Deep-Ideation: Generating Novel Research Ideas with Evolving LLM Agents on Scientific Concept Network

Anonymous ACL submission

Abstract

Novel research ideas play a critical role in advancing scientific inquiries. Recent advancements in Large Language Models (LLMs) have demonstrated their potential to generate novel research ideas by leveraging large-scale scientific literature. However, previous works face substantial challenges since they heavily rely on the textual content of scientific literature, but overlook the rich semantics and high-order connections embedded in scientific concept networks. To address these challenges, we propose **Deep-Ideation**, a powerful LLM agent that generates high-quality research ideas via iteratively searching for novel yet feasible combinations on scientific concept network. Our framework introduces an **Explore-Expand-Evolve** workflow to facilitate continuous dynamic interaction with the scientific concept network. This mechanism drives a structural evolution from concept accumulation to deep innovation, utilizing an Idea Stack to track the ideation trajectory and align it with overall research trends. Furthermore, we fine-tune a Critic Model on real-world reviewer feedback to align with expert standards, rigorously steering ideation toward greater novelty and feasibility. Extensive experiments demonstrate that Deep-Ideation achieves publication-quality performance, with about **81.5%** of generated ideas surpassing the acceptance scores of top AI conferences. Furthermore, expert human evaluations corroborate these findings, confirming that Deep-Ideation produces professional-grade research proposals. Our code is open-sourced for reproducibility at `https://anonymous.4open.science/r/Deep_Ideation-E385`.

1 Introduction

The emerging agentic power of Large Language Models (LLMs) (Li et al., 2025b; Wu et al., 2025b; Zhao et al., 2025) has inspired wide range of researchers to design LLM agents to automate scientific discovery (Wang et al., 2023; Lu et al., 2024; Peng et al., 2025), which is often known as AI scientist systems (Gottweis et al., 2025; Yu et al., 2024; Qi et al., 2024). Ideation, the ability to generate novel yet feasible research ideas, is arguably one of the most important capabilities, as it shapes the direction of scientific inquiry and influences the course of human progress (Coccia, 2019; Langley, 1987). Consequently, leveraging these agentic capabilities to automate the ideation process has emerged as a central research focus, promising to accelerate the pace of scientific innovation (Si et al., 2024; Li et al., 2024; Pu et al., 2025; Si et al., 2025a).

However, most existing frameworks rely exclusively on the textual content of scientific literature (Yang et al., 2024; Su et al., 2024; Qi et al., 2023). This approach overlooks the structural richness of the scientific concept network—a framework where human scientists have long constructed meaningful relationships between scientific concepts through literature, forming a comprehensive conceptual landscape. Although some studies have attempted to incorporate this network data (Wang et al., 2024b,a; Baek et al., 2024), they typically restrict usage to a one-off context retrieval at the initial stage. Consequently, they fail to dynamically re-query the network for fresh insights during the subsequent iterative optimization of ideas. Consequently, these approaches capture only immediate concept associations and fail to continuously and dynamically interact with the network to acquire new structural knowledge as the ideation process evolves, thereby missing the complex, high-order relationships essential for deep innovation.

However, effectively implementing this vision is non-trivial. First, mining meaningful high-order neighbors is challenging, as naively expanding the search depth often introduces excessive noise that obscures truly innovative connections. Second, achieving continuous dynamic interaction is complex, requiring the agent to adaptively align

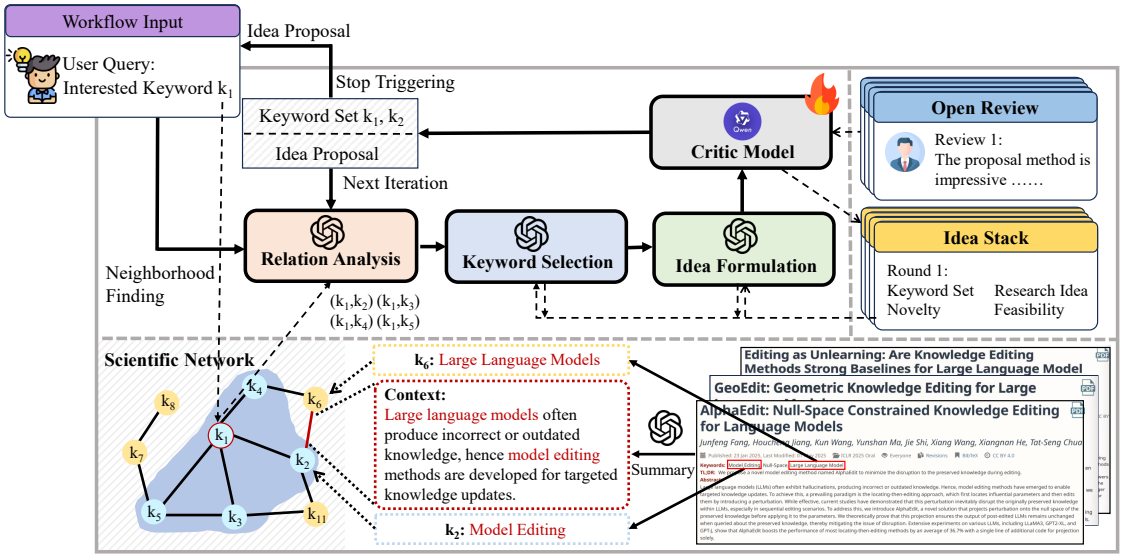

Figure 1: Illustration of the construction of the scientific concept network and the Deep-Ideation process.

its network exploration with the evolving research context to maintain logical coherence.

To systematically tackle these hurdles, we propose **Deep-Ideation**, an agentic framework grounded in a massive scientific concept network where nodes are instantiated as research keywords extracted from over 100,000 papers. The construction of the scientific network and the Deep-Ideation process are shown in Figure 1. Our framework introduces a dynamic Explore-Expand-Evolve workflow to facilitate continuous dynamic interaction with the scientific concept network. In this evolutionary trajectory, the Explore and Expand phases fuel the process by actively identifying and integrating new concepts from high-order connections, thereby building the necessary semantic breadth for innovation. Crucially, this accumulation lays the groundwork for the Evolve phase, which drives structural evolution by strategically replacing concepts to escape local optima. This mechanism moves beyond simple keyword matching, transforming the ideation process into an adaptive optimization cycle. To support this complex evolution, an Idea Stack records the entire trajectory to maintain logical coherence, while a Critic Model—fine-tuned on peer reviews—steers the generated ideas toward expert-level novelty and feasibility.

Experimental results across ten major AI conferences demonstrate that Deep-Ideation achieves a significant breakthrough, outperforming state-of-the-art baselines by an average margin of **6.70%**. Notably, about **81.5%** of these generated ideas surpass the average acceptance scores of top-tier conferences, marking a capability leap from mere generation to high-quality innovation. This quantitative success is further corroborated by expert human evaluation, which confirms that our system offers genuine scientific inspiration. Furthermore, comprehensive ablation studies validate the effectiveness of the refinement components designed within our framework.

The key contributions of this work are as follows:

- We construct a massive scientific concept network from over 100,000 papers and a comprehensive real-world review dataset. These datasets will be made publicly available for the research community to foster further collaboration and exploration.
- We propose Deep-Ideation, an agentic framework capable of continuously interacting with the scientific network. This mechanism allows the agent to dynamically capture meaningful high-order information from the graph.
- We demonstrate through extensive experiments that our method achieves superior performance, producing research ideas that surpass the average acceptance scores of top-tier AI conferences.

2 Problem Formulation

2.1 Definition of Scientific Concept Network

Scientific knowledge is traditionally stored in the textual content of massive literature collections. While rich in detail, this unstructured format presents information linearly, often obscuring the

intricate dependencies and logical structures underlying scientific progress. In contrast, organizing this knowledge as a graph offers distinct advantages (Badalyan et al., 2024): it transforms isolated documents into a structured map, explicitly representing complex relationships and facilitating efficient navigation through the logic of scientific inquiry (Wang et al., 2023; Sourati and Evans, 2023).

Formally, we define the Scientific Concept Network as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. The node set $\mathcal{V} = \{v_1, v_2, ..., v_n\}$ represents the universe of scientific concepts, which are instantiated in our framework as concrete keywords extracted from the literature. The edge set $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ represents the connections between these keywords. Specifically, an edge $(v_i, v_j) \in \mathcal{E}$ exists if keywords v_i and v_j co-occur in the literature. Crucially, rather than compressing these connections into a static scalar weight, we define the attribute of each edge as the specific set of source papers $\mathcal{D}_{ij} = \{p \mid v_i, v_j \in p\}$, where p represents a paper containing both keywords. This design ensures that the edge acts as a carrier of raw scientific evidence, allowing the specific relationship between v_i and v_j to be dynamically analyzed and synthesized from the contexts in $\mathcal{D}_{ij}$ during the ideation process.

