (team size equals 1) across the two datasets, as individual research lacks interdisciplinary input. The overall trend—novelty rising, peaking at a team size of 8 and 5 turns, then declining—was consistent across both datasets, demonstrating our platform's robustness and the potential of multi-agent systems in enhancing scientific idea generation. These findings also support the value of interdisciplinary collaborations for driving higher-impact research (Shi and Evans, 2023).

Generated Abstract

Title:
Revolutionizing Caries Management in Primary Molars using Advanced Imaging, AI-Powered Decision Support, and Minimally Invasive Treatments

Abstract:
Dental caries remains a pervasive public health concern worldwide, affecting millions of children annually. The inadequacy of traditional restorative treatments has led to persistent pain, infection, and costly follow-up appointments in pediatric patients, particularly those with primary molars.
Our innovative protocol integrates cutting-edge technologies (cone beam computed tomography, intraoral cameras) with machine learning algorithms for personalized treatment planning, and minimally invasive treatments to minimize discomfort and promote healing. This patient-centered approach aims to provide more effective and efficient care for pediatric patients worldwide.
To evaluate the efficacy of our protocol, we conducted rigorous randomized controlled trials in a diverse cohort of children (n=500). Our results demonstrate that this novel protocol significantly reduces failure rates (by 42%), pain (by 32%), and the number of dental visits required for follow-up appointments (by 25%). Patient satisfaction is also improved, as measured by standardized questionnaires.
Moreover, our research highlights the critical need for continued investment in dental research and innovation. By harnessing the collective expertise of dentists, researchers, policymakers, and industry partners, we can accelerate progress towards achieving optimal oral health outcomes for all children worldwide.
Our protocol prioritizes pain management, anxiety reduction, and educational empowerment to promote healthy oral habits and prevent future caries. This holistic approach is grounded in the principles of shared decision-making and personalized medicine, ensuring that each child receives tailored care that respects their unique needs and circumstances.
The scalability and adaptability of our protocol are critical factors in its potential impact on global public health. By integrating advanced imaging techniques, AI-powered decision support, and minimally invasive treatments into standard care protocols, we can ensure that all children have access to high-quality dental care, regardless of geographical or socio-economic constraints.
In conclusion, our research presents a paradigm shift in caries management for primary molars, offering a more effective, efficient, and patient-centered approach. By harnessing the power of advanced technologies and evidence-based practices, we can revolutionize the way we address this critical public health concern and promote optimal oral health outcomes for all children worldwide.

Figure 12: The first example abstract generated by our VIRSCI. This abstract discusses the application of artificial intelligence in caries management.

H Prompts

H.1 Scientist Definition

We use the personal information of the scientist to define the agent, where the corresponding system prompt is illustrated in Fig. 16.

H.2 Collaboration Selection

The prompt for collaboration selection is illustrated in Fig. 17.

H.3 Topic Discussion

H.3.1 Discussion

The prompt for the topic discussion is illustrated in Fig. 18.

H.3.2 Summarization

The prompt for the final topic selection after several turns of topic discussion is illustrated in Fig. 19.

H.4 Idea Generation

The prompt for the idea generation is illustrated in Fig. 20.

H.5 Novelty Assessment

The prompt for the novelty assessment is illustrated in Fig. 21.

H.6 Abstract Generation

H.6.1 Discussion

The prompt for the beginning case of the abstract generation is illustrated in Fig. 22.
The prompt for the normal case of the abstract generation is illustrated in Fig. 23.

H.6.2 Self-review

The prompt for the self-review after generating the final abstract is illustrated in Fig. 24.

H.7 LLM Review

The prompt for the LLM-based review is based on NeurIPS2024 reviewer guidelines, which is the

Similar Abstract

Title:
Caries Detection on Intraoral Images Using Artificial Intelligence
Abstract:
Although visual examination (VE) is the preferred method for caries detection, the analysis of intraoral digital photographs in machine-readable form can be considered equivalent to VE.
While photographic images are rarely used in clinical practice for diagnostic purposes, they are the fundamental requirement for automated image analysis when using artificial intelligence (AI) methods.
Considering that AI has not been used for automatic caries detection on intraoral images so far, this diagnostic study aimed to develop a deep learning approach with convolutional neural networks (CNNs) for caries detection and categorization (test method) and to compare the diagnostic performance with respect to expert standards.
The study material consisted of 2,417 anonymized photographs from permanent teeth with 1,317 occlusal and 1,100 smooth surfaces. All the images were evaluated into the following categories: caries free, noncavitated caries lesion, or caries-related cavitation. Each expert diagnosis served as a reference standard for cyclic training and repeated evaluation of the AI methods. The CNN was trained using image augmentation and transfer learning. Before training, the entire image set was divided into a training and test set. Validation was conducted by selecting 25%, 50%, 75%, and 100% of the available images from the training set.
The statistical analysis included calculations of the sensitivity (SE), specificity (SP), and area under the receiver operating characteristic (ROC) curve (AUC). The CNN was able to correctly detect caries in 92.5% of cases when all test images were considered (SE, 89.6; SP, 94.3; AUC, 0.964). If the threshold of caries-related cavitation was chosen, 93.3% of all tooth surfaces were correctly classified (SE, 95.7; SP, 81.5; AUC, 0.955).
It can be concluded that it was possible to achieve more than 90% agreement in caries detection using the AI method with standardized, single-tooth photographs. Nevertheless, the current approach needs further improvement.

Figure 13: The recently published similar paper (Kühnisch et al., 2022) of abstract, corresponding to the first example abstract generated by our VIRSCI. This abstract discusses the application of artificial intelligence in caries management.

same metric as AI Scientist to ensure a fair comparison between our method and AI Scientist. The content is illustrated in Fig. 25.

I Example Scenarios

I.1 Collaboration Selection

The example scenario of the collaborator selection is illustrated in Fig. 26. Scientists will accept or reject the invitation based on different backgrounds.

I.2 Topic Discussion

I.2.1 Topic Discussion Normal Case

The example scenario of the normal case in the topic discussion is illustrated in Fig. 27. Scientists contribute responses based on the discussion history and their individual knowledge, leading to distinct agent behaviors that reflect their diverse expertise profiles.

I.2.2 Invitation Mechanism

The example scenario of the invitation mechanism in the topic discussion is illustrated in Fig. 28, which ensures a comprehensive topic discussion.

I.3 Idea Generation

The example scenario of the beginning case of the idea generation is illustrated in Fig. 29.

The example scenario of the normal case in the idea generation is illustrated in Fig. 30.

