CoI-Pro: Scientific Idea Generation via Retrieval over Research Trajectories

Anonymous authors
Paper under double-blind review

Abstract

Scientific idea generation is crucial for automated research and scientific progress. Existing methods typically follow a three-step pipeline: first constructing idea chains or graphs for a given topic, then selecting an idea generation direction, and finally producing concrete ideas. However, these methods fail to explicitly incorporate historical research trajectories from other topics, which often leads to low-quality or incremental ideas. In contrast, human researchers frequently propose innovative ideas by drawing on research trajectories from other topics, rather than confining themselves to their own topic. For instance, suppose Topic A has a research trajectory of "supervised learning → unsupervised learning → semi-supervised learning." If a new Topic B is currently at the "unsupervised learning" stage, an automated idea generation model without referencing Topic A's trajectory would likely remain confined to proposing ideas within unsupervised learning. Conversely, if provided with Topic A's trajectory as a reference, the model is far more likely to generate a novel idea extending to semi-supervised learning, thereby substantially enhancing both the quality and innovativeness of the generated idea. Thus, inspired by the above intuition, we propose a novel method, called CoI-Pro, by retrieving relevant historical research trajectories in other topics to guide idea generation for a target topic. Extensive experiments show that CoI-Pro achieves average preference scores of 57.7%, 74.9%, and 58.9% against AI-Researcher, AI-Scientist-v2, and the CoI Agent, respectively, under three LLM judges on 50 temporally separated computer-science topics.

1 Introduction

Scientific idea generation is becoming a critical stage of automated research. Large language models (LLMs) already support literature search, hypothesis formulation, method implementation, experimentation, analysis, and manuscript preparation (Lu et al., 2024; Yamada et al., 2025; Tang et al., 2025). These capabilities can substantially reduce the effort required after a research direction has been chosen. Their value, however, still depends on the quality of that initial direction: faithful implementation and rigorous experimentation cannot compensate for a question that is incremental, weakly motivated, or unlikely to produce a meaningful advance. As automated systems assume broader responsibility for research execution, deciding *what should be investigated next* remains an upstream bottleneck.

Existing methods address this bottleneck by grounding idea generation in scientific literature. Retrieval-based systems select relevant papers, entities, or textual inspirations and use them to propose or refine candidate ideas (Wang et al., 2024a;b; Baek et al., 2025; Hu et al., 2025). More recent approaches organize multiple papers into structured representations of research development. Chain of Ideas arranges citation-linked papers into ordered trajectories so that a model can follow successive advances (Li et al., 2024), while graph-based methods preserve richer relations among papers, motivations, challenges, and methods (Lei et al., 2025; Gao et al., 2026; Li et al., 2026). These approaches construct idea chains or graphs, select a direction for idea generation explicitly or implicitly, and generate a concrete idea. The structure supplies evidence about the target topic's existing trajectory.

However, these methods do not explicitly retrieve successor innovations from other topics to guide direction selection. After reading a chain or graph, the generator must still select a direction from an

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

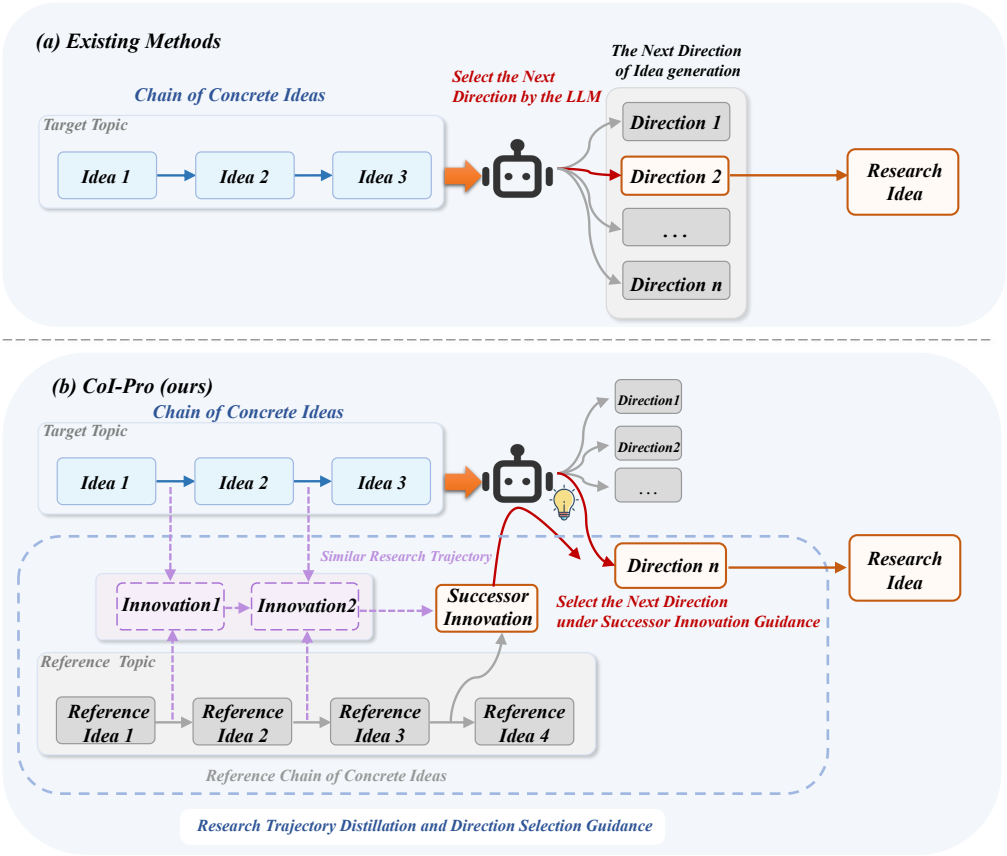

Figure 1: Comparison between existing methods and CoI-Pro. In existing methods, the next direction is determined by the model itself, leading to broad and weakly constrained direction selection. In CoI-Pro, a retrieved successor innovation guides the next direction; the direction is then grounded in the target chain to produce a concrete research idea. For ease of illustration, only the chain-based case for existing methods is shown.

open-ended space. This can leave effective next steps unexplored, especially when useful innovations have already emerged along similar trajectories in other topics. Research histories can provide evidence about both how a sequence of ideas progressed and which innovation followed it. Retrieving this evidence can guide the generator toward a concrete advance rather than a loosely connected extension of the target literature.

Our key insight is that research trajectories from other topics can guide the next step in idea generation. We distill each transition between concrete ideas into an *innovation move*, consisting of the limitation that prompted a change and the strategy used to address it. An ordered sequence of these moves forms a *research trajectory*, which captures successive innovations while omitting topic-specific technical details. When the target trajectory resembles a historical prefix, the innovation immediately following that prefix is its *successor innovation*. We retrieve this successor together with its historical prefix so that the guidance reflects the preceding sequence of ideas.