2.2 Formulating Research Ideation Problem

Novel ideas are often conceptualized as the cohesive synthesis of seemingly unrelated knowledge fragments (Koestler, 1964; Lee and Chung, 2024; Benedek et al., 2012). In the context of scientific research, these knowledge fragments are encapsulated as nodes within the Scientific Concept Network. Consequently, we frame the ideation challenge not merely as a static text generation task, but as a strategic navigation process designed to discover distant yet synergistic concept combinations that address gaps in the current research landscape.

Formally, we define the research ideation problem as an iterative optimization process over the network $\mathcal{G}$. Let $\mathcal{K}_t \subseteq \mathcal{V}$ denote the subset of keywords selected at iteration t. The ideation process evolves through a series of state transitions, where the system continuously refines the concept set to approach an optimal configuration. Specifically, the transition from t to $t+1$ is modeled as a two-stage process:

$$(\mathcal{K}_{t+1}, \mathcal{H}'_t) = \psi(\mathcal{K}_t, \mathcal{G}, \mathcal{H}_t)$$
$$(P_{t+1}, \mathcal{H}_{t+1}) = \phi(\mathcal{K}_{t+1}, \mathcal{H}'_t)$$

Here, $\mathcal{H}_t$ represents the optimization history at the start of the iteration. In the first stage, the function ψ acts as the transition operator, updating the keyword set to $\mathcal{K}_{t+1}$ (e.g., via expansion or substitution) while recording this topological change into an intermediate history state $\mathcal{H}'_t$. In the second stage, ϕ maps the updated concepts into a coherent research proposal P_{t+1} and finalizes the history update to $\mathcal{H}_{t+1}$ by appending the newly generated proposal. The objective is to reach a terminal state T that maximizes a scientific utility function $\mathcal{J}(P_T)$, which quantifies the proposal's novelty and feasibility based on established scientific standards. In practice, $\mathcal{J}(\cdot)$ is realized by the Critic Model introduced in Section 3.3, which scores each candidate proposal along the novelty and feasibility axes by imitating reviewer reasoning at every iteration.

3 Methods

3.1 Overview

To enable continuous and dynamic interaction with the scientific concept network and capture meaningful high-order neighbors for deep innovation, we propose Deep-Ideation. As illustrated in Figure 2, the entire workflow involves an iterative Explore-Expand-Evolve process governed by an Idea Stack to track ideation trajectory and a Critic Model to align with expert standards. To systematically realize this agentic workflow, we first formalize the underlying structural foundation that stores scientific knowledge, followed by the specific functional modules designed to navigate and reason over this topological space.

3.1.1 Scientific Concept Network

The Scientific Concept Network serves as the structural foundation of our framework, instantiated as a graph where nodes represent research keywords and edges encapsulate literature evidence. To establish the node set $\mathcal{V}$, we utilize LLM to extract representative research keywords from the titles, abstracts and introductions of collected academic papers. Subsequently, edges are formed based on co-occurrence; specifically, if two keywords appear within the same paper, they are connected. Crucially, consistent with the definition in Section 2.1, each edge stores the specific set of source papers $\mathcal{D}_{ij}$ as its attribute. This design preserves the raw contextual evidence, allowing the agent to dynamically retrieve and analyze the specific relationships

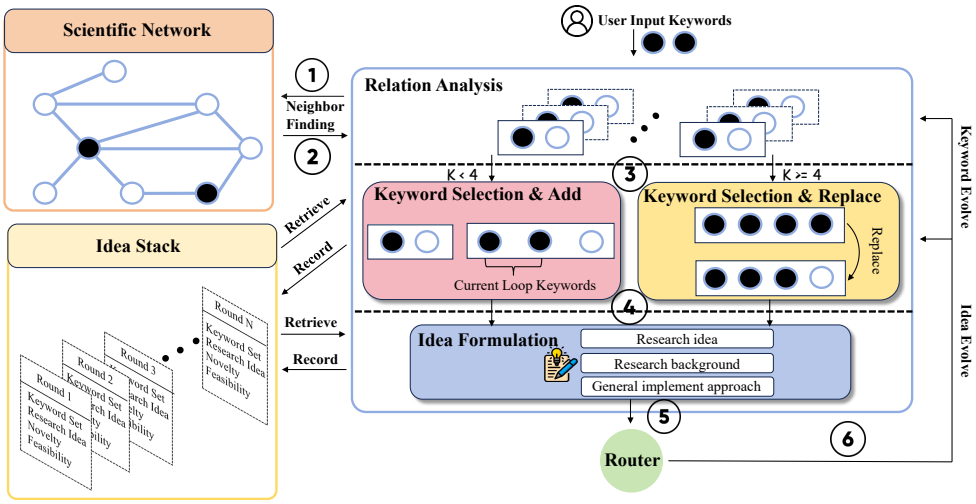

Figure 2: Overview of our Deep Ideation framework. In this figure, we set the maximum size of the keyword set to 4.

between concepts during the ideation process. The details of the keyword extraction pipeline are provided in Appendix A.5.

3.1.2 Core Processing Modules

Driven by the foundational network, Deep-Ideation employs three specialized modules to navigate the semantic space and synthesize innovations:

The **Relation Analysis Module** is designed to dynamically interpret the semantic connections between keywords by leveraging the raw evidence stored on the network edges. When the agent examines an edge connecting two keywords, the module retrieves the corresponding set of source papers $\mathcal{D}_{ij}$. To efficiently capture the essence of the relationship while managing input length, we extract specific high-information sections—titles, abstracts, and introductions—from these documents. This aggregated context is then fed into an LLM, which synthesizes a concise summary of the relationship, explicitly describing how the concepts interact.

The **Keyword Selection Module** acts as a strategic filter for network navigation, utilizing an LLM to evaluate candidate neighbors based on their semantic relations. To determine how to update the current keyword set $\mathcal{K}_t$, this module dynamically delegates execution to one of two specialized subunits based on the set's target length L_{target}:

- Keyword Selection & Expand: Triggered when the set is not full ($|\mathcal{K}_t| < L_{target}$). This unit functions to supplement the current keyword set, selecting the most semantically valuable neighbor to add into $\mathcal{K}_t$.
- Keyword Selection & Evolve: Triggered when the set reaches capacity ($|\mathcal{K}_t| = L_{target}$). This unit functions to optimize the conceptual composition, identifying a superior external candidate to replace a less contributing incumbent in $\mathcal{K}_t$.

The **Idea Formulation Module** translates the selected keywords into a coherent research proposal containing three sections: Research Background (motivation and existing gaps), Research Idea (core innovation and hypothesis), and General Implementation Approach (technical feasibility and execution strategy). Requiring an explicit Implementation Approach in every proposal anchors each idea in a concrete execution path, directly improving the practical feasibility of the generated outputs.

3.2 Network-Guided Idea Evolution

We operationalize this process through an Explore-Expand-Evolve workflow that harmonizes knowledge acquisition with structural optimization. The Explore and Expand phases fuel the evolutionary growth by continuously injecting diverse high-order neighbors into the context. Building on this foundation, the Evolve phase facilitates qualitative refinement, dynamically restructuring the concept set to escape local optima and synthesize a coherent scientific breakthrough. The complete algorithmic procedure is formalized in Appendix A.1.

3.2.1 Explore

The Explore phase initiates the iteration by identifying potential directions for conceptual expansion. Given the current keyword set $\mathcal{K}_t$, the agent first queries the Scientific Concept Network to retrieve the set of immediate neighboring nodes $\mathcal{N}(\mathcal{K}_t)$. For each candidate neighbor $v_j \in \mathcal{N}(\mathcal{K}_t)$ connected to an existing keyword $v_i \in \mathcal{K}_t$, the sys-

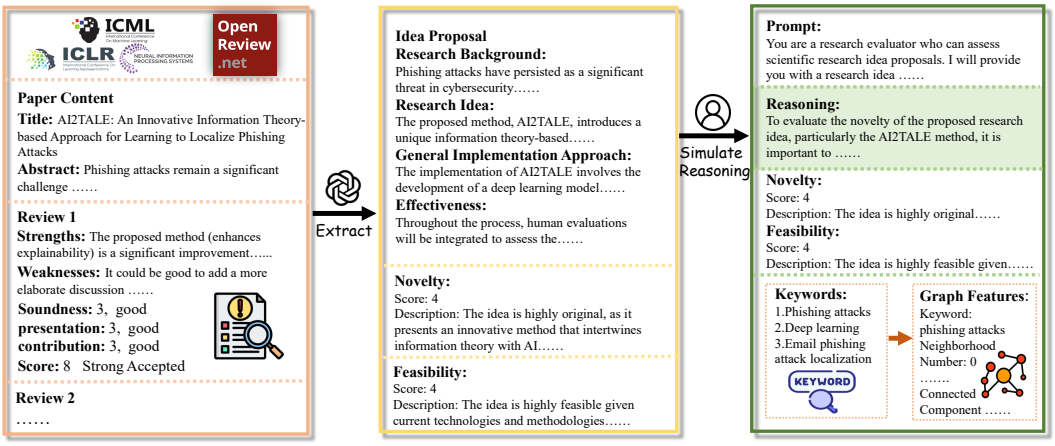

Figure 3: The construction process of the training data for the Critic Model

tem retrieves the raw literature evidence $\mathcal{D}_{ij}$ stored on the connecting edge. Crucially, to transform this raw data into actionable insights, the system invokes the Relation Analysis Module (denoted as Φ_{rel}). This module processes the evidence to generate a concise semantic summary r_{ij} describing the specific interaction between the concepts:

$$r_{ij} = \Phi_{rel}(\mathcal{D}_{ij})$$

Consequently, the output of this phase is a set of semantically annotated candidates $\mathcal{C}_{explore} = \{(v_j, r_{ij}) \mid v_j \in \mathcal{N}(\mathcal{K}_t)\}$. This structured representation ensures that the subsequent selection step is grounded in interpreted scientific logic rather than mere topological connectivity.