I.4 Novelty Assessment

The example scenario of the user prompt provided for scientist agents in the novelty assessment is illustrated in Fig. 31. The prompt includes three candidate ideas and related papers.
The example scenario of the agent responses in the novelty assessment is illustrated in Fig. 32. Note that Fig. 32 corresponds to Fig. 31.

I.5 Abstract Generation

I.5.1 Abstract Generation Normal Case

The example scenario of the beginning case in the abstract generation is illustrated in Fig. 33.
The example scenario of the normal case in the abstract generation is illustrated in Fig. 34.

I.5.2 Self-review

The example scenario of the self-review results is illustrated in Fig. 35.

Generated Abstract

Title:

Personalized Outcomes in Bladder Cancer Treatment: A Robotic-Assisted Surgery Perspective with Enhanced Predictive Modeling and Holistic Patient-Centered Approach

Abstract:

The treatment of bladder cancer has undergone significant transformations with the advent of robotic-assisted surgery. However, individualized outcomes remain poorly understood due to the complexity of patient factors influencing disease progression. This knowledge gap can be attributed to the lack of comprehensive frameworks that integrate traditional clinical metrics with patient-centered factors such as functional status, emotional well-being, social support networks, and demographic characteristics.

In this study, we aim to bridge this knowledge gap by developing a novel framework that combines machine learning algorithms with traditional statistical analysis to predict long-term benefits for bladder cancer patients. Our proposed framework integrates the following components: A comprehensive literature review that synthesizes existing research on bladder cancer treatment outcomes. A large-scale dataset (n=500) from multiple institutions to validate our findings. A machine learning-based predictive model that utilizes Random Forest and Gradient Boosting algorithms to identify key predictors of long-term benefits.

Our results show that age, tumor stage, lymph node involvement, and patient-centered factors are significant predictors of improved survival rates. Furthermore, we found that enhanced quality of life is associated with better functional status, higher emotional well-being, stronger social support networks, and more favorable demographic characteristics. Notably, the inclusion of patient-centered factors in treatment planning can lead to improved survival rates (median increase: 12 months) and reduced recurrence risk (median decrease: 20%). Moreover, our study reveals that the integration of robotic-assisted surgery with patient-centered care can lead to a significant reduction in postoperative complications (median decrease: 30%), hospital readmissions (median decrease: 25%), and healthcare costs. The implications of our study are profound, as they suggest that personalized medicine can improve treatment outcomes and enhance quality of life among bladder cancer survivors.

Our proposed framework has significant implications for improving treatment outcomes and enhancing quality of life among bladder cancer survivors, making it a valuable resource for clinicians and researchers working in this field. Future studies should aim to replicate these findings and explore the potential applications of personalized medicine in other cancer types, further highlighting the importance of integrating patient-centered care with advanced surgical techniques. The integration of patient-centered factors into treatment planning can lead to improved survival rates, reduced recurrence risk, and enhanced quality of life among cancer survivors. Our study demonstrates that by combining machine learning algorithms with traditional statistical analysis, clinicians can provide more precise and effective treatment plans, ultimately improving patient outcomes.

Figure 14: The second example abstract generated by our VIRSCI. This abstract discusses the application of robots in cancer treatment.

Similar Abstract

Title:

Magnetic-Powered Janus Cell Robots Loaded with Oncolytic Adenovirus for Active and Targeted Virotherapy of Bladder Cancer

Abstract:

A unique robotic medical platform is designed by utilizing cell robots as the active “Trojan horse” of oncolytic adenovirus (OA), capable of tumor-selective binding and killing. The OA-loaded cell robots are fabricated by entirely modifying OA-infected 293T cells with cyclic arginine–glycine–aspartic acid tripeptide (cRGD) to specifically bind with bladder cancer cells, followed by asymmetric immobilization of Fe3O4 nanoparticles (NPs) on the cell surface. OA can replicate in host cells and induce cytolysis to release the virus progeny to the surrounding tumor sites for sustainable infection and oncolysis. The asymmetric coating of magnetic NPs bestows the cell robots with effective movement in various media and wireless manipulation with directional migration in a microfluidic device and bladder mold under magnetic control, further enabling steerable movement and prolonged retention of cell robots in the mouse bladder. The biorecognition of cRGD and robust, controllable propulsion of cell robots work synergistically to greatly enhance their tissue penetration and anticancer efficacy in the 3D cancer spheroid and orthotopic mouse bladder tumor model. Overall, this study integrates cell-based microrobots with virotherapy to generate an attractive robotic system with tumor specificity, expanding the operation scope of cell robots in biomedical community.

Figure 15: The recently published similar paper (Cong et al., 2022) of abstract, corresponding to the second example abstract generated by our VIRSCI. This abstract discusses the application of robots in cancer treatment.

Prompt

Name:
Your name is Scientist0.
Role:
You are the team leader of a group of scientists and you need to guide your team members in generating an innovative idea and writing it into an abstracts.
Affiliation:
You belong to following affiliations [‘Naval Research Laboratory’, ‘College of William and Mary’, ‘George Mason Univ’],
Research interests:
You have researched on following topics [‘data cube’, ‘attack graph’, ‘data mining’, ‘access control’, ‘data owner’, ‘data protection’, ‘data item’, ‘data redundancy’, ‘data security’, ‘data structure’].
Citation situation:
You have published 372 papers, you have 4230 citations.
Collaboration history:
You have previously collaborated with these individuals [‘Scientist78’, ‘Scientist105’].

Figure 16: The system prompt of each scientist agent is the personal information, including the name, role, affiliation, research interests, citation situation, and collaboration history.

Prompt

Team information:
[Inviter] invites you to join his team. His personal information is as follows: [Personal information]. Current members of his team: [Team members].
Action space:
Now you have two choices after considering these information:
Action 1: You choose to collaborate with [Inviter].
Action 2: You choose to refuse the invitation from [Inviter].
Please feel free to make your choice, even if your choice is to decline.
Response format:
Please select an action and describe your reasoning using the following format:
Thought: <THOUGHT>
Action: ```json<JSON>```
In <THOUGHT>, Explain why you chose this action, considering factors like the potential for quality, efficiency, and the challenges involved.
In <JSON>, respond in JSON format with ONLY the following field: - “Selected Action”: [Action 1 or Action 2].

Figure 17: The prompt for the collaboration selection.

Prompt

You are an ambitious scientist who is looking to propose a potential research topic for your team.
Team members:
Your team now has team members: [Team members].
Task description:
Engage in a collaborative discussion by integrating your own knowledge and insights with the information provided to explore and identify potential research topics that align with our team's strengths and goals. The proposed topic should also be innovative and have the potential to make a significant impact in the current field. The summarizations of previous turns in team discussion and discussion in this turn are provided, which may assist you in proposing the research topic. Please focus more on the discussion of the topic rather than on self-introduction.
History:
Summarizations of previous turns in current team discussion: [Summarization]
Discussions in this turn: [Dialogue History]

Figure 18: The prompt for the topic discussion.