As illustrated in Figure 1, we propose **CoI-Pro**, a retrieval-augmented framework for scientific idea generation. Given a topic, CoI-Pro constructs a citation-grounded Chain of Ideas (CoI) that records the concrete development of the relevant literature. It distills the transitions into innovation moves and queries a *research trajectory memory* with their ordered trajectory. The memory pairs historical trajectory prefixes with their observed successor innovations. Retrieved successor innovations provide guidance for direction selection, which the generator combines with the target topic and concrete CoI to formulate a research proposal. The memory contains 1,512 validated research tra-

jectories expanded into 4,241 prefix–successor records. This design connects the circumstances under which a strategy was used historically to the concrete setting in which a new idea is proposed.

We evaluate CoI-Pro on 50 computer-science topics derived from papers published after 2024. The evaluation compares CoI-Pro with three automated-research systems—AI-Researcher (Tang et al., 2025), AI-Scientist-v2 (Yamada et al., 2025), and the CoI Agent (Li et al., 2024)—using pairwise judgments of novelty, significance, clarity, feasibility, and expected effectiveness. CoI-Pro obtains the highest point-estimate ranking among the four systems under DeepSeek-V4-Pro, GPT-5, and Claude-Opus-5. Under Claude-Opus-5, its preference scores against AI-Researcher, AI-Scientist-v2, and the CoI Agent are 59.3%, 74.7%, and 59.7%, respectively. A component study shows that innovation move distillation and retrieval are most effective when combined, while controlled comparisons across three random seeds show that semantically retrieved trajectories provide more useful guidance than randomly selected ones. Human evaluation further supports the main comparison and the reliability of the model-based evaluation protocol.

Our contributions are threefold:

- We distill research histories into research trajectories, representing reusable innovation experience as ordered moves that connect prior limitations to subsequent strategies.
- We introduce a research trajectory memory and propose CoI-Pro, which retrieves historically observed successor innovations following similar trajectories and grounds them in a target topic's concrete literature context.
- We validate CoI-Pro through comparisons with three strong automated-research systems, three independent LLM judges, human evaluation, component ablations, retrieval controls, and a case study tracing how a retrieved successor innovation informs a generated idea.

2 Related Work

Automated scientific discovery and idea generation. LLMs increasingly support complete research workflows. The AI Scientist automates idea formulation, implementation, experimentation, and manuscript preparation (Lu et al., 2024); AI-Scientist-v2 extends this workflow with agentic tree search (Yamada et al., 2025); and AI-Researcher targets autonomous innovation through literature synthesis, hypothesis formulation, experimental validation, and iterative refinement (Tang et al., 2025). Alongside these end-to-end systems, other work studies idea generation as a distinct stage. Large-scale human evaluation shows that LLMs can produce ideas judged novel by researchers, while also revealing limitations such as reduced diversity across repeated generations (Si et al., 2025). ResearchAgent refines candidate ideas with academic-graph evidence and reviewing agents (Baek et al., 2025), NOVA interleaves ideation with planned knowledge acquisition (Hu et al., 2025), and MoRI learns motivation-to-methodology reasoning for scientific ideation (Gu et al., 2026). CoI-Pro focuses on the upstream choice of a research direction and supplies that choice with guidance derived from observed scientific progressions.

Literature-grounded scientific ideation. Retrieval provides external evidence for generating ideas beyond a model's internal knowledge. SciMON retrieves inspirations from earlier papers and iteratively improves novelty (Wang et al., 2024a). SciPIP combines semantic, entity-based, and citation-based signals to retrieve relevant papers before proposal generation (Wang et al., 2024b). ResearchAgent connects a seed paper to publications and entities in an academic knowledge graph (Baek et al., 2025), while NOVA allows the model to plan which external knowledge to acquire during iterative ideation (Hu et al., 2025). These methods retrieve papers, entities, or textual inspirations that enrich the evidence available to the generator. CoI-Pro retrieves successor innovations together with their matched historical trajectory prefixes. This record supplies a candidate strategy together with the progression in which it was used, giving direction selection an empirical reference in the research history.

Structured representations of scientific development. Recent work organizes multiple related ideas to expose how research changes over time. Chain of Ideas constructs ordered citation chains and uses their progression to anticipate a subsequent idea (Li et al., 2024). MotivGraph-SoIQ models relations among research motivations and combines the resulting graph with Socratic interaction (Lei et al., 2025). Graphs of Research uses citation-evolution graphs as supervision for idea

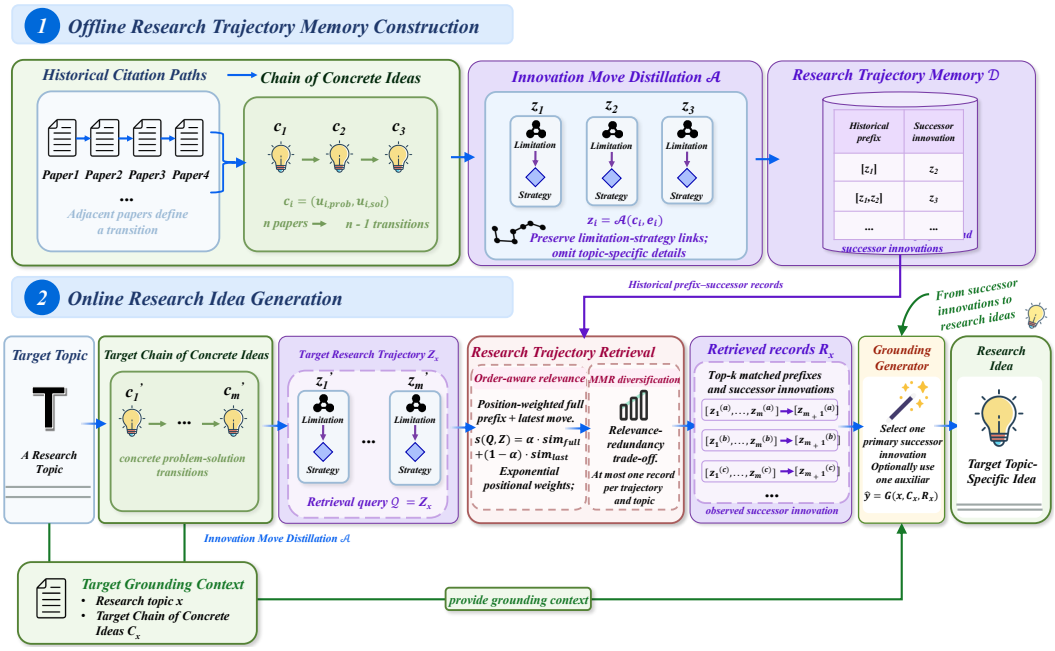

Figure 2: Overview of CoI-Pro. Offline, citation-grounded concrete idea chains are distilled into research trajectories and expanded into historical prefix–successor records. Online, the target trajectory retrieves relevant records through position-weighted similarity and MMR diversification. The generator selects one primary successor innovation and optionally one auxiliary innovation, then combines this guidance with the target topic and concrete chain to generate a research idea.

generation, and Graph2Idea retrieves graph-structured contexts containing richer relations among prior works (Gao et al., 2026; Li et al., 2026). Chains make sequential dependencies explicit, while graphs preserve branches, convergence, and broader relational context. These representations are primarily supplied to a model as context or supervision, after which the model determines the next direction. CoI-Pro distills concrete research transitions into innovation moves and indexes their ordered trajectories. Matching a target progression to historical prefixes makes their observed successor innovations directly retrievable, while the concrete target literature supplies the context for instantiating the selected strategy.