3.2.2 Expand

The Expand phase is triggered when the size of the current keyword set $|\mathcal{K}_t|$ has not yet reached the target length L_{target}. Its primary objective is to incrementally enrich the concept set by integrating the most semantically valuable neighbor identified in the Explore phase.

To ensure the agent maintains logical consistency throughout this accumulation process, we introduce the Idea Stack ($\mathcal{H}_t$), a memory structure designed to record the historical evolution of the ideation process. Specifically, it tracks the iterative changes in both keyword selections and proposal drafts from previous rounds. Crucially, this mechanism addresses the challenge of achieving continuous dynamic interaction, as it allows the agent to adaptively align its network exploration with the evolving research context to maintain logical coherence.

Functionally, the process begins with the Keyword Selection & Expand Module (Φ_{exp}), which evaluates the candidates $\mathcal{C}_{explore}$ from the previous phase together with the current history $\mathcal{H}_t$. The module selects the optimal keyword v^* that maximizes semantic synergy:

$$v^* = \Phi_{exp}(\mathcal{C}_{explore}, \mathcal{H}_t)$$

Subsequently, the keyword set is updated to $\mathcal{K}_{t+1} = \mathcal{K}_t \cup \{v^*\}$. Utilizing this enriched context, the Idea Formulation Module (Φ_{gen}) synthesizes a new research proposal P_{t+1}. Finally, both the selected keyword and the generated proposal are pushed onto the stack, updating the history state for the next iteration:

$$P_{t+1} = \Phi_{gen}(\mathcal{K}_{t+1}, \mathcal{H}_t)$$
$$\mathcal{H}_{t+1} = \mathcal{H}_t \oplus (v^*, P_{t+1})$$

3.2.3 Evolve

The Evolve phase acts as a dual-path optimization mechanism, triggered once the keyword set reaches its saturation point ($|\mathcal{K}_t| = L_{target}$). At this stage, the system focuses on iterative refinement rather than accumulation. Since each proposal is synthesized from a keyword combination, a suboptimal output has two possible causes: a *structural weakness* in the keyword combination itself, where the current set fails to expand into the adjacent possible space (Swanson, 1986; Lee and Chung, 2024; Johnson, 2011); or an adequate combination whose textual formulation is poor. To determine the most effective optimization strategy, we introduce a Router Module that assesses the current state to decide between two pathways: Keyword Evolution ($d_{keyword}$, modifying the conceptual composition) or Proposal Evolution ($d_{proposal}$, refining the textual realization).

Formally, the router acts as a decision function $d = \Phi_{router}(\mathcal{K}_t, P_t, \mathcal{H}_t)$, where the decision outcome $d \in \{d_{keyword}, d_{proposal}\}$.

Keyword Evolution: If the router detects structural weaknesses in the concept combination($d = d_{keyword}$), the Keyword Selection & Evolve Module (Φ_{evo}) is activated to refine the conceptual composition. This unit evaluates the external candidates $\mathcal{C}_{explore}$ against the current set $\mathcal{K}_t$ to clearly identify a substitution pair—a superior external candidate v_{in} and the least contributing incumbent v_{out}:

$$(v_{in}, v_{out}) = \Phi_{evo}(\mathcal{C}_{explore}, \mathcal{K}_t, \mathcal{H}_t)$$

Subsequently, the keyword set is updated by executing this substitution: $\mathcal{K}_{t+1} = (\mathcal{K}_t \setminus \{v_{out}\}) \cup \{v_{in}\}$.

Proposal Evolution: Conversely, if the router finds the current concepts robust but the narrative execution suboptimal ($d = d_{proposal}$), the system maintains the current keyword set ($\mathcal{K}_{t+1} = \mathcal{K}_t$). The agent then focuses solely on re-synthesizing the proposal to improve logical coherence or explore alternative perspectives.

Following this routing decision, the Idea Formulation Module generates the updated proposal P_{t+1} based on the determined $\mathcal{K}_{t+1}$. The entire evolution step—including the router's decision and the resulting changes—is recorded in the Idea Stack:

$$P_{t+1} = \Phi_{gen}(\mathcal{K}_{t+1}, \mathcal{H}_t)$$
$$\mathcal{H}_{t+1} = \mathcal{H}_t \oplus (d, \mathcal{K}_{t+1}, P_{t+1})$$

To ensure convergence and efficiency, the iterative process terminates if the Novelty and Feasibility scores fail to improve for three consecutive rounds or if the maximum iteration limit(T_{max}) is reached. Upon termination, the proposal with the highest historical score is selected as the final output.

3.3 Expert-Aligned Evolutionary Steering

While the Explore-Expand-Evolve workflow establishes a comprehensive mechanism for network navigation and idea generation, the system still lacks a crucial feedback signal to steer the optimization direction. Without this explicit guidance, the iterative process risks becoming blind, potentially drifting toward valid but scientifically trivial combinations despite the structural integrity of the search. Although standard Large Language Models could be employed to provide this feedback, they often fall short in replicating the nuanced evaluative reasoning employed by expert reviewers, generating surface-level assessments that lack deep, domain-specific understanding.

To overcome this limitation, we designed a "Scientific Reasoning Simulation" prompt to enable the LLM to provide review feedback aligned with human reviewers' scientific thinking. In this prompt, the LLM is directed to simulate a reviewer's cognitive process, evaluating an idea's novelty and feasibility based on existing research. This simulated reasoning is structured into training data to fine-tune the Critic Model, aligning its evaluations with peer review standards. The process of constructing the training data is referred to in Figure 3. Finally, the generated novelty and feasibility feedback is explicitly recorded in the Idea Stack ($\mathcal{H}_{t+1}$) to be utilized in the subsequent iteration. Feasibility thus enters the iterative optimization objective directly alongside novelty, ensuring the two criteria are improved in tandem rather than traded off.

4 Experiments

4.1 Experimental Setup

Dataset. For the dataset, we curated a collection of about 100,000 research papers from major AI conferences over the past decade. These papers were grouped into four categories: DL, NLP, CV, and General AI. To train the Critic Model, we collected the latest peer review data from top-tier AI conferences (ICLR, NeurIPS, ICML) via OpenReview. This strict temporal selection is critical because scientific novelty is time-sensitive—an idea considered innovative in 2024 may be obsolete by 2025. The dataset details are provided in the Appendix A.2.

Baselines. We compare our approach with several prominent methods in AI-driven scientific discovery, including the embedding-based method Sci. Net. Emb. (Sourati and Evans, 2023) and LLM-based methods such as SciMON (Wang et al., 2024a), SciAgents (Ghafarollahi and Buehler, 2025), MOOSE-Chem (Yang et al., 2024), Zero-Shot Hypothesis Proposers (Qi et al., 2023), ResearchAgent (Baek et al., 2024), research ideation agent of (Si et al., 2025b) and papers accepted in the latest year from major AI conferences. More details are presented in Appendix A.3.