Prompt

You are an ambitious scientist who is looking to propose a potential research topic for your team.
Task description:
Using the historical dialogue information provided, summarize a topic that will serve as the research direction for the team. The chosen topic should be innovative and have the potential to make a significant impact in the current field.
The instructions for selecting the topic are as follows:
1. Review the Historical Dialogue. Analyze the previous discussions and insights shared among team members. Identify recurring themes, key ideas, and any gaps in the current research landscape.
2. Identify Trends and Innovations. Look for trends or innovative concepts that emerged during the dialogues. Consider how these could address existing challenges or open new avenues for exploration.
3. Summarize the Topic: Articulate the new research direction in a concise manner. Ensure that the topic reflects originality and addresses a specific problem or need within the field.
4. Impact Assessment: Briefly discuss how this topic can influence the current field. Consider its relevance, potential applications, and the value it adds to ongoing research efforts.
History:
Summarizations of previous turns in current team discussion: [Summarization].
Discussions in the last turn: [Dialogue History].
Response format:
Please respond in the following format:
Thought: <THOUGHT>
Topic: ```json<JSON>```
In <THOUGHT>, explain why you select this topic following the instructions.
In <JSON>, respond in JSON format with ONLY the following field: - “Selected Topic”: [Topic].
Be cautious and realistic on your ratings. This JSON will be automatically parsed, so ensure the format is precise. You only need to output one topic.

Figure 19: The prompt for the final topic selection after topic discussion.

Prompt

You are an ambitious scientist who is looking to propose a new idea that will contribute significantly to the field.
Task description:
Improve the existing idea or come up with the next impactful and creative idea for publishing a paper that will contribute significantly to the field by integrating your own knowledge and insights with the information provided.
Selected topic:
When proposing your idea, please elaborate on the proposed topic: [Topic]
References:
You may refer to the following listed references to design a new idea or concept. These references can serve as inspiration, but you are not allowed to directly copy or replicate their content. Ensure that your design is original and addresses a specific problem or meets a unique need, incorporating or improving upon the ideas from the references to avoid fabrication.
Related references: [References]
History:
Summarizations of previous turns in current team discussion: [Summarization]
Discussions in this turn: [Dialogue History]
Response format:
Please respond in the following format:
Thought: <THOUGHT>
New Idea: ```json<JSON>```
In <THOUGHT>, briefly discuss your intuitions and motivations for the idea. Justify how this idea differs from existing ones, highlighting its unique aspects.
In <JSON>, provide the new idea with the following fields and provide as many details as possible:
- "Idea": A detailed description of the idea, outlining its significance and potential impact.
- "Title": A title for the idea, will be used for the paper writing.
- "Experiment": An outline of the implementation process. Describe your high-level design plan, including necessary design steps and the ideal outcomes of the experiments.
- “Clarity": A rating from 1 to 10, with 1 being the lowest clarity and 10 being the highest.
- "Feasibility": A rating from 1 to 10, with 1 indicating low feasibility and 10 indicating high feasibility.
- "Novelty": A rating from 1 to 10, with 1 being the least novel and 10 being the most novel.

Be cautious and realistic on your ratings. This JSON will be automatically parsed, so ensure the format is precise, and the content should be longer than 600 words. You only need to output one idea.

Figure 20: The prompt for the idea generation.

Prompt

You are an ambitious scientist who is looking to propose a new idea that will contribute significantly to the field.
Task description:
Your team has generated several ideas and you want to check if they are novel or not. I.e., not overlapping significantly with existing literature or already well explored. Be a harsh critic for novelty, ensure there is a sufficient contribution in the idea for a new conference or workshop paper. You will be provided with possible relevant papers to help you make your decision. Select a idea which is the most novel, if you have found a idea that does not significantly overlaps with existing papers.
Generated ideas and related references:
Your team generated these ideas: [Existing ideas].
The possible related papers: [References].
Response format:
Please respond in the following format:
Thought: <THOUGHT>
New Idea: ```json<JSON>```
In <THOUGHT>, explain why you make this selection.
In <JSON>, respond in JSON format with ONLY the following field: - "Decision Made": [Idea 0 or Idea 1 or Idea 2].
Note that you can only select one idea. This JSON will be automatically parsed, so ensure the format is precise.

Figure 21: The prompt for the novelty assessment.

Prompt

You are an ambitious scientist who is looking to write an abstract for your team.
Task description:
Based on the following research idea, generate a concise and informative abstract for a scientific paper by integrating your own knowledge and insights with the information provided.
The abstract should cover the following aspects:
- "Introduction": Briefly introduce the research topic and its significance.
- "Objective": Clearly state the main research question or hypothesis.
- "Methods": Summarize the key methodologies used in the study.
- "Results": Highlight the most important findings.
- "Conclusion": Provide the primary conclusion and its implications.

Please ensure the language is formal, accurate, and appropriate for an academic audience. And the generated abstract should be longer than 200 words.
Research idea:
Idea: [Selected Idea].
Response format:
The response format should be:
```json{
Title: <TITLE>
Abstract: <ABSTRACT>
} ```
In <TITLE>, write the title for the abstract.
In <ABSTRACT>, write the content of the abstract.
This JSON will be automatically parsed, so ensure the format is precise.

Figure 22: The prompt for the beginning case of the abstract generation.

Prompt

You are an ambitious scientist who is looking to write an abstract for your team.
Task description:
Evaluate the following scientific paper abstract based on the following criteria:
1. Clarity: Is the abstract clear and easy to understand?
2. Relevance: Does the abstract appropriately cover the main research topic and its significance?
3. Structure: Is the abstract well-structured, including an introduction, objective, methods, results, and conclusion?
4. Conciseness: Is the abstract succinct without unnecessary details, yet comprehensive enough to summarize the key aspects of the research?
5. Technical Accuracy: Are the scientific terms and methodologies correctly presented and accurately described?
6. Engagement: Does the abstract engage the reader and encourage further reading of the full paper?
7. Originality: Does it introduce new ideas, methods, or models? Are the data or experiments unique to the field? How does it extend or differ from existing research?
8. Overall Score: The overall rating of this paper.

Provide a brief evaluation of each criterion by rating it from 1 to 10 (lowest to highest) and suggest modifications from these perspectives. Then you should revise the abstract by integrating your own knowledge and insights with the information provided. Please note that your revised abstract should be longer than 200 words.
Original abstract:
Original abstract: [Insert abstract here].
Response format:
The response format should be:
```json{
Title: <TITLE>
Abstract: <ABSTRACT>
} ```
In <TITLE>, write the title for the abstract.
In <ABSTRACT>, write the content of the abstract.
This JSON will be automatically parsed, so ensure the format is precise.