3 Method

3.1 Overview

CoI-Pro guides direction selection with successor innovations retrieved from similar historical trajectories. Given a topic x, it (1) constructs a citation-grounded Chain of Ideas (CoI) C_x, (2) distills its concrete transitions into a research trajectory Z_x, (3) retrieves historical prefixes and their observed successor innovations $\mathcal{R}_x$ from a research trajectory memory $\mathcal{D}$, and (4) grounds the retrieved successor innovations in the target's concrete context to generate an idea $\hat{y}$. Memory construction is offline; all four stages for a target topic are performed online.

Figure 2 summarizes the offline memory construction and online generation procedure.

3.2 From Chains of Concrete Ideas to Research Trajectories

Citation-grounded CoI. Following Chain of Ideas (Li et al., 2024), we organize papers relevant to x into an ordered citation path $P_x = (p_1, \ldots, p_n)$. For every adjacent pair, we extract a concrete transition $c_i = (u_{i,\text{prob}}, u_{i,\text{sol}})$: $u_{i,\text{prob}}$ identifies the limitation of p_i that motivates the later work,

and $u_{i,\text{sol}}$ describes how p_{i+1} addresses it. The concrete CoI is $C_x = (c_1, \ldots, c_m)$, where $m = n - 1$.

Although the surrounding literature is naturally a graph, we use ordered chains as retrieval units. Graphs preserve richer relations, but their size, branching pattern, and local topology vary substantially across research topics. These structural differences can dominate comparison and obscure similarity in the underlying progression. A chain retains sufficient trend information through an ordered sequence of limitations and responses, while providing a simpler and more consistent structure that is easier to compare and reuse. Our choice concerns the retrieval unit rather than the nature of scientific development; multiple chains may be extracted from a larger literature graph.

Innovation move distillation. Concrete transitions contain paper names, models, datasets, and implementation details that can dominate semantic matching. We distill each transition into an *innovation move*, consisting of the underlying limitation and the strategy used to address it. An LLM-based operator $\mathcal{A}$ produces

$$z_i = \mathcal{A}(c_i, e_i) = (a_{i,\text{prob}}, a_{i,\text{sol}}), \tag{1}$$

where e_i is the supporting paper and citation evidence. The first component, $a_{i,\text{prob}}$, describes the *limitation*: its cause and consequence. The second, $a_{i,\text{sol}}$, describes the *strategy*: the functional change used to address that limitation. The ordered moves form the research trajectory $Z_x = (z_1, \ldots, z_m)$. Each move must remain faithful to its concrete transition and preserve the connection between limitation and strategy while omitting topic-specific implementation details. An innovation move records a strategy observed in the research history; its usefulness for a new topic depends on that topic's context.

3.3 Research Trajectory Memory

We construct $\mathcal{D}$ from historical citation paths using the same concrete-transition extraction and move-distillation procedure. Each research trajectory is expanded into records pairing a prefix with its successor innovation, the innovation move observed immediately after that prefix. For a historical trajectory $Z^{(j)} = (z_1^{(j)}, \ldots, z_{L_j}^{(j)})$,

$$\mathcal{D} = \bigcup_j \left\{ \left(Z_{1:t}^{(j)}, z_{t+1}^{(j)} \right) \right\}_{t=1}^{L_j - 1}. \tag{2}$$

Thus, (z_1, z_2, z_3) yields the records $((z_1), z_2)$ and $((z_1, z_2), z_3)$. Each prefix provides the historical context in which the associated successor innovation occurred. We retain this context so that retrieval selects a strategy based on the progression preceding it. Corpus construction, filtering, temporal boundaries, and memory statistics are reported in the experimental setup.

3.4 Retrieval over Research Trajectories

The target research trajectory Z_x acts as the query. We encode each move's limitation and strategy with a text encoder ϕ and normalize its vector, $\mathbf{e}_i = \phi(z_i)$. For a prefix $Z_{1:t}$, we assign normalized exponential positional weights and represent both its ordered progression and its latest move:

$$\omega_i^{(t)} = \frac{\exp(i/t)}{\sum_{j=1}^{t} \exp(j/t)}, \qquad \mathbf{h}_{\text{full}}(Z_{1:t}) = \text{norm}\left(\sum_{i=1}^{t} \omega_i^{(t)} \mathbf{e}_i \right), \qquad \mathbf{h}_{\text{last}}(Z_{1:t}) = \mathbf{e}_t. \tag{3}$$

The increasing weights make the prefix representation sensitive to move order while smoothly emphasizing more recent developments. Normalizing positions by prefix length keeps the maximum last-to-first weight ratio below e, so earlier moves continue to contribute. The separate last-move representation explicitly captures the current research frontier. For query Q and candidate prefix Z, relevance combines the two views:

$$s(Q, Z) = \alpha\, \mathbf{h}_{\text{full}}(Q)^\top \mathbf{h}_{\text{full}}(Z) + (1 - \alpha)\, \mathbf{h}_{\text{last}}(Q)^\top \mathbf{h}_{\text{last}}(Z), \tag{4}$$

where α balances global progression and the current research state.

We then use maximal marginal relevance to reduce redundant guidance. Given selected records S, the next record is

$$r^* = \arg\max_{r \notin S} \left[\lambda s(Q, r) - (1 - \lambda) \max_{r' \in S} \mathbf{h}_{\text{full}}(r)^\top \mathbf{h}_{\text{full}}(r') \right]. \tag{5}$$

We retain at most one record from each historical trajectory and topic. The final set $\mathcal{R}_x$ contains the top-k matched prefixes and their observed successor innovations.

3.5 Grounding Retrieved Successor Innovations

The retrieved successor innovations guide direction selection for idea generation in the target topic. The generator combines them with the concrete topic context:

$$\hat{y} = G(x, C_x, \mathcal{R}_x), \tag{6}$$

where G is an instruction-following generator. It selects one retrieved successor innovation as the primary guide, may use at most one additional move when useful, and instantiates the selected limitation and strategy within the technical path described by C_x. The output is a structured proposal containing a title, motivation, problem statement, methodology, contributions, experimental design, and limitations. In this way, retrieved successor innovations guide *which direction* to pursue, while the topic and concrete CoI determine *how* to realize it.

4 Experiments

We evaluate CoI-Pro through three questions: **(RQ1)** Does it generate stronger ideas than existing automated-research systems? **(RQ2)** How do innovation move distillation and successor-innovation retrieval contribute to its performance? **(RQ3)** Does semantic retrieval provide more useful guidance than adding an arbitrary historical trajectory?