Implementation Details. In the Deep-Ideation framework, we use GPT-4o-mini for all the components, while Qwen3-8B is used for Critic Model. We evaluate the **Novelty** and **Feasibility** of the generated ideas using five advanced models: GPT-4o, Gemini-2.5-Flash, Grok-3, DeepSeek-V3.1, and Qwen3-235B-A22B. The final performance score

Method	DL			NLP			CV			General AI			Overall		
	Nov.	Fea.	Avg.	Nov.	Fea.	Avg.	Nov.	Fea.	Avg.	Nov.	Fea.	Avg.	Nov.	Fea.	Avg.
Accepted Papers	3.72	3.93	3.83	3.70	3.95	3.83	3.73	3.86	3.78	3.68	3.90	3.79	3.71	3.91	3.81
Sci. Net. Emb.	3.34	3.53	3.44	3.24	3.61	3.43	2.64	3.44	3.04	3.26	3.57	3.42	3.12	3.53	3.33
Scimon	3.19	3.48	3.34	3.31	3.65	3.24	2.50	3.02	2.76	3.44	3.52	3.48	3.11	3.42	3.27
SciAgents	2.93	3.63	3.28	2.86	3.69	3.28	2.73	3.65	3.19	2.75	3.27	3.01	2.82	3.46	3.14
MOOSE-Chem	3.53	3.33	3.43	3.43	3.22	3.33	3.34	3.21	3.28	3.46	3.07	3.27	3.44	3.21	3.33
Zero-Shot HP	2.80	3.65	3.23	2.73	3.46	3.10	2.78	3.61	3.20	2.81	3.51	3.16	2.78	3.57	3.18
ResearchAgent	3.38	3.22	3.30	3.54	3.33	3.44	3.48	3.28	3.38	3.43	3.25	3.34	3.46	3.47	3.47
Research Ideation	3.48	3.60	3.54	3.57	3.73	3.65	3.47	3.67	3.57	3.46	3.60	3.53	3.50	3.65	3.58
Deep-Ideation	**3.79***	**3.86***	**3.83***	**3.70***	**3.92***	**3.81***	**3.81***	**3.89***	**3.85***	**3.73***	**3.90***	**3.82***	**3.76***	**3.88***	**3.82***
Improvement↑	7.37%	5.75%	8.19%	3.64%	5.09%	4.38%	9.48%	5.99%	7.84%	7.80%	8.33%	8.22%	7.43%	6.30%	6.70%

Table 1: Performance of Deep-Ideation with LLM as Judge compared to Baselines across Different AI Domains. Bold and underline indicate the best and second best performance(except Accepted Papers). * implies the improvements over the second-best results are statistically significant (p-value < 0.05).

Method	DL			NLP			CV			General AI			Overall		
	Nov.	Fea.	Avg.	Nov.	Fea.	Avg.	Nov.	Fea.	Avg.	Nov.	Fea.	Avg.	Nov.	Fea.	Avg.
Accepted Papers	3.57	3.74	3.66	3.63	3.80	3.72	3.70	3.56	3.63	3.54	3.72	3.63	3.61	3.71	3.66
Sci. Net. Emb.	3.15	3.24	3.19	3.11	3.24	3.18	2.94	3.25	3.10	3.22	3.35	3.29	3.11	3.27	3.19
Scimon	2.94	2.85	2.90	2.96	3.16	3.06	3.11	3.04	3.08	2.89	3.24	3.07	2.98	3.07	3.02
SciAgents	3.19	3.20	3.20	3.11	3.19	3.15	3.15	3.14	3.15	2.72	3.09	2.91	3.04	3.16	3.10
MOOSE-Chem	3.24	3.20	3.22	3.07	3.23	3.15	3.35	3.38	3.37	3.35	3.21	3.28	3.25	3.26	3.25
Zero-Shot HP	2.72	3.22	2.97	2.64	3.22	2.93	3.11	3.31	3.21	2.76	3.11	2.94	2.81	3.22	3.01
ResearchAgent	3.22	3.26	3.24	3.24	3.41	3.33	3.22	3.19	3.20	3.23	3.50	3.47	3.23	3.34	3.28
Research Ideation	3.27	3.30	3.29	3.26	3.48	3.37	3.36	3.23	3.30	3.29	3.44	3.37	3.30	3.36	3.33
Deep-Ideation	**3.65***	**3.72***	**3.69***	**3.61***	**3.76***	**3.69***	**3.74***	**3.57***	**3.66***	**3.61***	**3.81***	**3.71***	**3.65***	**3.72***	**3.68***

Table 2: Performance of Deep-Ideation with human evaluation compared to Baselines across Different AI Domains. Bold and underline indicate the best and second best performance(except Accepted Papers). * implies the improvements over the second-best results are statistically significant (p-value < 0.05).

is averaged across these models. More details, including an inference-cost analysis, are provided in Appendices A.4 and A.12.

4.2 Experimental Results

4.2.1 LLM as Judge Evaluation Results

As shown in Table 1, Deep-Ideation demonstrates a significant improvement across all domains. Specifically, Deep-Ideation shows a substantial increase in the Avg. score: 8.19% in the DL domain, 4.38% in the NLP domain, 7.84% in the CV domain, and 8.22% in the General AI domain compared to the best-performing baseline. Notably, beyond these average metric improvements, our analysis reveals that about 81.5% of the generated ideas surpass the average acceptance scores of top-tier AI conferences. Our framework's significant lead in Novelty validates its capacity to discover distant, non-obvious combinations beyond the reach of static retrieval, while the high Feasibility scores indicate that the Critic Model and Idea Stack effectively prevent the semantic drift common in unguided generation. Since the seeds are sampled at random (Appendix A.4), the reported statistical significance confirms that the gains hold consistently across domains regardless of the starting anchor. Additional analyses—parameter sensitivity, case study, scaling with stronger backbones, and per-run standard deviations—are deferred to Appendices A.7, A.8, A.11, and A.6.

4.2.2 Human Evaluation Result

To further evaluate the effectiveness of the Deep-Ideation framework, a human evaluation is conducted involving 54 researchers, and details are provided in the Appendix A.9). The human evaluation results are presented in Table 2.

The results shows that our method consistently outperformed the best baseline, demonstrating Deep Ideation's ability to generate more valuable and well-structured ideas. Qualitative feedback specifically highlights the framework's ability to generate novel proposals. Human reviewers commended the system for bridging distinct sub-fields to create innovative yet logical combinations. Technically, this cross-domain synergy validates the utility of mining high-order neighbors, demonstrating that accessing distant nodes in the concept network

Method	DL			NLP			CV			General AI			Overall		
	Nov.	Fea.	Avg.	Nov.	Fea.	Avg.	Nov.	Fea.	Avg.	Nov.	Fea.	Avg.	Nov.	Fea.	Avg.
Deep-Ideation(full)	3.79	3.86	3.83	3.70	3.92	3.81	3.81	3.89	3.85	3.73	3.90	3.82	3.76	3.88	3.82
w/o Evolve	3.74	3.68	3.71	3.59	3.82	3.71	3.64	3.80	3.72	3.66	3.74	3.70	3.66	3.76	3.71
w/o Critic Model	3.61	3.64	3.63	3.43	3.63	3.53	3.58	3.63	3.61	3.52	3.62	3.57	3.54	3.63	3.59

Table 3: Ablation study of Deep-Ideation across different AI domains.

is essential for breaking disciplinary silos. An analysis of the correlation between LLM and human evaluations is provided in Appendix A.10.

4.3 Ablation Study of Deep-Ideation

To validate the effectiveness of Evolve Mechanism and Critic-Model in Deep Ideation, we conducted an ablation study. The results of this experiment are presented in Table 3.

Effectiveness of Evolve Mechanism. The drop in the "w/o Evolve" setting (Table 3) underscores the necessity of iterative refinement. Without Evolve, the agent is restricted to greedy expansion, where early suboptimal keyword choices become permanent constraints; the Evolve module breaks this dependency by allowing keyword substitution to escape local optima.

Effectiveness of Critic Model. The simultaneous decline in both metrics confirms the Critic Model's role as a quality discriminator. Without this feedback, the search loses the incentive to pursue distant connections (lowering Novelty) and the ability to reject hallucinatory associations (lowering Feasibility).

5 Related Works

5.1 LLMs for Scientific Research

Large Language Models are transforming specific stages of the scientific lifecycle, ranging from information synthesis (Team et al., 2025; Schmidgall and Moor, 2025; Wu et al., 2025a) to code optimization (Romera-Paredes et al., 2024; Zhai et al., 2025; Lange et al., 2025; Liu et al., 2025). WebThinker (Li et al., 2025a) automates literature review by dynamically navigating the web. Virtual Lab (Swanson et al., 2025) generates ideas through multi-agent collaboration. AlphaEvolve (Novikov et al., 2025) uses evolutionary coding agents to optimize algorithms beyond human baselines. AutoSurvey (Wang et al., 2024c) automates the retrieval and drafting of literature surveys. Beyond individual stages, recent efforts automate the entire research loop: DeepScientist (Weng et al., 2025) performs goal-oriented discovery via Bayesian optimization, while The AI Scientist-v2 (Yamada et al., 2025) achieves end-to-end autonomy with the first peer-reviewed AI-generated manuscript.