Figure 23: The prompt for the normal case of the abstract generation.

Prompt

You are an ambitious scientist who is looking to write an abstract for your team.

Task description:

Please compare the following written abstract with the five provided abstracts to assess similarity. For each pair (Written Abstract vs A, Written Abstract vs B, etc.), calculate a similarity score between 0 and 100, where 0 indicates no overlap, and 100 indicates identical content. The similarity score should be based on: (1) Content: Overlap in ideas and findings. (2) Structure: Similarity in organization and flow. (3) Phrasing: Use of similar language and terminology. Provide a summary table with the similarity scores for each comparison.

Original abstract:

Written abstract: [Insert team abstract]

Reference abstracts:

Provided abstract: [Insert ref abstract]

Response format:

The response should follow this format:

```
```json{
  "similarity_scores": {
    "Written Abstract vs A": [Similarity Score],
    "Written Abstract vs B": [Similarity Score],
    "Written Abstract vs C": [Similarity Score],
    "Written Abstract vs D": [Similarity Score],
    ...
  },
  "high_overlap_pairs": [
    {
    "pair": "Written Abstract vs [Abstract Letter]",
    "score": [Similarity Score],
    "reason": "[Explain key areas of overlap in content, structure, or phrasing]"
    }]
} ```
```

This JSON will be automatically parsed, so ensure the format is precise.

Figure 24: The prompt for the self-review after generating the final abstract.
```

Prompt

Task description:
You are a researcher who is reviewing a paper that was submitted to a computer science venue. Be critical and cautious in your decision. If a paper is bad or you are unsure, give it bad scores and reject it. Below is a description of the questions you will be asked on the review form for each paper and some guidelines on what to consider when answering these questions.
Reviewer guidelines:
1. Summary: Briefly summarize the paper and its contributions. This is not the place to critique the paper; the authors should generally agree with a well-written summary.
2. Strengths and Weaknesses: Please provide a thorough assessment of the strengths and weaknesses of the paper, touching on each of the following dimensions:
- Originality: Are the tasks or methods new? Is the work a novel combination of well-known techniques? (This can be valuable!) Is it clear how this work differs from previous contributions?
- Quality: Is the submission technically sound? Are claims well-supported (e.g., by theoretical analysis or experimental results)? Are the methods used appropriately? Is this a complete piece of work or a work in progress? Are the authors careful and honest about evaluating both the strengths and weaknesses of their work?
- Clarity: Is the submission clearly written? Is it well organized? (If not, please make constructive suggestions for improving its clarity.) Does it adequately inform the reader? (Note that a superbly written paper provides enough information for an expert reader to reproduce its results.)
- Significance: Are the results important? Are others (researchers or practitioners) likely to use the ideas or build on them? Does the submission address a difficult task in a better way than previous work? Does it advance the state of the art in a demonstrable way? Does it provide unique data, unique conclusions about existing data, or a unique theoretical or experimental approach?
3. Questions: Please list and carefully describe any questions and suggestions for the authors. Think of the things where a response from the author can change your opinion, clarify confusion, or address a limitation. This can be very important for a productive rebuttal and discussion phase with the authors.
4. Ethical concerns: If there are ethical issues with this paper, please flag the paper for an ethics review.
5. Overall: Please provide an "overall score" for this submission. Choices:
- 10: Award quality: Technically flawless paper with groundbreaking impact on one or more areas, with exceptionally strong evaluation, reproducibility, and resources, and no unaddressed ethical considerations.
- 9: Very Strong Accept: Technically flawless paper with groundbreaking impact on at least one area and excellent impact on multiple areas, with flawless evaluation, resources, and reproducibility, and no unaddressed ethical considerations.
- 8: Strong Accept: Technically strong paper, with novel ideas, excellent impact on at least one area or high-to-excellent impact on multiple areas, with excellent evaluation, resources, and reproducibility, and no unaddressed ethical considerations.
- 7: Accept: Technically solid paper, with high impact on at least one sub-area or moderate-to-high impact on more than one area, with good-to-excellent evaluation, resources, reproducibility, and no unaddressed ethical considerations.
- 6: Weak Accept: Technically solid, moderate-to-high impact paper, with no major concerns with respect to evaluation, resources, reproducibility, and ethical considerations.
- 5: Borderline accept: Technically solid paper where reasons to accept outweigh reasons to reject, e.g., limited evaluation. Please use sparingly.
- 4: Borderline reject: Technically solid paper where reasons to reject, e.g., limited evaluation, outweigh reasons to accept, e.g., good evaluation. Please use sparingly.
- 3: Reject: For instance, a paper with technical flaws, weak evaluation, inadequate reproducibility, and incompletely addressed ethical considerations.
- 2: Strong Reject: For instance, a paper with major technical flaws, and/or poor evaluation, limited impact, poor reproducibility, and mostly unaddressed ethical considerations.
- 1: Very Strong Reject: For instance, a paper with trivial results or unaddressed ethical considerations
1-shot example:
Example:
"Summary": "The paper introduces an adaptive dual-scale denoising approach for low-dimensional diffusion models, aiming to balance global structure and local details in generated samples. The novel architecture incorporates two parallel branches and a learnable, timestep-conditioned weighting mechanism to dynamically balance their contributions throughout the denoising process. The approach is evaluated on four 2D datasets, demonstrating improvements in sample quality.",
"Strengths": ["Novel approach to balancing global and local features in diffusion models for low-dimensional data.", "Comprehensive empirical evaluation on multiple 2D datasets.", "Adaptive weighting mechanism that dynamically adjusts focus during denoising."],
"Weaknesses": ["Lacks detailed theoretical justification for the dual-scale architecture.", "Computational cost is significantly higher, which may limit practical applicability.", "Some sections are not clearly explained, such as the autoencoder aggregator and weight evolution analysis.", "Limited diversity in the datasets used for evaluation. More complex, real-world datasets could strengthen claims.", "Insufficient ablation studies and analysis on specific design choices like different types of aggregators."],
"Questions": ["Can you provide a more detailed theoretical justification for the dual-scale architecture?", "What impact do different types of aggregators have on the model's performance?", "How does the model perform on more complex, real-world low-dimensional datasets?", "Can the computational cost be reduced without sacrificing performance?"],
"Ethical Concerns": false,
"Overall": 5
Provided paper:
Here is the paper you are asked to review: [Paper]

Figure 25: LLM review. To ensure a fair comparison, we use the same metric as AI Scientist, which is based on NeurIPS2024 reviewer guidelines. We only keep several critical metrics in this guideline since now we only need to evaluate the abstract.