4.1 Experimental Setup

Topics and temporal separation. We collect papers using 40 Semantic Scholar queries and build the research trajectory memory only from papers published no later than 2024. This yields 1,541 concrete chains, of which 1,512 yield validated research trajectories and 4,241 prefix–successor records. We evaluate on 50 topics derived from computer-science papers published after 2024. The source papers span NeurIPS (9), CVPR (6), AAAI (4), ICCV (4), and venues including ACL, EMNLP, ECCV, ICML, Interspeech, ICLR, KDD, ICRA, and RSS. For each target paper, `gpt-4o-mini` extracts one self-contained description of its research topic, without the paper's specific technical approach, using the prompt in Appendix C.3.1. This extracted description is the complete topic supplied identically to all systems; the remaining target-paper content is excluded from chain construction, retrieval, and generation prompts. We use the fixed `gpt-4o-2024-11-20` checkpoint for innovation move distillation and idea generation, so both the accessible literature and generator precede the evaluation targets.

Baselines and implementation. We compare against AI-Researcher (Tang et al., 2025), AI-Scientist-v2 (Yamada et al., 2025), and the CoI Agent (Li et al., 2024). All methods use `gpt-4o-2024-11-20` for generation. CoI construction follows the best-performing hyperparameter configuration reported by the CoI Agent, without tuning on our evaluation topics. Because the systems expose different native schemas, differences in presentation could otherwise confound pairwise evaluation. We therefore losslessly normalize every proposal into the same ordered idea format before evaluation; the formatter only reorganizes source content under common headings and cannot introduce new technical claims.

Evaluation protocol. We adopt Idea Arena, introduced by the CoI Agent (Li et al., 2024). For each topic, every method pair is judged twice with reversed presentation order along novelty, significance, clarity, feasibility, and expected effectiveness. We run the complete four-system arena independently with DeepSeek-V4-Pro, GPT-5, and Claude-Opus-5. For wins W, ties T, and losses L, preference is

$$\text{Pref} = \frac{W + 0.5T}{W + T + L}. \tag{7}$$

Table 1: Main Idea Arena results. Panel (a) reports $W/T/L$ from the perspective of CoI-Pro; Pref. pools all five dimensions. Panel (b) reports four-system Elo ratings.

(a) CoI-Pro against each baseline

Judge	Opponent	Nov.	Sig.	Cla.	Feas.	Eff.	Pref.
DeepSeek-V4-Pro	AI-Researcher	51/1/48	40/0/60	58/4/38	68/0/32	55/0/45	54.9
	AI-Scientist-v2	67/2/31	72/4/24	77/1/22	74/0/26	77/0/23	74.1
	CoI Agent	57/0/43	57/0/43	48/2/50	73/0/27	67/0/33	60.6
GPT-5	AI-Researcher	65/3/32	53/3/44	58/3/39	58/1/41	55/0/45	58.8
	AI-Scientist-v2	68/1/31	78/1/21	85/3/12	65/2/33	80/0/20	75.9
	CoI Agent	61/0/39	48/0/52	46/3/51	67/2/31	57/0/43	56.3
Claude-Opus-5	AI-Researcher	58/14/28	51/10/39	61/4/35	48/8/44	59/3/38	59.3
	AI-Scientist-v2	61/15/24	72/13/15	79/6/15	59/11/30	79/2/19	74.7
	CoI Agent	48/21/31	48/10/42	50/5/45	73/5/22	56/6/38	59.7

(b) Four-system Elo ranking

Method	DeepSeek-V4-Pro	GPT-5	Claude-Opus-5	Mean rank
CoI-Pro	1073	1076	1081	1.0
AI-Researcher	1027	1024	1011	2.3
CoI Agent	1006	1013	1029	2.7
AI-Scientist-v2	895	887	879	4.0

We also report Elo ratings fitted from all pairwise outcomes. For the main comparison with existing systems, 95% confidence intervals are percentile bootstrap intervals obtained by resampling the 50 topics with replacement 10,000 times. A topic, together with all of its dimensions and presentation orders, is retained as one resampling unit.

4.2 Comparison with Existing Systems

Table 1 gives both the dimension-level outcomes against each baseline and the four-system Elo ranking. Each $W/T/L$ entry contains 100 decisions from CoI-Pro's perspective: 50 topics in two presentation orders. CoI-Pro has a higher point-estimate preference than every baseline under every judge and obtains the highest point-estimate Elo in all three arenas. Its opponent-specific preference is 54.9–59.3% against AI-Researcher, 74.1–75.9% against AI-Scientist-v2, and 56.3–60.6% against the CoI Agent. The most consistent gains occur in feasibility and expected effectiveness, supporting our premise that an observed successor innovation constrains the next direction more effectively than literature context alone. AI-Researcher remains the closest competitor, particularly in significance.

The topic-level 95% confidence intervals for preference against AI-Researcher, AI-Scientist-v2, and the CoI Agent are, respectively, $[47.3, 62.3]$, $[66.2, 81.5]$, and $[51.5, 69.5]$ under DeepSeek-V4-Pro; $[52.0, 65.5]$, $[70.2, 81.4]$, and $[47.9, 64.6]$ under GPT-5; and $[53.3, 65.2]$, $[68.8, 80.2]$, and $[53.7, 65.8]$ under Claude-Opus-5.

4.3 Component Analysis

We form a 2×2 design over innovation move distillation (A) and retrieval (R). Every variant receives the topic and complete concrete target CoI. $A1R0$ additionally receives its distilled research trajectory; $A0R1$ receives five concrete prefix–successor records; and the full $A1R1$ retrieves five research-trajectory prefix–successor records using the target research trajectory as its query. All variants use one prompt template, output schema, model, and decoding configuration; disabled evidence fields are left empty, while $A0R1$ replaces distilled external records with concrete ones.

Table 2 shows that $A1R1$ is preferred to each alternative overall. Distillation alone supplies a second description of the known history but no evidence about what follows; concrete retrieval introduces source-specific successors whose technical details may not fit the target. These outcomes are consistent with complementary roles: distillation supports comparison of research trajectories, while retrieval supplies experience about a successor innovation. The complete concrete CoI remains available for grounding in all variants. Clarity remains close to indifference after format normalization, whereas the full method's largest gains appear in feasibility and expected effectiveness.

Table 2: Component analysis under GPT-5. Entries are $W/T/L$ for $A1R1$ against each alternative; Pref. pools the five dimensions.

Contrast	Nov.	Sig.	Cla.	Feas.	Eff.	Pref.
$A1R1$ vs. $A0R0$	63/1/36	59/1/40	49/0/51	62/1/37	60/0/40	58.9
$A1R1$ vs. $A1R0$	55/5/40	60/2/38	52/0/48	81/0/19	66/0/34	63.5
$A1R1$ vs. $A0R1$	60/4/36	60/3/37	46/0/54	63/3/34	63/0/37	59.4

Table 3: Semantic retrieval versus forced rank-1 random guidance under GPT-5. Entries are $W/T/L$ from the perspective of semantic retrieval.