5.2 LLMs for Scientific Ideation

Recent LLM-based ideation systems have advanced automated discovery from several angles. SciAgents (Ghafarollahi and Buehler, 2025) refines hypotheses via multi-agent collaboration. MOOSE-Chem (Yang et al., 2024) rediscovers hidden chemistry knowledge with LLMs. Qi et al. (Qi et al., 2023) demonstrate the zero-shot hypothesis generation capability of LLMs. CycleResearcher (Weng et al., 2024) introduces a self-supervised feedback loop with automated reviews. However, these methods largely overlook the structural guidance of scientific networks. While SciMon (Wang et al., 2024a) and ResearchAgent (Baek et al., 2024) leverage citation or academic graphs to augment ideas with retrieved entities, both restrict their use of the graph to a one-off context retrieval at the initial stage and do not dynamically re-query it during iterative optimization. Deep-Ideation fundamentally differs by establishing continuous dynamic interaction with the scientific network. Building on the theory of the "adjacent possible" (Johnson, 2011), this continuous interaction lets the agent systematically explore the expanding boundary of knowledge and recombine fresh structural combinations that one-off retrieval pipelines never observe.

6 Conclusion

In this paper, we propose Deep-Ideation, driven by the motivation to uncover high-order neighbors and establish continuous dynamic interaction with the scientific network. Leveraging an iterative Explore-Expand-Evolve workflow governed by an Idea Stack and Critic Model, our approach actively aligns its search trajectory with the evolving research context. Experiments demonstrate that Deep Ideation outperforms leading baselines, achieving superior Novelty through deep graph exploration while maintaining high Feasibility via rigorous, feedback-driven optimization.

Limitations

While Deep-Ideation interacts with the scientific concept network continuously throughout the optimization process, its structural rigidity lies strictly in the predefined procedural pipeline rather than in the agent's decision-making: within this workflow the agent retains full autonomy to evaluate candidates and steer the search toward the evolving research context. To elevate this autonomy further, our future work will move to a fully Agentic LLM framework that allows the agent to autonomously plan and adjust its high-level actions—choosing to Explore, Expand, or Evolve dynamically—on the fly, taking a step toward more fluid scientific discovery.

References

Anna Badalyan, Nicolò Ruggeri, and Caterina De Bacco. 2024. Structure and inference in hypergraphs with node attributes. *Nature Communications*, 15(1):7073.

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

Mathias Benedek, Tanja Könen, and Aljoscha C Neubauer. 2012. Associative abilities underlying creativity. *Psychology of Aesthetics, Creativity, and the Arts*, 6(3):273.

Mario Coccia. 2019. Why do nations produce science advances and new technology? *Technology in society*, 59:101124.

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

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, and 1 others. 2025. Towards an ai co-scientist. *arXiv preprint arXiv:2502.18864*.

Steven Johnson. 2011. *Where good ideas come from: The natural history of innovation*. Penguin.

Arthur Koestler. 1964. The act of creation.

Robert Tjarko Lange, Yuki Imajuku, and Edoardo Cetin. 2025. Shinkaevolve: Towards open-ended and sample-efficient program evolution. *arXiv preprint arXiv:2509.19349*.

Pat Langley. 1987. *Scientific discovery: Computational explorations of the creative processes*. MIT press.

Byung Cheol Lee and Jaeyeon Chung. 2024. An empirical investigation of the impact of chatgpt on creativity. *Nature Human Behaviour*, 8(10):1906–1914.

Long Li, Weiwen Xu, Jiayan Guo, Ruochen Zhao, Xingxuan Li, Yuqian Yuan, Boqiang Zhang, Yuming Jiang, Yifei Xin, Ronghao Dang, and 1 others. 2024. Chain of ideas: Revolutionizing research via novel idea development with llm agents. *arXiv preprint arXiv:2410.13185*.

Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yongkang Wu, Ji-Rong Wen, Yutao Zhu, and Zhicheng Dou. 2025a. Webthinker: Empowering large reasoning models with deep research capability. *arXiv preprint arXiv:2504.21776*.

Zijian Li, Xin Guan, Bo Zhang, Shen Huang, Houquan Zhou, Shaopeng Lai, Ming Yan, Yong Jiang, Pengjun Xie, Fei Huang, and 1 others. 2025b. Webweaver: Structuring web-scale evidence with dynamic outlines for open-ended deep research. *arXiv preprint arXiv:2509.13312*.

Gang Liu, Yihan Zhu, Jie Chen, and Meng Jiang. 2025. Scientific algorithm discovery by augmenting alphaevolve with deep research. *arXiv preprint arXiv:2510.06056*.

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

Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco JR Ruiz, Abbas Mehrabian, and 1 others. 2025. Alphaevolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*.

Yongqian Peng, Yuxi Ma, Mengmeng Wang, Yuxuan Wang, Yizhou Wang, Chi Zhang, Yixin Zhu, and Zilong Zheng. 2025. Probing and inducing combinational creativity in vision-language models. *arXiv preprint arXiv:2504.13120*.

Kevin Pu, KJ Kevin Feng, Tovi Grossman, Tom Hope, Bhavana Dalvi Mishra, Matt Latzke, Jonathan Bragg, Joseph Chee Chang, and Pao Siangliulue. 2025. Ideasynth: Iterative research idea development through evolving and composing idea facets with literature-grounded feedback. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–31.

Biqing Qi, Kaiyan Zhang, Haoxiang Li, Kai Tian, Sihang Zeng, Zhang-Ren Chen, and Bowen Zhou. 2023. Large language models are zero shot hypothesis proposers. *arXiv preprint arXiv:2311.05965*.

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*.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M Pawan Kumar, Emilien Dupont, Francisco JR Ruiz, Jordan S Ellenberg, Pengming Wang, Omar Fawzi, and 1 others. 2024. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475.

Samuel Schmidgall and Michael Moor. 2025. Agentrxiv: Towards collaborative autonomous research. *arXiv preprint arXiv:2503.18102*.

Chenglei Si, Tatsunori Hashimoto, and Diyi Yang. 2025a. The ideation-execution gap: Execution outcomes of llm-generated versus human research ideas. *arXiv preprint arXiv:2506.20803*.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2024. Can llms generate novel research ideas. *A large-scale human study with*, 100.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2025b. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. In *International Conference on Learning Representations*, volume 2025, pages 94003–94092.

Jamshid Sourati and James A Evans. 2023. Accelerating science with human-aware artificial intelligence. *Nature human behaviour*, 7(10):1682–1696.

Haoyang Su, Renqi Chen, Shixiang Tang, Zhenfei Yin, Xinzhe Zheng, Jinzhe Li, Biqing Qi, Qi Wu, Hui Li, Wanli Ouyang, and 1 others. 2024. Many heads are better than one: Improved scientific idea generation by a llm-based multi-agent system. *arXiv preprint arXiv:2410.09403*.

Don R Swanson. 1986. Undiscovered public knowledge. *The Library Quarterly*, 56(2):103–118.

Kyle Swanson, Wesley Wu, Nash L Bulaong, John E Pak, and James Zou. 2025. The virtual lab of ai agents designs new sars-cov-2 nanobodies. *Nature*, pages 1–3.

Tongyi DeepResearch Team, Baixuan Li, Bo Zhang, Dingchu Zhang, Fei Huang, Guangyu Li, Guoxin Chen, Huifeng Yin, Jialong Wu, Jingren Zhou, and 1 others. 2025. Tongyi deepresearch technical report. *arXiv preprint arXiv:2510.24701*.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, and 1 others. 2023. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60.

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

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

Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Qingsong Wen, Wei Ye, and 1 others. 2024c. Autosurvey: Large language models can automatically write surveys. *Advances in neural information processing systems*, 37:115119–115145.

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

Yixuan Weng, Minjun Zhu, Qiujie Xie, Qiyao Sun, Zhen Lin, Sifan Liu, and Yue Zhang. 2025. Deepscientist: Advancing frontier-pushing scientific findings progressively. *arXiv preprint arXiv:2509.26603*.

Jialong Wu, Baixuan Li, Runnan Fang, Wenbiao Yin, Liwen Zhang, Zhengwei Tao, Dingchu Zhang, Zekun Xi, Gang Fu, Yong Jiang, and 1 others. 2025a. Webdancer: Towards autonomous information seeking agency. *arXiv preprint arXiv:2505.22648*.

Qianhui Wu, Kanzhi Cheng, Rui Yang, Chaoyun Zhang, Jianwei Yang, Huiqiang Jiang, Jian Mu, Baolin Peng, Bo Qiao, Reuben Tan, and 1 others. 2025b. Gui-actor: Coordinate-free visual grounding for gui agents. *arXiv preprint arXiv:2506.03143*.

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

Zonglin Yang, Wanhao Liu, Ben Gao, Tong Xie, Yuqiang Li, Wanli Ouyang, Soujanya Poria, Erik Cambria, and Dongzhan Zhou. 2024. Moose-chem: Large language models for rediscovering unseen chemistry scientific hypotheses. *arXiv preprint arXiv:2410.07076*.