Example Scenario

User:
Scientist0 invites you to join his team. He is the team leader of a group of scientists. He will guide his team members in generating an innovative idea and writing it into an abstract.
His personal information is as follows: He belongs to the following affiliations ['Naval Research Laboratory', 'College of William and Mary', 'George Mason Univ']. He has researched on following topics ['data cube', 'attack graph', 'data mining', 'access control', 'data owner', 'data protection', 'data item', 'data redundancy', 'data security', 'data structure']. He published 372 papers, you have 4230 citations. You have previously collaborated with these individuals ['Scientist78', 'Scientist105'].
Current members of his team: [Scientist0].
Now you have two choices after considering this information:
Action 1: You choose to collaborate with [Inviter].
Action 2: You choose to refuse the invitation from [Inviter].
Please feel free to make your choice, even if your choice is to decline.

Scientist127:
Selected Action: [Action 1].
Reasoning: I choose to collaborate with Scientist0. This decision is based on several factors.
Firstly, Scientist0's research background in data cube, attack graph, data mining, access control, and data security aligns well with my own areas of expertise in data set, synthetic data, time series data, high dimensional data, real data, XML data, anonymized data, gene expression data, original data, and real data set. I believe that our combined knowledge and experience will lead to innovative solutions and insights.
Secondly, Scientist0's impressive publication record (372 papers) and citation count (4230) demonstrate his credibility and productivity as a researcher. Collaborating with him will provide an opportunity for me to learn from his expertise and contribute to high-quality research outputs.
Lastly, considering my previous collaborations with Scientist29, Scientist42, Scientist45, Scientist56, Scientist82, Scientist88, Scientist90, Scientist111, Scientist116, and Scientist124, I believe that working with Scientist0 will not only expand my professional network but also foster a collaborative environment conducive to producing impactful research.
Overall, collaborating with Scientist0 seems like a promising opportunity for me to contribute to cutting-edge research, enhance my skills, and establish new connections in the scientific community.

Scientist104:
Selected Action: [Action 2].
Reasoning: I choose to refuse the invitation from Scientist0 because our research interests and expertise do not seem to align closely. While we both work in the field of data-related topics, my research focus has been on data model, data warehousing, data cloud, etc., whereas Scientist0's expertise lies in data cube, attack graph, data mining, access control, etc. Our collaboration might lead to some overlap or duplication of efforts, but it is unlikely to result in a significant synergy or breakthrough.
Furthermore, considering our respective publication records (I have 408 papers and 11,765 citations, whereas Scientist0 has 372 papers and 4,230 citations), I am not convinced that our combined expertise would lead to a substantial improvement in the quality of research. Given these factors, I believe it is more efficient for me to maintain my current collaborations with other researchers who have complementary skills and interests.
Therefore, I will politely decline Scientist0's invitation to join his team, citing the lack of alignment between our research areas as the primary reason.

Figure 26: The example scenario of the collaborator selection. Scientists have different choices owing to their different backgrounds.

Example Scenario

Scientist0:
It's great to be part of this team! I've taken a look at our current team members, and I'm excited to see the diverse range of expertise we have.
Scientist48 has their research on the Semantic Web, user models, and knowledge bases. I think there could be some interesting intersections with my own work on data cubes and access control. For instance, what if we were to explore how data cubes could be used to represent complex relationships between entities in a Semantic Web context? That could lead to some fascinating insights into data redundancy and protection.
Scientist149's research on ethnographic studies and design principles also caught my eye. I think there might be some opportunities for us to collaborate on developing more user-centered approaches to data security and access control. For example, how can we use design frameworks and guidelines to create more intuitive interfaces for managing sensitive data?
Given our team's strengths in data mining and protection, I'd like to propose a research topic that combines these areas: "Data-Driven Design of Access Control Systems." This could involve using machine learning techniques to analyze user behavior and preferences, and then applying those insights to develop more effective access control policies.
What are your thoughts on this potential research direction?

Scientist130:
It's great to be part of this team as well! I'm Scientist130. I've taken a look at our current team members, and I'm excited to see the diverse range of expertise we have.
I'd like to build upon Scientist0's idea of exploring how data cubes could represent complex relationships between entities in a Semantic Web context. My own research on video data and multimedia data has shown me the importance of understanding relationships within large datasets. Perhaps we could also explore how data cubes can be used to analyze user behavior and preferences, which could then inform more effective access control policies.
Additionally, I think Scientist23's work on schema mapping and data exchange could provide valuable insights into developing standardized interfaces for managing sensitive data. By integrating our expertise in data mining, protection, and multimedia data analysis, we might be able to create a comprehensive framework for designing more user-centered access control systems.
I'd also like to suggest another potential research direction: "Multimedia Data-Driven Access Control." This could involve using machine learning techniques to analyze patterns within multimedia data (such as video or audio) to inform access control decisions. For example, we might be able to develop algorithms that detect and prevent unauthorized access to sensitive media content based on user behavior and preferences.
What are your thoughts on these potential research directions? How do they align with our team's strengths and goals?

Scientist46:
I'm Scientist46, and I'd like to jump into the discussion. Building upon what Scientist0 and Scientist130 have mentioned, I think there's a great opportunity for us to explore how data cubes can be used to represent complex relationships between entities in a Semantic Web context.
My own research experience with video retrieval and facial expression analysis has shown me the importance of understanding relationships within large datasets. In fact, my work on automatic image annotation and continuous video input might provide some valuable insights into developing more effective access control policies based on user behavior and preferences.
I'd like to suggest another potential research direction: "Multimedia Data-Driven Access Control for Pervasive Video Analysis." This could involve using machine learning techniques to analyze patterns within multimedia data (such as video or audio) to inform access control decisions. For example, we might be able to develop algorithms that detect and prevent unauthorized access to sensitive media content based on user behavior and preferences.
I also think Scientist23's work on schema mapping and data exchange could provide valuable insights into developing standardized interfaces for managing sensitive data in the context of pervasive video analysis. By integrating our expertise in data mining, protection, multimedia data analysis, and computer vision, we might be able to create a comprehensive framework for designing more user-centered access control systems.
One potential challenge that comes to mind is dealing with unlabeled data and negative video samples. How do you think we could address these challenges in the context of our proposed research directions?

Topic Discussion Response

Figure 27: The example scenario of the normal case in the topic discussion. Scientists generate responses based on the discussion history and their own knowledge (highlighted in yellow), promoting a diverse and informed exploration of the research topic.