Seed	Nov.	Sig.	Cla.	Feas.	Eff.	Pref.
21	44/2/54	58/0/42	61/4/35	63/1/36	62/0/38	58.3
42	44/5/51	52/2/46	58/1/41	60/1/39	57/1/42	55.2
123	40/3/57	51/1/48	63/3/34	65/3/32	61/1/38	57.1
Mean	42.7/3.3/54.0	53.7/1.0/45.3	60.7/2.7/36.7	62.7/1.7/35.7	60.0/0.7/39.3	56.9

4.4 Semantic Retrieval versus Random Guidance

To distinguish targeted retrieval from the generic benefit of extra context, we compare the top-ranked semantic trajectory with a randomly selected trajectory designated as the primary direction. The generation procedure and retrieval budget are otherwise identical, and random selection is repeated with seeds 21, 42, and 123. Semantic retrieval is preferred for every seed (Table 3), averaging 56.9%. Random guidance receives higher novelty preferences, whereas semantic guidance is consistently stronger in significance, clarity, feasibility, and expected effectiveness. A randomly imposed trajectory is usually farther from the target's established progression and therefore encourages unusual combinations or larger conceptual jumps. Consequently, a random trajectory can receive a higher novelty score precisely because it induces a more surprising departure, even when the resulting mechanism is weakly motivated or technically mismatched. Semantic retrieval deliberately constrains this search space to historically compatible progressions; it may sacrifice some surface-level novelty, but produces directions that are more coherent with the target trajectory and more likely to be meaningful and realizable. The overall gain therefore comes from retrieving a progression that fits the target trajectory, rather than merely adding a surprising direction.

4.5 Human Validation

Six computer-science researchers independently evaluate the same 150 CoI-Pro–baseline pairs. Each pair is assigned to four evaluators, with method identities hidden and order randomized, using the same five dimensions as Idea Arena. We aggregate the four judgments into one label per pair and dimension. Table 4 reports both human preference and exact label agreement. Human evaluators prefer CoI-Pro to all three baselines. Their mean pairwise agreement is 71.3%, while the three Idea Arena judges agree with human consensus by 69.7–71.4%. The 70.4% mean model–human agreement is close to human–human agreement, indicating that the model-based ranking tracks human preferences at approximately the level at which independent researchers agree with one another.

4.6 Case Study

Appendix A traces one complete generation for efficient reinforcement learning of long-chain-of-thought reasoning. The target CoI progresses from constrained policy updates to a clipped surrogate and then to critic-free, group-relative feedback. CoI-Pro retrieves a neural-architecture-search trajectory with an analogous progression from expensive independent search to shared optimization and finer-grained credit assignment. Its observed successor innovation organizes alternatives hierarchically and optimizes sampled local choices directly. Grounding this pattern in the target topic yields BranchGRPO, which constructs budget-matched shared-prefix rollout trees and assigns sibling-relative advantages at the reasoning decisions where candidate trajectories diverge. Both displayed CoIs are strictly chronological and their adjacent citation links are verified.

Table 4: Human validation. Panel (a) reports consensus judgments from CoI-Pro's perspective. Panel (b) reports exact win/tie/loss agreement.

(a) Human preference

Opponent	$W/T/L$	Pref.
AI-Researcher	140/11/99	58.2
AI-Scientist-v2	183/11/56	75.4
CoI Agent	160/9/81	65.8

(b) Evaluation agreement (%)

Comparison	Agreement
Human–Human	71.3
DeepSeek-V4-Pro–Human	69.7
GPT-5–Human	71.4
Claude-Opus-5–Human	70.2
Idea Arena mean	70.4

5 Conclusion

We presented CoI-Pro, a retrieval-augmented framework that guides scientific idea generation with successor innovations retrieved from historical trajectories. CoI-Pro constructs a citation-grounded Chain of Ideas, distills its transitions into innovation moves, retrieves historical prefixes with similar progressions and their observed successor innovations, and grounds the selected strategies in the target topic's concrete technical context. Across 50 temporally separated computer-science topics, CoI-Pro obtains the highest point-estimate ranking among four systems under three independent LLM judges. Component and retrieval-control experiments support the complementary roles of move distillation and retrieval, and show that semantically matched trajectories provide more useful guidance than randomly selected ones. Human evaluation further supports both the main comparison and the reliability of Idea Arena.

Our evaluation measures proposed ideas before their full experimental execution, and performance depends on the coverage and quality of the citation chains and research trajectory memory. Future work will validate generated ideas in closed-loop research workflows and extend the retrieval unit from ordered trajectories to graph-structured research histories. A key challenge is to retain useful branching, convergence, and dependency information without allowing differences in graph topology to obscure the semantic progression. Distilling and aligning innovation moves within graphs may enable CoI-Pro to reuse experience from richer representations of scientific development.

AI use statement

Generative AI tools were used to polish the manuscript, help identify relevant literature, assist with the implementation of portions of the codebase, and draft initial versions of some sections. They were also used to extract research trajectories from collected papers when constructing the research trajectory memory; this is the use reported under synthetic dataset generation in the submission form. All AI-assisted outputs were reviewed, verified, and revised by the authors, who take full responsibility for the final content, analyses, claims, and conclusions.

Ethics statement

The human evaluation involved six computer-science researchers whose expertise spans multiple research directions. We report only aggregate judgments and do not disclose personally identifying information.

Reproducibility statement

We provide the complete CoI-Pro algorithm, model and decoding configurations, retrieval hyperparameters, generation and evaluation prompts, construction and filtering criteria, temporal boundaries, statistical procedures, and full pairwise results in the appendices. These specifications document the chain construction, retrieval, generation, and evaluation procedures used in the reported experiments.

REFERENCES

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. ResearchAgent: Iterative research idea generation over scientific literature with large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6709–6738. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.342.

Xuanyi Dong and Yi Yang. Searching for a robust neural architecture in four GPU hours. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1761–1770, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Dong_Searching_for_a_Robust_Neural_Architecture_in_Four_GPU_Hours_CVPR_2019_paper.html.

Songyang Gao, Yinghui Xia, Siyi Liu, and Hui Xiong. Graphs of research: Citation evolution graphs as supervision for research idea generation. *arXiv preprint arXiv:2605.14790*, 2026.

Chenyang Gu, Jiahao Cheng, Meicong Zhang, Pujun Zheng, Jinquan Zheng, and Guoxiu He. MoRI: Learning motivation-grounded reasoning for scientific ideation in large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 34838–34869. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1609.

Xiang Hu, Hongyu Fu, Jinge Wang, Yifeng Wang, Zhikun Li, Renjun Xu, Yu Lu, Yaochu Jin, Lili Pan, and Zhenzhong Lan. NOVA: An iterative planning framework for enhancing scientific innovation with large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 21330–21359. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.1099.

Xinping Lei, Tong Zhou, Yubo Chen, Kang Liu, and Jun Zhao. MotivGraph-SoIQ: Integrating motivational knowledge graphs and socratic dialogue for enhanced LLM ideation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 2913–2933, 2025.

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

Xu Li, Hanzhe Tu, and Xun Han. Graph2Idea: Retrieval-augmented scientific idea generation with graph-structured contexts. *arXiv preprint arXiv:2606.09105*, 2026.