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*.

Yi Zhai, Zhiqiang Wei, Ruohan Li, Keyu Pan, Shuo Liu, Lu Zhang, Jianmin Ji, Wuyang Zhang, Yu Zhang, and Yanyong Zhang. 2025. \(x\)-evolve: Solution space evolution powered by large language models. *arXiv preprint arXiv:2508.07932*.

Xuyang Zhao, Shiwan Zhao, Hualong Yu, Liting Zhang, and Qicheng Li. 2025. Agentcdm: Enhancing multi-agent collaborative decision-making via ach-inspired structured reasoning. *arXiv preprint arXiv:2508.11995*.

A Appendix

A.1 Algorithmic Details

For clarity, we summarize all key notations used in the algorithm below:

- $\mathcal{G}$: the Scientific Concept Network;
- $\mathcal{K}_t$: the keyword set at iteration t;
- L_{target}: the target size of the keyword set;
- T_{max}: the maximum number of iterations;
- $\mathcal{C}_{explore}$: the set of relation-annotated candidate neighbors obtained in the Explore phase;
- $\mathcal{H}_t$: the Idea Stack, recording the full history of keyword changes, proposals, and scores up to iteration t;
- P_t: the research proposal generated at iteration t;
- s_t: the combined novelty and feasibility score assigned by the Critic Model at iteration t;
- d: the routing decision, either $d_{keyword}$ (Keyword Evolution) or $d_{proposal}$ (Proposal Evolution);
- Φ_{exp}: Keyword Selection & Expand Module;
- Φ_{evo}: Keyword Selection & Evolve Module;
- Φ_{gen}: Idea Formulation Module;
- Φ_{router}: Router Module;
- $\hat{\mathcal{J}}_\theta$: the fine-tuned Critic Model.

Algorithm 1 presents the complete iterative procedure of Deep-Ideation. The algorithm makes explicit that (i) a research proposal is generated after **every** keyword set update, including during the Expand phase before the set reaches its target size; (ii) the Expand branch only **grows** the keyword set, whereas the Keyword Evolution branch performs an in-place **replacement**; and (iii) the Router decides between Keyword Evolution and Proposal Evolution based on whether the current keyword combination exhibits a structural weakness.

A.2 Dataset Details

The dataset consists of 108,825 research papers sourced from major AI conferences over the past decade, including ICLR, NeurIPS, ICML, ACL, NAACL, CVPR, ICCV, AAAI, and IJCAI. These papers were grouped into four categories:

- **DL (short for Deep Learning):** ICLR, NeurIPS, ICML
- **NLP (short for Natural Language Process)**: ACL, NAACL
- **CV (Computer Vision)**: CVPR, ICCV
- **General AI**: AAAI, IJCAI

The detailed distribution of collected papers across each conference is summarized in Table A.4.

Conference	Count	Percentage
NeurIPS	19,186	17.6%
CVPR	17,433	16.0%
AAAI	14,978	13.8%
ICML	13,613	12.5%
ICLR	11,329	10.4%
IJCAI	8,178	7.5%
ACL	7,900	7.3%
EMNLP	6,869	6.3%
ICCV	5,989	5.5%
NAACL	3,350	3.1%
Total	**108,825**	**100.0%**

Table A.4: Statistics of the collected research papers from major AI conferences.

In constructing the Scientific Concept Network, we employed a strictly temporal strategy by using all collected papers except those from the most recent year. This "leave-last-year-out" approach ensures that the concept network is built solely on historical knowledge, thereby preventing information leakage from the evaluation period.

For each paper, 3-4 keywords were extracted, forming the basis of a scientific network constructed from keyword co-occurrence. During the idea proposal generation process, an initial keyword was selected from a specific domain. The resulting idea proposal was then classified according to the domain from which the initial keyword was drawn.

A.3 Baseline Methods

We benchmark our approach against several prominent methods in AI-driven scientific discovery:

- **Sci. Net. Emb.:** (Sourati and Evans, 2023) This method integrates human expertise into AI models to enhance predictions of future scientific breakthroughs, particularly in data-scarce contexts. By considering the distribution of human

Algorithm 1: Deep-Ideation: Explore-Expand-Evolve

Input: seed keyword v_0, Scientific Concept Network $\mathcal{G}$, target keyword set size L_{target}, max iterations T_{max}, Critic Model $\hat{\mathcal{J}}_\theta$

Output: best proposal P^*

$\mathcal{K}_0 \leftarrow \{v_0\}$; $\mathcal{H}_0 \leftarrow \varnothing$; $P^* \leftarrow \varnothing$; $s^* \leftarrow -\infty$;

for $t = 0, 1, \ldots, T_{max} - 1$ **do**

  `// - Explore -`

  $\mathcal{C}_{explore} \leftarrow \text{Explore}(\mathcal{K}_t, \mathcal{G})$; `// retrieve and annotate neighboring candidates`

  if $|\mathcal{K}_t| < L_{target}$ **then**

    `// - Expand (grow only) -`

    $v^* \leftarrow \Phi_{exp}(\mathcal{C}_{explore}, \mathcal{H}_t)$; `// select the best candidate to add`

    $\mathcal{K}_{t+1} \leftarrow \mathcal{K}_t \cup \{v^*\}$;

  else

    `// - Evolve -`

    $d \leftarrow \Phi_{router}(\mathcal{K}_t, P_t, \mathcal{H}_t)$; `// diagnose: structural weakness or textual issue?`

    if $d = d_{keyword}$ **then**

      `// Keyword Evolution (replacement)`

      $(v_{in}, v_{out}) \leftarrow \Phi_{evo}(\mathcal{C}_{explore}, \mathcal{K}_t, \mathcal{H}_t)$;

      $\mathcal{K}_{t+1} \leftarrow (\mathcal{K}_t \setminus \{v_{out}\}) \cup \{v_{in}\}$;

    else

      `// Proposal Evolution (refine text only)`

      $\mathcal{K}_{t+1} \leftarrow \mathcal{K}_t$;

    end

  end

  `// - Generate & Score -`

  $P_{t+1} \leftarrow \Phi_{gen}(\mathcal{K}_{t+1}, \mathcal{H}_t)$; `// synthesize proposal from current keywords`

  $s_{t+1} \leftarrow \hat{\mathcal{J}}_\theta(P_{t+1})$; `// Critic scores novelty + feasibility`

  if $s_{t+1} > s^*$ **then**

    $s^* \leftarrow s_{t+1}$; $P^* \leftarrow P_{t+1}$;

  end

  $\mathcal{H}_{t+1} \leftarrow \mathcal{H}_t \oplus (\mathcal{K}_{t+1}, P_{t+1}, s_{t+1})$;

  `// - Convergence check -`

  if *score s has not improved for 3 consecutive iterations* **then**

    break;

  end

end

return P^*;

expertise, it improves AI-driven predictions beyond traditional research content.

- **SciMON:** (Wang et al., 2024a) SciMON focuses on optimizing neural language models for novelty. It iteratively refines generated hypotheses by comparing them with existing literature, aiming to improve both the technical depth and originality of the generated ideas.
- **SciAgents:** (Ghafarollahi and Buehler, 2025) This method employs ontological knowledge graphs and multi-agent systems to autonomously generate and refine hypotheses. By uncovering interdisciplinary connections, it accelerates material discovery and fosters new research avenues.
- **MOOSE-Chem:** (Yang et al., 2024) MOOSE-Chem applies a structured framework to generate hypotheses in chemistry. It demonstrates the potential of LLMs in rediscovering scientifically valuable insights and advancing the hypothesis generation process in the field of chemistry.

Title: Enhancing Multi-Face Forgery Detection and Deepfake Prevention through Reinforcement Learning

Research Background:
.....deepfake content becomes increasingly sophisticated, particularly in scenarios involving multiple individuals, traditional detection methods.....
Research Idea:
.....introduce a dual-task model capable of performing simultaneous face detection and segmentation.....adaptability through RL techniques..... select the most effective segmentation strategies based on real-time input.....
General Implementation Approach:
.....Develop a dual-task CNN architecture that integrates both face detection and segmentation tasks.....allow the model to dynamically choose the most appropriate segmentation strategy .....

Title: Enhancing Adaptive Self-Evolution and Preference Alignment with Reinforcement Learning through Noise Contrastive Estimation

Research Background:
..... depend on human-annotated data, which can suffer from bias, limitations..... Noise Contrastive Estimation emerges as a promising method.....as demonstrated in "Noise Contrastive Alignment of Language Models with Explicit Rewards. ".....
Research Idea:
..... incorporating a structured set of multimodal instructions to serve as a foundation ..... Implementing NCE to enhance the robustness of preference alignment.....
General Implementation Approach:
..... specifically employing MPO to aggregate diverse user preferences ..... Integrate NCE into the training process.....to differentiate between relevant and irrelevant data points.....