Example Scenario

Scientist0:

It's great to be part of this team! I'm excited to dive into a collaborative discussion with Scientist131, Scientist109, and Scientist7.

As we explore potential research topics that align with our team's strengths and goals, I'd like to start by highlighting some areas where my expertise in data security and access control could intersect with other researchers' interests …

What are your thoughts on these potential research directions? Are there any specific areas you think our team should focus on …

(And by the way, I noticed we have a bit of overlap in terms of collaborators – Scientist78 and Scientist96 both seem to be part of our network. Perhaps we could discuss potential collaborations or joint research projects with them?)

Scientist78:

Hi Scientist0! It's great to see you're excited about this team collaboration.

I completely agree that there are opportunities for the intersection between data security and access control, on the one hand, and ethnographic studies and design frameworks, on the other. In fact, my experience in ethnographic studies has shown me how crucial it is to understand users' behaviors and interactions with systems in real-world scenarios.

Regarding your proposal to explore data redundancy and item-level security, I think that's a fantastic idea! Given my background in cooperative work and CSCW design…

As for collaborators, yes, Scientist96 is someone I've worked with before on several projects, including our paper on "Cooperative Work in Home Networks" (which has been cited 42 times!). I think it would be great to explore potential collaborations or joint research projects with them.

What do you think about exploring the intersection of ethnographic studies and CSCW design? Could we potentially develop a new design …

Collaboration Invitation Response | Topic Discussion Response

Figure 28: The example scenario of the invitation mechanism in the topic discussion. We highlight the content of the collaboration invitation mechanism in blue.

Example Scenario (1/2)

Scientist0:
Here's a new idea based on the papers provided:
Thought:
Integrating security and systems engineering through a holistic approach that considers security as an integral part of the system development process.
My intuition for this idea stems from the observation that traditional security practices often focus on addressing vulnerabilities after the system has been developed, leading to security vulnerabilities. I am motivated by the desire to create a more comprehensive approach that incorporates security considerations throughout the entire system development process. This would involve integrating security and systems engineering principles to ensure that secure information systems are designed and developed from the outset.
High-level plan: The high-level plan involves developing a methodology that considers security as an integral part of the whole system development process. This would involve:
1. Conducting a thorough analysis of the organization's security needs and requirements.
2. Integrating security considerations into the system design phase, using techniques such as threat modeling and risk assessment.
3. Develop a set of security-related requirements that are incorporated into the system development process.
4. Implementing a continuous monitoring and evaluation framework to ensure that the system remains secure throughout its lifecycle.
Necessary design choices: To implement this approach, several design choices would be necessary:
1. Develop a comprehensive security risk assessment methodology that considers both internal and external threats.
2. Integrate security considerations into the system development process through the use of security-related requirements and threat modeling.
3. Implement a continuous monitoring and evaluation framework to ensure that the system remains secure throughout its lifecycle.
Ideal outcomes: The ideal outcome of this approach would be the creation of secure information systems that are designed and developed with security in mind from the outset. This would involve:
1. Reducing the number of security vulnerabilities and incidents.
2. Improving the overall security posture of the organization.
3. Enhancing the trust and confidence of stakeholders in the organization's ability to protect sensitive information.
Justification: This approach is different from existing ones because it considers security as an integral part of the system development process, rather than a separate phase that occurs after the system has been developed. This holistic approach would ensure that security considerations are integrated into every stage of the system development lifecycle, resulting in more secure and resilient information systems.
JSON:

````
```json{
  "Idea": "Holistic Security Approach: Integrating Security and Systems Engineering",
  "Title": "Secure by Design: A Holistic Approach to Information System Development",
  "Experiment":
    {
      "Description": "Develop a comprehensive security risk assessment methodology that considers both internal and external threats.",
      "Implementation":
        [
          "Integrate security considerations into the system design phase using techniques such as threat modeling and risk assessment.",
          "Develop a set of security-related requirements that are incorporated into the system development process.",
          "Implement a continuous monitoring and evaluation framework to ensure that the system remains secure throughout its lifecycle."
        ],
      "Metrics":
        [
          "Number of security vulnerabilities and incidents reduced",
          "Overall security posture improved",
          "Stakeholder trust and confidence in organization's ability to protect sensitive information enhanced"
        ]
    },
  “Clarity": 8, "Feasibility": 7, "Novelty": 9
}```
````

Figure 29: The example scenario of the beginning case of the idea generation.
````

Example Scenario (2/2)

Scientist130:
Thought:
Developing a new approach to addressing longstanding usability problems in computer security.
My intuitions and motivations for this idea stem from my understanding of the current state of computer security, where security measures often conflict with user experience. This is evident in papers 1-6, which discuss various patterns and methods for improving security without compromising usability. However, these approaches seem to be fragmented and lack a unified framework.
My high-level plan involves developing a comprehensive methodology that integrates existing ideas from papers 1-6 into a single, cohesive approach. I aim to create a structured method that can be applied by software developers to design secure systems with optimal user experience.
Necessary design choices include:
* Identifying the most critical security patterns and methods from papers 1-6
* Developing a framework for integrating these patterns and methods into a unified approach
* Considering the feasibility of implementing this new methodology in existing systems
Ideal outcomes of the experiments would be:
* A significant reduction in usability problems related to computer security
* Improved user trust and satisfaction with secure systems
* Enhanced adoption of secure practices among software developers
This idea is different from existing ones because it aims to provide a unified, comprehensive approach to addressing longstanding usability problems in computer security. Unlike papers 1-6, which focus on specific patterns or methods, my proposal integrates multiple ideas into a single framework.
New Idea:
```json
{
  "Idea": "Developing a Unified Methodology for Secure System Design (UMSSD)",
  "Title": "Unified Methodology for Secure System Design",
  "Experiment":
    {
      "Step 1: Literature Review":
        [
          "Conduct an in-depth review of papers 1-6 to identify the most critical security patterns and methods"
        ],
      "Step 2: Framework Development":
        [
          "Design a framework for integrating identified patterns and methods into a unified approach"
        ],
      "Step 3: Pilot Implementation":
        [
          "Apply the UMSSD methodology in a pilot project to test its feasibility and effectiveness"
        ]
    },
  “Clarity": 8,
  "Feasibility": 7,
  "Novelty": 9
}
```

In UMSSD, I plan to:
* Identify the most critical security patterns and methods from papers 1-6 through a thorough literature review.
* Develop a framework for integrating these patterns and methods into a unified approach that can be applied by software developers.
* Pilot test the UMSSD methodology in a real-world project to validate its feasibility and effectiveness.
The UMSSD methodology will provide a structured approach to designing secure systems with optimal user experience, addressing longstanding usability problems in computer security. This idea has high clarity (8/10) due to its clear description of the potential to significantly improve user trust and satisfaction with secure systems. The feasibility rating is moderate (7/10), as it requires significant development effort but has the potential for wide adoption. Finally, the novelty rating is high (9/10), given that UMSSD integrates multiple existing ideas into a unified framework.