Chenxi Liu, Barret Zoph, Maxim Neumann, Jonathon Shlens, Wei Hua, Li-Jia Li, Li Fei-Fei, Alan Yuille, Jonathan Huang, and Kevin Murphy. Progressive neural architecture search. In *European Conference on Computer Vision*, 2018. URL https://arxiv.org/abs/1712.00559.

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

Hieu Pham, Melody Y. Guan, Barret Zoph, Quoc V. Le, and Jeff Dean. Efficient neural architecture search via parameter sharing. In *Proceedings of the 35th International Conference on Machine Learning*, 2018. URL https://arxiv.org/abs/1802.03268.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015. URL https://arxiv.org/abs/1502.05477.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models, 2024.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can LLMs generate novel research ideas? a large-scale human study with 100+ NLP researchers. In *International Conference on Learning Representations*, 2025. arXiv:2409.04109.

Jiabin Tang, Lianghao Xia, Zhonghang Li, and Chao Huang. AI-researcher: Autonomous scientific innovation. *arXiv preprint arXiv:2505.18705*, 2025.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. SciMON: Scientific inspiration machines optimized for novelty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 279–299. Association for Computational Linguistics, 2024a. doi: 10.18653/v1/2024.acl-long.18.

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

Sirui Xie, Hehui Zheng, Chunxiao Liu, and Liang Lin. SNAS: Stochastic neural architecture search. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1812.09926.

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

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, et al. DAPO: An open-source LLM reinforcement learning system at scale. In *Advances in Neural Information Processing Systems*, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/a4277440d50f1f15d2cb4c14f7e0c0d2-Abstract-Conference.html.

APPENDIX

Contents

Section	Contents	Page
A	Case Study: From a Successor Innovation to a Concrete Idea	12
B	Full CoI-Pro Algorithm	14
C	Implementation Details and Prompts	14
D	Idea Arena and Statistical Details	21
E	Complete Main-Comparison Results	22
F	Complete Ablation Round-Robin Results	23

A CASE STUDY: FROM A SUCCESSOR INNOVATION TO A CONCRETE IDEA

We expand the generation trace summarized in Section 4.6. The held-out target is *DAPO: An Open-Source LLM Reinforcement Learning System at Scale* (Yu et al., 2025), published at NeurIPS 2025. CoI-Pro receives only the topic: *improve the stability and efficiency of reinforcement learning for long-chain-of-thought reasoning in large language models*. The target paper and its contents are excluded from the CoI, retrieval memory, and generator context. Table 5 contains only work available by 2024, ordered by first public date, and every successor cites its predecessor. Table 6 gives the rank-1 retrieved prefix and its observed successor innovation. Matching blue and green cells denote the two analogous transitions; violet denotes the additional historical transition transferred to the target topic.

Table 5: Strictly chronological target-topic CoI for efficient reinforcement learning of long reasoning traces. Bold text gives each innovation move's limitation and strategy; roman text gives their concrete instantiations.

Concrete ideas in the target topic	Innovation moves and concrete transitions
Trust Region Policy Optimization (Schulman et al., 2015). A trust-region constraint stabilizes policy learning by limiting the size of each update.	**Limitation: Stable learning depends on an expensive constrained update.** Concrete problem: TRPO repeatedly solves a constrained second-order trust-region subproblem. **Innovation strategy: Replace the explicit constraint with a lightweight conservative surrogate.** Concrete solution: PPO clips the likelihood-ratio surrogate to discourage excessively large policy changes.
Proximal Policy Optimization Algorithms (Schulman et al., 2017). A clipped first-order objective retains conservative updates while simplifying optimization.	**Limitation: Auxiliary estimation and global feedback make structured credit assignment costly and noisy.** Concrete problem: PPO-style reasoning optimization relies on a separate value model and whole-response advantages. **Innovation strategy: Compare related candidates directly to obtain relative, attributable update signals.** Concrete solution: GRPO normalizes rewards within response groups and removes the learned critic.
DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models (Shao et al., 2024). Group-relative advantages support critic-free reinforcement learning for mathematical reasoning.	

The target trajectory first simplifies a costly but stable update and then replaces auxiliary estimation with relative feedback. Retrieval returns the two-edge neural-architecture-search prefix in Table 6 at rank 1 (score 0.837). Its transitions have the same ordering: reuse structure to avoid repeated global optimization, then attach finer feedback to interdependent local decisions. The historical research trajectory contains one further successor innovation, shown in violet.

Table 6: Retrieved CoI prefix and its observed successor innovation. Formatting follows Table 5; blue and green mark analogous transitions, and violet marks the successor innovation used for generation.

Concrete ideas in the retrieved trajectory	Innovation moves and concrete transitions
Progressive Neural Architecture Search (Liu et al., 2018). Candidate cells are explored progressively, but each evaluated architecture still requires substantial separate optimization.	**Limitation: Re-optimizing related alternatives independently wastes computation.** Concrete problem: Progressive search repeatedly trains and evaluates many candidate architectures. **Innovation strategy: Reuse one shared parameterization across alternatives.** Concrete solution: ENAS represents candidates as subgraphs of one shared network and reuses their weights.
Efficient Neural Architecture Search via Parameter Sharing (Pham et al., 2018). Candidate subgraphs share weights, sharply reducing the cost of architecture evaluation.	**Limitation: Delayed global rewards give weak credit to interdependent local choices.** Concrete problem: A reinforcement-learning controller receives an architecture-level reward only after sampling and evaluating a child network. **Innovation strategy: Jointly optimize structured alternatives with locally decomposable feedback.** Concrete solution: SNAS learns operation weights and a stochastic architecture distribution together, exposing local credit through differentiable losses.
SNAS: Stochastic Neural Architecture Search (Xie et al., 2019). Stochastic, differentiable search jointly learns architectural choices and network parameters with local credit assignment.	**Limitation: Optimizing all competing choices or relying on long-horizon search remains expensive and unstable.** Concrete problem: Updating a complete supergraph mixes competing operations, whereas reinforcement-learning search receives delayed feedback. **Innovation strategy: Sample one structured subgraph differentiably and send immediate feedback to its selected choices.** Concrete solution: GDAS samples a subgraph from an architecture DAG and optimizes its sampler directly with validation loss.
Searching for a Robust Neural Architecture in Four GPU Hours (Dong & Yang, 2019). A differentiable architecture sampler selects one subgraph per iteration and learns the search rule end to end.	

From a retrieved successor innovation to a target-specific mechanism. The first matched transition maps architecture weight sharing to a lightweight policy-update mechanism that avoids repeatedly solving the full constrained optimization problem. The second maps local credit for architectural operations to critic-free relative credit among reasoning candidates. The retrieved successor innovation then contributes a new operation: organize alternatives as a hierarchy, sample local branches while reusing shared computation, and attach feedback to the choices that distinguish successful from unsuccessful candidates. Grounding maps the architecture DAG to a shared-prefix reasoning tree, sampled subgraphs to sibling continuations, shared network states to cached reasoning prefixes, and validation feedback to terminal verifier rewards.