Figure A.4: A case study of idea proposal generated by Deep Ideation.

- **Zero-Shot Hypothesis Proposers:** (Qi et al., 2023) This method explores the ability of LLMs to propose valid hypotheses without prior fine-tuning. It showcases the capability of LLMs to generate novel scientific ideas from unseen literature, pushing the boundaries of zero-shot hypothesis generation.

- **ResearchAgent:** (Baek et al., 2024) ResearchAgent combines iterative idea generation with LLM-based review agents to refine scientific proposals. It represents a comprehensive approach to supporting researchers in the ideation process, enhancing both the creativity and rigor of generated ideas.

- **Si et al.:** (Si et al., 2025b) This method instantiates a minimalist LLM ideation agent organized into three components—retrieval-augmented paper retrieval that conditions generation on topic-relevant literature, over-generate-and-deduplicate idea generation, and an LLM-based idea ranker that selects the final outputs.

- **Accepted Papers:** We also include the latest accepted papers from major AI conferences as baselines. Here, we input the title, abstract, and introduction of each paper into the model and asked it to organize the content into an idea proposal format.

A.4 More Details about Implementation

In the Deep-Ideation workflow, GPT-4o-mini (released on July 18, 2024) is used as the backbone model for the modules of Relation Analysis, Keyword Selection, Idea Formulation. Since our evaluation dataset (the test set) consists of the latest papers published after this date, using this model ensures that there is no data leakage from the test set into the model's pre-training data. For the Critic Model, Qwen3-8B is used, which is fine-tuned with Low-Rank Adaptation (LoRA) on a training dataset of 4278 examples. Each training instance follows the prompt–thinking–answer template adopted by current Large Reasoning Models (illustrated in the right panel of Figure 3): the prompt presents the candidate idea proposal to be judged, the thinking segment records the simulated reviewer reasoning before the final decision, and the answer holds the resulting novelty and feasibility scores. Fine-tuning is performed with the standard cross-entropy objective conventionally used in LLM training. This fine-tuning process enhances the model's ability to evaluate the novelty and feasibility of the generated ideas.

For the evaluation phase, we constructed a dedicated test set to rigorously assess the generation capabilities of each method. Specifically, for each of the four domains (DL, NLP, CV, General AI), we randomly selected 50 representative keywords from the validation partition (i.e., the data from the most recent year) to serve as the initial inputs for idea generation. These initial keywords are drawn entirely at random from the validation partition, so every reported number is averaged over randomized starting anchors rather than hand-picked seeds.

To assess the quality of the generated ideas, five advanced large language models are used for evaluation: GPT-4o, Gemini-2.5-Flash, Grok-3, DeepSeek-V3.1, and Qwen3-235B-A22B. The final performance scores are derived by averaging the results across these models, ensuring a robust and comprehensive evaluation of the ideas' quality. It is worth noting that the optimization loop and the benchmark evaluation are deliberately decoupled: search-time feedback comes exclusively from the small fine-tuned Qwen3-8B Critic, whereas test-

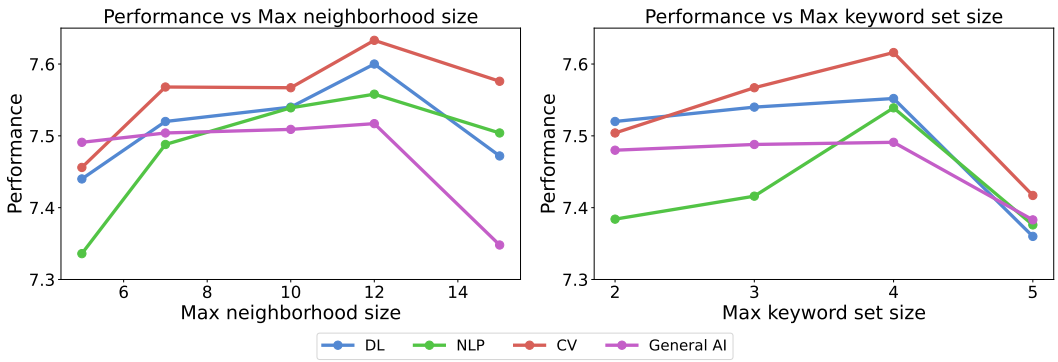

Figure A.5: Effect of max neighborhood size and max keyword set size, where performance is the sum of novelty and feasibility.

time scoring is delegated to the disjoint ensemble of five frontier judges listed above, applied identically to all baselines. The framework therefore does not directly optimize against the test oracle. The double-blind human evaluation with 54 domain experts (Appendix A.9) further corroborates this design, showing consistent gains over all baselines and confirming that the improvements are not an artifact of the automated metric.

A.5 Details of Keyword Extraction

Although the underlying connectivity signal is keyword co-occurrence, the keyword set itself is produced by a controlled two-step LLM prompt rather than by raw n-gram mining. We initially defined a broad, standardized set of AI concepts (e.g., "Large Language Models"). For each collected paper, the LLM is instructed to first map the paper to relevant broad concepts from this predefined set to establish foundational network connectivity. Subsequently, the LLM extracts highly specific, method-related keywords directly from the paper's content to capture its unique scientific contribution. This dual-layered strategy guarantees that the extracted keywords possess low noise, avoid total overlap, and balance the broad connectivity required for graph construction with the distinct specificity needed for innovation.

A.6 Standard Deviation of Evaluation Scores

To provide a more comprehensive picture of the evaluation stability, we report the mean and standard deviation of LLM-as-judge scores for Deep-Ideation and the best-performing baseline across multiple runs. As shown in Table A.5, Deep-Ideation consistently outperforms the best baseline across all domains and metrics.

A.7 Analysis

In this section, we analyzed key elements of the Deep Ideation framework. Max neighborhood size controls the breadth of knowledge sampled for each keyword, while max keyword set size defines the number of keywords used to generate the idea proposal. The results are shown in Figure A.5.

A.7.1 Effect of max neighborhood size.

As shown in Figure A.5 (left), Deep Ideation performs best when the maximum neighborhood size is set to 12. When smaller, the limited scientific knowledge surrounding each keyword restricts the agent's ability to capture comprehensive insights, diminishing the quality and depth of the final idea proposals. Conversely, increasing the neighborhood size beyond 12 expands the agent's knowledge boundary, but may lead to information overload, making it difficult for the agent to prioritize the most valuable insights and causing the focus of the generated ideas to become diluted.

A.7.2 Effect of max keyword set size

Figure A.5 (right) illustrates that, when the keyword set size is small, the knowledge breadth and diversity of the final idea proposal are limited, which results in ideas that lack innovation and depth, often failing to address the complexities of scientific problems. In contrast, increasing the keyword set size too much leads to overly complex relationships between the keywords, causing the ideas to become disjointed or unnatural, with forced connections that undermine clarity. However, when the keyword set size is set to 4, the performance improves significantly, indicating that a balanced set allows the agent to capture sufficient diversity and depth while keeping the generated ideas focused and logically connected.

Method	DL			NLP			CV			General AI		
	Nov.	Fea.	Avg.	Nov.	Fea.	Avg.	Nov.	Fea.	Avg.	Nov.	Fea.	Avg.
Best Baseline	$3.53_{\pm.03}$	$3.65_{\pm.03}$	$3.54_{\pm.03}$	$3.57_{\pm.02}$	$3.73_{\pm.02}$	$3.65_{\pm.02}$	$3.48_{\pm.04}$	$3.67_{\pm.03}$	$3.57_{\pm.03}$	$3.46_{\pm.03}$	$3.60_{\pm.03}$	$3.53_{\pm.03}$
Deep-Ideation	$\mathbf{3.79}_{\pm.05}$	$\mathbf{3.86}_{\pm.05}$	$\mathbf{3.83}_{\pm.04}$	$\mathbf{3.70}_{\pm.05}$	$\mathbf{3.92}_{\pm.05}$	$\mathbf{3.81}_{\pm.04}$	$\mathbf{3.81}_{\pm.06}$	$\mathbf{3.89}_{\pm.05}$	$\mathbf{3.85}_{\pm.05}$	$\mathbf{3.73}_{\pm.06}$	$\mathbf{3.90}_{\pm.05}$	$\mathbf{3.82}_{\pm.05}$

Table A.5: Mean scores with standard deviations (subscripts) for Deep-Ideation and the best baseline across multiple experimental runs.

A.8 Case Study

These case studies in Figure A.4 demonstrate Deep Ideation's ability to generate novel solutions. In multi-face forgery detection and deepfake prevention, the dual-task model innovatively combines face detection and segmentation with reinforcement learning, dynamically selecting the best strategies in real-time. In adaptive self-evolution and preference alignment, incorporating Noise Contrastive Estimation (NCE) into reinforcement learning offers a novel approach to overcoming biases in human-annotated data, improving preference robustness. Additionally, in the right case, the idea proposal cites a relevant paper, demonstrating how Deep Ideation effectively incorporates existing research to refine and enhance its generated ideas.