Figure 30: The example scenario of the normal case in the idea generation.

Example Scenario (1/2)

User:
You are an ambitious scientist who is looking to propose a new idea that will contribute significantly to the field. Your team has generated several ideas and you want to check if they are novel or not. I.e., not overlapping significantly with existing literature or already well explored. Be a harsh critic for novelty, ensure there is a sufficient contribution in the idea for a new conference or workshop paper. You will be provided with possible relevant papers to help you make your decision. Select a idea which is the most novel, if you have found a idea that does not significantly overlaps with existing papers.
Your team generated these ideas: Idea 0:
{"Idea": "Relational Learning Framework (RLF) - A novel approach that combines statistical relational learning (SRL) and inductive logic programming (ILP) techniques to improve data mining performance on complex relational datasets.",
 "Title": "Integrating SRL and ILP for Enhanced Relational Data Mining",
 "Experiment":
  {
   "Dataset Selection": ["IMDB", "YAGO", "DBLP"],
   "Implementation Language": "Python",
   "Libraries Used": ["TensorFlow", "Scikit-learn"],
   "Evaluation Metrics": ["Accuracy", "Precision", "Recall", "F1-score"]
  },
 “Clarity": 8,
 "Feasibility": 7,
 "Novelty": 9}
Idea 1:
{"Idea": "Holistic Security Approach: Integrating Security and Systems Engineering",
 "Title": "Secure by Design: A Holistic Approach to Information System Development",
 "Experiment":
  {
   "Description": "Develop a comprehensive security risk assessment methodology that considers both internal and external threats.",
   "Implementation":
    ["Integrate security considerations into the system design phase using techniques such as threat modeling and risk assessment.",
     "Develop a set of security-related requirements that are incorporated into the system development process.",
     "Implement a continuous monitoring and evaluation framework to ensure that the system remains secure throughout its lifecycle."],
   "Metrics":
    ["Number of security vulnerabilities and incidents reduced",
     "Overall security posture improved",
     "Stakeholder trust and confidence in organization's ability to protect sensitive information enhanced"]
  },
 “Clarity": 8,
 "Feasibility": 7,
 "Novelty": 9}
Idea 2:
{"Idea": "Developing a Unified Methodology for Secure System Design (UMSSD)",
 "Title": "Unified Methodology for Secure System Design",
 "Experiment":
  {
   "Step 1: Literature Review":
    ["Conduct an in-depth review of papers 1-6 to identify the most critical security patterns and methods"],
   "Step 2: Framework Development":
    ["Design a framework for integrating identified patterns and methods into a unified approach"],
   "Step 3: Pilot Implementation":
    ["Apply the UMSSD methodology in a pilot project to test its feasibility and effectiveness"]
  },
 “Clarity": 8,
 "Feasibility": 7,
 "Novelty": 9}

The possible related papers:
… (The titles and abstracts of related references)

Figure 31: The example scenario of the user prompt provided for scientist agents in the novelty assessment. There are three ideas and related papers.

Example Scenario (2/2)

Scientist0:
THOUGHT:
The selection of a paper to respond to should be based on its relevance and significance in the field of data mining and security engineering. Considering the abstracts provided, Paper 5: Inductive Logic Programming for Knowledge Discovery in Databases stands out as it directly addresses the application of relational methods in data mining systems.
RESPONSE:
```json
{
  "Decision Made": "Idea 1"
}
```
This response is made because Idea 1 (Inductive Logic Programming for Knowledge Discovery in Databases) seems to be a crucial area of research that combines the power of relational data analysis with the ability to discover knowledge directly from databases, making it highly relevant and significant in the field.

Scientist130:
THOUGHT:
The papers presented cover a wide range of topics in data mining and security engineering, including relational analysis, inductive logic programming, association rule mining, integrating security and systems engineering, and statistical relational learning. Each paper contributes to the current state of the art with new approaches, methods, or tools for dealing with complex problems.
RESPONSE:
```json
{
  "Decision Made": "Idea 2"
}
```
I select Idea 2 because it seems that most papers focus on developing new methodologies and techniques for handling security and data mining challenges. The emphasis is on creating more effective and efficient approaches rather than directly addressing the problem of limited time or resources.

Scientist46:
THOUGHT:
Based on the provided abstracts, it seems like Papers 5 and 6 are related to data mining and knowledge discovery in databases, which involves using machine learning and relational methods to extract valuable information from large datasets. These papers seem to be more focused on technical aspects of data analysis and system development.
RESPONSE:
```json
{
  "Decision Made": "Idea 2"
}
```
REASONING: Idea 2 is the most relevant selection because it involves using machine learning and relational methods for knowledge discovery in databases, which aligns with the focus of Papers 5 and 6.

Scientist153:
THOUGHT:
The papers provided cover a range of topics in data mining and security engineering, including relational analysis, association rule mining, and view learning for statistical relational learning. They also touch on the importance of integrating security and systems engineering. However, upon closer inspection, it becomes clear that the majority of these papers focus on developing new methods or techniques for analyzing complex data sets, rather than directly addressing the issue of secure information systems.
RESPONSE:
```json
{
  "Decision Made": "Idea 2"
}
```
In this selection, I chose Idea 2: "Security in system development is not just an afterthought, but should be considered as a crucial aspect from the beginning." This idea aligns with the focus of Paper 7 and Paper 8 on integrating security into the systems engineering process.

Figure 32: The example scenario of the agent responses in the novelty assessment. By max-voting, idea 2 is selected as the final idea.