Generated proposal: BranchGRPO. Motivation. Group-relative optimization compares complete responses to the same prompt, but long reasoning traces often share many correct steps and diverge at only a few decisions. One response-level advantage consequently rewards or penalizes an entire trace, obscuring the branch that caused success or failure. Independently sampling more complete responses also recomputes similar prefixes without resolving this attribution problem.

Method. BranchGRPO redistributes the standard group rollout budget into a shared-prefix tree. At reasoning-step boundaries, several continuations are forked from the same cached prefix; later boundaries may branch again while the total generated-token budget remains fixed. Terminal verifier rewards are propagated to every branch. For an edge leaving prefix v, BranchGRPO averages the rewards of its descendant leaves and normalizes this value against the edge's siblings, producing a local relative advantage for the tokens generated on that edge. Tokens before the first fork retain the global group-relative advantage, and a support-aware interpolation falls back toward that global estimate when a branch has too few descendants. The resulting advantages use the usual clipped

policy objective. BranchGRPO therefore introduces decision-specific credit without a learned critic or extra rollout tokens, while prefix-cache reuse reduces redundant generation.

B Full CoI-Pro Algorithm

CoI-Pro separates memory construction from target-time generation. The offline procedure converts historical citation paths into a searchable research trajectory memory; the online procedure uses a target trajectory to retrieve observed successor innovations and ground their strategies in a new proposal.

Offline memory construction. Given a collection of historical citation paths $\mathcal{P}$, we perform the following steps:

1. For every adjacent paper pair in each path, extract a concrete transition $c_i = (u_{i,\text{prob}}, u_{i,\text{sol}})$ from the papers' ideas and citation context.
2. Apply the move-distillation operator to obtain $z_i = (a_{i,\text{prob}}, a_{i,\text{sol}})$, retaining the causal methodological change while removing paper names, datasets, application entities, and implementation-specific objects.
3. For every research trajectory $Z = (z_1, \ldots, z_L)$, emit the records $(Z_{1:t}, z_{t+1})$ for $t = 1, \ldots, L-1$.
4. Encode each transition as the text `Problem: {limitation}` followed by `Solution: {innovation strategy}`. Store the normalized exponentially position-weighted sum of the transition embeddings as the full-prefix vector and the final move embedding as the last-edge vector.

Online idea generation. For a target topic x, the online procedure is:

1. Construct a citation-grounded concrete CoI C_x and extract its research trajectory Z_x.
2. Encode Z_x into full-prefix and last-edge vectors. Score every eligible memory record using Equation 4.
3. Retain the 100 highest-relevance candidates, apply the MMR rule in Equation 5, and select $k = 5$ records. Records containing the query paper are excluded, and at most one record is returned from each historical trajectory and topic.
4. Supply the target topic, the complete concrete target trajectory, and the five retrieved prefix–successor records to the grounding model.
5. The grounding model selects one successor innovation as the primary guide and may use at most one additional direction. It instantiates the selected innovation strategy using only information available in the target topic and CoI, and returns one structured research proposal.

C Implementation Details and Prompts

C.1 Model and Decoding Configuration

Table 7 records the configurations used to produce the reported results. Target innovation move extraction, historical-library move distillation, and final idea generation all use `gpt-4o-2024-11-20`. All generations are parsed against their required schemas before being accepted.

Table 7: Model and decoding configuration. A dash indicates a deterministic embedding or retrieval operation.

Stage	Model	Temp.	Max tokens	Retries	Timeout
Target move distillation	`gpt-4o-2024-11-20`	0.3	4,000	3	API default
Library move distillation	`gpt-4o-2024-11-20`	0.2	2,600	4	API default
Trajectory embedding	`BAAI/bge-base-en-v1.5`	–	512 per edge	–	–
Final idea generation	`gpt-4o-2024-11-20`	0.6	8,000	5	300 s
Idea Arena judging	DeepSeek-V4-Pro / GPT-5 / Claude-Opus-5	0.7	16,000	5	180 s

The embedding model uses revision `a5beb1e3e68b9ab74eb54cfd186867f64f240e1a`, produces 768-dimensional vectors, and is applied independently to each innovation move with a maximum sequence length of 512 tokens. Model outputs containing a `</think>` delimiter are reduced to the text following that delimiter before parsing.

C.2 Retrieval Configuration

For every transition, the indexed text is

$$\texttt{Problem:}\quad a_{\text{prob}}\quad \texttt{Solution:}\quad a_{\text{sol}}.$$

The encoder produces an ℓ_2-normalized vector for each transition. Full-prefix vectors use the normalized exponential positional weights $\omega_i^{(t)} = \exp(i/t)/\sum_{j=1}^{t}\exp(j/t)$ from Equation 3; last-edge vectors are the normalized final transition vectors. We use full-prefix weight $\alpha = 0.4$, last-edge weight $1-\alpha = 0.6$, MMR coefficient $\lambda = 0.8$, an initial relevance pool of 100 records, and $k = 5$. Successors later than 2024 and records containing the query paper are filtered before selection. Selection admits at most one record from a historical trajectory and at most one record for a repeated historical topic.

C.3 Prompt Specifications

The following blocks report the prompts used for target-paper topic extraction, innovation move distillation, final idea generation, output normalization, and Idea Arena evaluation. Braced expressions denote per-example fields inserted at runtime. The original prompt text and serialization names are retained for reproducibility: `abstract_problem` and `abstract_solution` encode a move's limitation and innovation strategy, respectively. The corresponding uppercase output tags follow the same convention.

C.3.1 Target-paper topic extraction

We apply the following prompt once to each held-out target paper using `gpt-4o-mini`. The returned topic is used as the input to each idea-generation system; no other target-paper text is provided.

System prompt.

```
You are a scientific-topic extractor. Given the content of one research paper,
identify only the research topic investigated by the paper. Do not extract the
paper's proposed method or use its solution as part of the topic. Do not propose
a new idea, evaluate the paper, or use external knowledge.

Write one concise, self-contained phrase or sentence that states the research
problem, task, phenomenon, or question being studied. Describe what the paper
investigates, not how it addresses the problem. Omit method names, architectures,
modules, algorithms, objectives, training procedures, implementation details,
datasets, baselines, metrics, numerical results, ablation findings, authors,
affiliations, venue information, citations, and generic performance claims. If
the paper title is centered on a named method, recover the underlying research
topic instead of repeating that method name. Do not truncate the output.

Return only one valid JSON object with exactly one string field:
{"topic": "<one concise, self-contained research topic>"}
```

User template.

```
[PAPER TITLE]
{target-paper title}

[PAPER ABSTRACT]
{target-paper abstract}

[AVAILABLE PAPER CONTENT]
{target-paper text or extracted sections}

Extract only the paper's research topic without its specific technical solution.
```