A.9 Human evaluation details

To ensure a rigorous and unbiased assessment of the generated ideas, we conducted a comprehensive human evaluation study. The details regarding participant recruitment, data sampling, and the evaluation protocol are described below.

Participants and Expertise. We invited a total of 54 evaluators from diverse research backgrounds. To ensure high-quality judgments, all participants were required to be PhD candidates with at least two publications in top-tier AI conferences (e.g., ACL, NeurIPS, CVPR, AAAI). The evaluators were assigned to the four domains (DL, NLP, CV, General AI) based on their specific research expertise to ensure they possessed the necessary background knowledge to judge the novelty and feasibility of the proposals.

Data Sampling and Assignment. Due to the high cognitive load of manually evaluating research proposals, evaluating the entire test set was impractical. Instead, we adopted a stratified random sampling strategy. From the test set of 50 keywords per domain (described in Appendix A.2), we randomly selected 10 keywords for each domain to serve as the basis for the human study. For each keyword, idea proposals were generated using Deep Ideation and all baseline methods. To ensure statistical robustness, we designed a distributed assignment scheme:

- **Coverage:** Each idea proposal was evaluated by at least three independent human experts.
- **Workload:** Each evaluator was assigned to review idea proposals generated from 3 distinct keywords within their domain. This design ensures that every keyword's resulting ideas are cross-verified by multiple experts while preventing evaluator fatigue.

Evaluation Protocol. The evaluation was conducted in a strictly double-blind manner. All method names were anonymized, and the order of proposals was randomized to prevent bias. Evaluators were asked to score each proposal on a Likert scale from 1 to 5 based on two criteria: Novelty (originality and distinctiveness from existing work) and Feasibility (technical soundness and practical realizability).

Instructions Given to Participants. To ensure consistency and objectivity in scoring, all evaluators were provided with a standardized rubric defining the scoring criteria for Novelty and Feasibility. The specific instructions and scale definitions were as follows:

Instruction

Task Overview:
You will be presented with a research idea proposal generated based on a set of keywords. Please evaluate the proposal independently along two dimensions using a 5-point Likert scale:
1. Novelty (1–5): How original and innovative the idea is compared to existing research.
- 5: Extremely novel and groundbreaking. The idea introduces highly original, unexplored concepts or radically shifts the direction of the field. Only award this score for ideas that truly break new ground, with no close prior work or that significantly alter

the research landscape.
- 4: Highly original. The idea is new and innovative but builds on existing research or concepts. It introduces a novel approach or expands on a well-established direction in a meaningful way. Be cautious when awarding this score; it should be reserved for ideas that show significant originality but do not fully redefine the field.
- 3: Moderately original. The idea brings some new insights but is largely based on existing work or closely follows well-established theories. Award this score for ideas that are derivative or offer incremental improvements without offering a substantial breakthrough.
- 2: Slightly original. The idea offers only minor variations to existing research or reworks known concepts with minimal variation. This score should be given for ideas that are only slightly different or when the novelty is quite limited.
- 1: Not original. The idea closely resembles existing research with little to no innovation. Reserve this score for ideas that simply replicate or overlap with established research, showing no originality.
2. Feasibility (1–5): How realistic and practical the idea is to implement in current scientific and technological conditions.
- 5: Fully feasible. The idea can be realistically executed with existing methods, data, and resources. The implementation plan is clear, actionable, and supported by available technologies. Only award this score for ideas that can be fully realized in the current research landscape without major technological or data challenges.
- 4: Highly feasible. The idea is feasible with current technologies, but may require some advancements or additional resources (e.g., access to new data or refinement of techniques). The idea is practical but not without challenges. This score should be given to ideas that are highly implementable, but may need some further refinement or additional support.
- 3: Moderately feasible. The idea faces significant practical challenges, requiring considerable advancements in technology or data. Award this score for ideas that are plausible but face notable barriers to implementation, whether technological, data-driven, or methodological.
- 2: Slightly feasible. The idea is difficult to implement with current resources and would need significant breakthroughs. Use this score for ideas that are highly ambitious or speculative, and would require substantial breakthroughs to become feasible.
- 1: Not feasible. The idea is impractical and unlikely to be implemented with current technologies or methods. This score should be given to ideas that are far-fetched, lacking a feasible path to implementation within the current scientific context.

A.10 Correlation Analysis between LLM and Human Evaluations

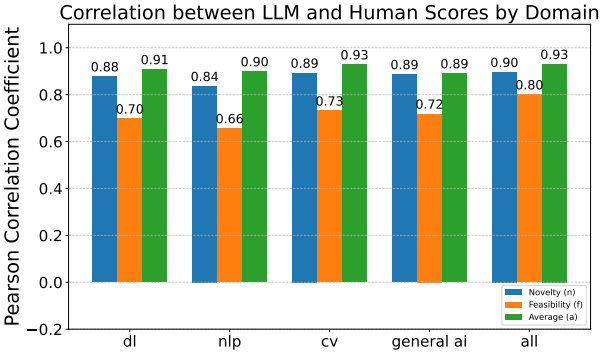

Figure A.6: Correlation between LLM and human scores by domain.

We analyzed the statistical relationship between LLM and human scores using the Pearson Correlation Coefficient, as shown in Figure A.6. The aggregated results indicate a strong alignment. Notably, the correlation for feasibility is consistently lower than for novelty; this likely stems from the fact that feasibility judgments rely heavily on subjective engineering experience, whereas novelty is more objectively verifiable against existing literature. Even so, feasibility still reaches $r = 0.80$ overall, reflecting a tight overall alignment with expert judgments and indicating that the system captures practical implementation hurdles reliably despite the subjectivity inherent to engineering assessment. Among domains, CV exhibits the highest alignment, likely because its innovations often center on explicit architectural patterns (e.g., specific

Method (GPT-5-mini)	DL			NLP			CV			General AI			Overall		
	Nov.	Fea.	Avg.	Nov.	Fea.	Avg.	Nov.	Fea.	Avg.	Nov.	Fea.	Avg.	Nov.	Fea.	Avg.
Best Baseline	3.75	3.86	3.81	3.70	3.87	3.79	3.69	3.82	3.76	3.71	3.79	3.75	3.71	3.84	3.78
Deep-Ideation	**3.98**	**4.11**	**4.05**	**3.85**	**4.13**	**3.99**	**3.98**	**4.06**	**4.02**	**3.94**	**4.13**	**4.04**	**3.94**	**4.11**	**4.03**

Table A.6: Performance with GPT-5-mini as the backbone. Deep-Ideation's overall score rises from 3.82 (GPT-4o-mini, Table 1) to 4.03, while continuing to outperform all baselines across every domain.

modules or visual encoders) that are structurally definable. In contrast, NLP research frequently involves abstract reasoning flows or complex semantic alignments, introducing greater ambiguity and variance in assessing their practical feasibility.

A.11 Scaling with Stronger Backbones

Our framework is model-agnostic, and a more capable backbone naturally enhances the logical coherence and depth of the generated ideas. To validate this empirically, we re-run the full pipeline with GPT-5-mini in place of GPT-4o-mini, keeping all other components unchanged. As shown in Table A.6, the substitution substantially improves the overall scores, indicating that Deep-Ideation continues to benefit from backbone improvements while retaining its state-of-the-art performance.

A.12 Inference Cost Analysis

Producing high-quality, scientifically grounded proposals necessarily requires more computation than single-pass generation. This mirrors the gap between standard Retrieval-Augmented Generation and multi-step DeepResearch-style pipelines, where the multi-iteration retrieval and validation are precisely what make the stronger outputs possible. Table A.7 reports per-idea costs against the two most relevant baselines: Deep-Ideation incurs higher latency and token cost than the simpler agents.

Method	Inference Latency	Token Costs
SciMON	19s	8k
ResearchAgent	106s	54k
Deep-Ideation	**127s**	**63k**

Table A.7: Inference latency and token costs per generated idea.

The cost should be read against the human-research alternative: Si et al. (2025b) report that each idea writer in their study was given 10 days and paid $300 per idea. Relative to that baseline, the additional token consumption of Deep-Ideation is negligible, while the resulting quality reaches the SOTA performance.

A.13 Potential Risks

The primary risk lies in the blurring distinction between AI-generated and human-conceived ideas. As the quality of automated generation improves, the academic community may face challenges in distinguishing the provenance of research proposals, potentially complicating authorship verification. We emphasize that Deep Ideation is designed as a copilot to augment human creativity, necessitating transparency and rigorous human oversight.

A.14 The Use of Large Language Models

In this work, only the grammar correction and sentence-level refinement of the manuscript were carried out using LLMs.