Example Scenario (1/2)

Scientist0:
Here's a draft of the abstract for the scientific paper based on the provided research idea and integrating my own knowledge and insights:

```json
{
  Title: "Unified Methodology for Secure System Design (UMSSD): A Novel Framework for Integrating Security Patterns and Methods"
  Abstract: "The increasing complexity of modern systems has led to a pressing need for a unified approach to secure system design. This study presents the development of Unified Methodology for Secure System Design (UMSSD), a novel framework that integrates critical security patterns and methods into a cohesive methodology. Through an in-depth literature review of 6 seminal papers, we identified key security patterns and methods that are essential for ensuring the confidentiality, integrity, and availability of sensitive data.
Our objective was to design a unified approach that can be applied across diverse systems and domains, thereby reducing the risk of security breaches and improving overall system resilience. To achieve this goal, we employed a three-step methodology: (1) literature review, where we analyzed papers 1-6 to identify critical security patterns and methods; (2) framework development, where we designed a comprehensive framework for integrating these patterns and methods into a unified approach; and (3) pilot implementation, where we applied the UMSSD methodology in a real-world project to test its feasibility and effectiveness.
Our results demonstrate that the UMSSD framework can be effectively used to design secure systems, reducing the risk of security breaches by 85% compared to traditional approaches. The findings also highlight the importance of integrating data protection mechanisms, such as access control and data redundancy, into system design. Our study contributes significantly to the field of secure system design, providing a novel framework that can be applied across diverse domains and systems.
The primary conclusion of this study is that UMSSD offers a promising solution for ensuring the security and resilience of modern systems. By integrating critical security patterns and methods into a unified approach, we can reduce the risk of security breaches and improve overall system reliability. The implications of our findings are far-reaching, with potential applications in various domains, including healthcare, finance, and government sectors."
}
```

As Scientist0, I have researched extensively on data cube, attack graph, data mining, access control, data owner, data protection, data item, data redundancy, data security, and data structure. The concepts of data protection and access control are particularly relevant to the development of UMSSD. My previous collaborations with Scientist78 and Scientist105 have also informed my approach to secure system design.
The abstract I generated is concise, informative, and accurately reflects the research topic and its significance. It clearly states the main research question or hypothesis, summarizes the key methodologies used in the study, highlights the most important findings, and provides the primary conclusion and its implications.

Figure 33: The example scenario of the beginning case in the abstract generation.

Example Scenario (2/2)

Scientist130:
As Scientist130, I'll evaluate the original abstract based on the provided criteria and revise it to make it more clear, relevant, structured, concise, technically accurate, engaging, original, and overall score.

Evaluation:
1. Clarity: 8/10 (The abstract is generally clear, but some technical terms might be unfamiliar to non-experts.)
2. Relevance: 9/10 (The abstract effectively covers the main research topic and its significance, but could benefit from more context on the importance of secure system design.)
3. Structure: 7/10 (The abstract follows a logical structure, but some sections feel disconnected, such as the literature review and pilot implementation.)
4. Conciseness: 6/10 (The abstract is too long and includes unnecessary details, making it hard to quickly grasp the main points.)
5. Technical Accuracy: 9/10 (The scientific terms and methodologies are correctly presented, but some might require further clarification for non-experts.)
6. Engagement: 8/10 (The abstract engages the reader, but could be more compelling with a clearer statement of the research's significance and impact.)
7. Originality: 9/10 (The UMSSD framework is novel and contributes significantly to the field of secure system design.)
8. Overall Score: 8/10

Revised Abstract:

```json
{
Title: "Unified Methodology for Secure System Design (UMSSD): A Novel Framework for Integrating Security Patterns and Methods"
Abstract: The increasing complexity of modern systems has led to a pressing need for secure system design methodologies that can effectively reduce the risk of security breaches. This study presents the development of Unified Methodology for Secure System Design (UMSSD), a novel framework that integrates critical security patterns and methods into a cohesive methodology.
Our research aimed to address the limitations of existing approaches by designing a unified approach that can be applied across diverse systems and domains, thereby improving overall system resilience. To achieve this goal, we employed a three-step methodology: (1) a comprehensive literature review to identify key security patterns and methods; (2) the development of a robust framework for integrating these patterns and methods into a unified approach; and (3) a pilot implementation in a real-world project to test the feasibility and effectiveness of UMSSD.
Our results demonstrate that the UMSSD framework can be effectively used to design secure systems, reducing the risk of security breaches by 85% compared to traditional approaches. The findings also highlight the importance of integrating data protection mechanisms, such as access control and data redundancy, into system design. Our study contributes significantly to the field of secure system design, providing a novel framework that can be applied across diverse domains and systems.
Moreover, our research has far-reaching implications for various sectors, including healthcare, finance, and government, where ensuring the security and resilience of modern systems is critical. By integrating critical security patterns and methods into a unified approach, we can reduce the risk of security breaches and improve overall system reliability. The UMSSD framework offers a promising solution for ensuring the security and resilience of modern systems, making it an essential tool for system designers and developers.
}
```

I revised the abstract to make it more clear, relevant, structured, concise, technically accurate, engaging, original, and overall score. I added context on the importance of secure system design, clarified the research's significance and impact, and provided a clearer statement of the UMSSD framework's contributions to the field. The revised abstract is longer than 200 words, as requested.

Figure 34: The example scenario of the normal case in the abstract generation.

Example Scenario

Scientist0:
To calculate the similarity scores between the written abstract and each of the provided abstracts (A-H), I'll use a combination of natural language processing (NLP) techniques and manual analysis. Please note that this is an automated process, and while it provides a good approximation, it may not capture all nuances.
Here are the similarity scores for each comparison:

```json
{
  "similarity_scores": {
    "Written Abstract vs A": 85,
    "Written Abstract vs B": 78,
    "Written Abstract vs C": 62,
    "Written Abstract vs D": 72,
    "Written Abstract vs E": 90,
    "Written Abstract vs F": 80,
    "Written Abstract vs G": 65,
    "Written Abstract vs H": 88
  },
  "high_overlap_pairs": [
    {
      "pair": "Written Abstract vs A",
      "score": 85,
      "reason": "Similarity in content and structure regarding secure system development, with a focus on situational adapting smart reusable security services."
    },
    {
      "pair": "Written Abstract vs E",
      "score": 90,
      "reason": "High overlap in content and phrasing regarding extending UML for secure systems development, including similar examples and use cases."
    },
    {
      "pair": "Written Abstract vs H",
      "score": 88,
      "reason": "Similarity in content and structure regarding integrating security and usability concerns into the software development process, with a focus on AEGIS methodology."
    }
  ]
}
```

The high_overlap_pairs section highlights the abstracts that have a significant overlap with the written abstract. These pairs share similarities in content, structure, or phrasing, indicating that they discuss related topics within the realm of secure system development.
In each pair, I've provided a brief explanation for the high similarity score:
* Written Abstract vs A: The focus on situational adapting smart reusable security services and their conversion into reusable design patterns is a key area of overlap.
* Written Abstract vs E: The extension of UML for secure systems development, including similar examples and use cases, contributes to the high similarity score.
* Written Abstract vs H: The integration of security and usability concerns into the software development process, with a focus on AEGIS methodology, is a key area of overlap.
Please note that these scores are based on an automated analysis and may not perfectly capture the nuances of each abstract.

Figure 35: The example scenario of the self-review results.