C.3.2 TARGET INNOVATION MOVE DISTILLATION

System prompt.

```
You are an expert scientific methodology analyst. Your task is to perform a deep
↪  "retroactive extraction" of the methodological evolution between two consecutive
↪  research papers within a specific topic.

**CORE LOGIC (CONCRETE TO ABSTRACT CoT):**
1. **Concrete Phase:** Identify the concrete problem Paper X faced (as claimed by Paper
↪  X+1) and the concrete solution Paper X+1 proposed.
2. **Abstract Phase:** Elevate these concrete specifics into domain-agnostic, abstract
↪  methodological concepts.

**RULES:**
1. **For Concrete Keys:** Keep the domain-specific jargon, algorithms, acronyms, and
↪  application scenarios.
2. **For Abstract Keys (STRICT RED-LINE):** Strip away ALL specific algorithms, dataset
↪  names, domain jargon, and explicit tool names. Describe the mechanisms, structural
↪  paradigms, computational boundaries, and underlying logic.
3. **ANTI-PLAGIARISM WARNING:** You MUST NOT copy phrases from the examples below (e.g.,
↪  do not use "cross-paradigm computing", "quantum", "holistic integration", "open-loop"
↪  unless genuinely applicable). Generate highly unique abstract concepts tailored
↪  specifically to the input text.
4. **Output Format:** You must output ONLY a valid JSON object containing exactly four
↪  keys: `concrete_problem`, `concrete_solution`, `abstract_problem`, and
↪  `abstract_solution`.

**EXAMPLE 1 (Hardware & Interaction):**
Topic: Educational Robotics
Paper X: Used simple rule-based robots to teach basic counting.
Paper X+1: Introduced AI-driven adaptive robots to dynamically adjust math difficulty
↪  based on student facial expressions.

Output 1:
{
  "concrete_problem": "The original approach relies on simple rule-based robots for
  ↪  teaching counting, which lack the ability to read student facial expressions and
  ↪  adjust difficulty, leading to a static curriculum that cannot respond to real-time
  ↪  student frustration.",
  "concrete_solution": "The target paper introduces AI-driven adaptive robots that capture
  ↪  real-time facial expressions to dynamically modulate the mathematical difficulty,
  ↪  creating a personalized learning loop.",
  "abstract_problem": "The structural bottleneck of an open-loop, deterministic execution
  ↪  framework. It relies on a unidirectional delivery mechanism that is context-blind,
  ↪  forcing a static execution path regardless of external variance or user state.",
  "abstract_solution": "A transition to a context-aware, dynamically adaptive framework.
  ↪  It integrates continuous, multidirectional external state sensors to create a
  ↪  closed-loop feedback mechanism, allowing the system to self-correct and personalize
  ↪  its execution path on the fly."
}

**EXAMPLE 2 (Data & Computing):**
Topic: Environmental Toxicology Prediction
Paper X: Confined to static predictions using traditional machine learning for single
↪  chemicals.
Paper X+1: Employs hybrid quantum-classical models to integrate multi-omics data for
↪  real-time threat detection.

Output 2:
{
  "concrete_problem": "The original paper is primarily confined to static predictions for
  ↪  isolated tasks using traditional machine learning. It struggles with heterogeneous
  ↪  data and lacks dynamic, real-time response capabilities.",
```

```
  "concrete_solution": "The target paper introduces multi-omics data integration handled
  ↪  by hybrid quantum-classical models, establishing a framework for real-time threat
  ↪  detection and adaptive risk management.",
  "abstract_problem": "The limitation of static, local modeling operating on classical
  ↪  computing architectures. An isolated feature extraction paradigm fails to capture
  ↪  non-linear interactions within dynamic systems, succumbing to the curse of
  ↪  dimensionality.",
  "abstract_solution": "A paradigm shift toward dynamic global integration and
  ↪  cross-paradigm computing. By merging cross-modal information and introducing
  ↪  next-generation computing architectures, it overcomes the processing limits of
  ↪  classical systems."
}
```

User template.

```
Topic: {target topic}

[Paper X]:
Title: {source-paper title}
Idea: {source-paper idea}

[Paper X+1]:
Title: {target-paper title}
Idea: {target-paper idea}

[Trend from X to X+1]:
{citation-grounded trajectory description}

Extract the transition in JSON format.
```

C.3.3 HISTORICAL-LIBRARY INNOVATION MOVE DISTILLATION

System prompt.

```
You are a scientific methodology analyst. Given two directly
connected papers, identify the intellectual transition from the source paper to
the target paper. First describe the concrete problem and solution, then express
their reusable methodological pattern.

CONTENT REQUIREMENTS:
- `concrete_problem` and `concrete_solution` retain the actual application,
  algorithm, architecture, data, and evidence needed to faithfully describe the
  two papers.
- `abstract_problem` explains the reusable limitation: its cause, the failure or
  bottleneck it creates, and when it matters. Include only elements supported by
  the concrete evidence.
- `abstract_solution` explains the reusable methodological shift: the mechanism,
  how information/control/feedback moves, and how that mechanism answers the
  stated cause. Include an important assumption, trade-off, or boundary when the
  evidence supports one.
- Preserve causal correspondence between the abstract problem and solution.
- Each abstract field must be substantive and self-contained, but use exactly the
  amount of detail needed. There is no target or minimum word count.

ABSTRACTION LEVEL:
- Remove paper, model, benchmark, dataset, and method names; acronyms; numerical
  settings; application-specific entities; and low-level implementation details.
- Abstract fields must not name an application area, profession, user population,
  task-specific object, or data modality. For example, replace clinical diagnosis,
  patients, human-AI dialogue, images, language, robots, or molecules with their
  functional roles such as decision process, interacting components, observations,
  sequential exchange, controlled actor, or structured entity.
- Retain discriminative, cross-domain methodological concepts when they define
  the innovation, such as sequential state, iterative feedback, hierarchical
  decomposition, adaptive allocation, uncertainty handling, constraint-aware
  optimization, multi-view evidence, retrieval, verification, or coordination.
- The abstract transition should transfer to at least one substantially different
  research area without changing its causal logic. It does not need to fit every
  imaginable field, and it should not be reduced to generic words such as
  "system", "process", or "improve performance".
- Prefer the most specific transferable description over both domain leakage and
  empty universal language.

STYLE:
- Write natural causal explanations rather than following a fixed five-part
```

```
  template. Vary sentence structure and do not mechanically repeat labels such as
  "structural limitation", "bottleneck", "methodological shift", or "this matters".
- Do not invent claims. Do not praise the target method. Do not mention these
  instructions or the transfer test.

Return only one JSON object with exactly these string keys:
concrete_problem, concrete_solution, abstract_problem, abstract_solution.
```

User template.

```
Topic: {historical-chain topic}

Source paper title: {source-paper title}
Source idea:
{source-paper idea}

Target paper title: {target-paper title}
Target idea:
{target-paper idea}

Citation context showing how the target builds on the source:
{citation context, or "(not available)"}
```

C.3.4 FINAL IDEA-GENERATION PROMPT

System prompt.

```
You are a senior computer-science researcher proposing one new, technically
grounded, and testable research idea for the supplied topic.

The target research topic defines the task and scope. The concrete historical
trajectory describes how existing work has developed, and its final solution is
the current starting point rather than the new contribution. Propose a meaningful
next step beyond it.

The retrieved records contain abstract directions that followed related research
trajectories. Select one direction as the main guide and, only when useful, at most
one additional direction. Transfer the underlying research pattern rather than
copying source-domain terminology, datasets, or components.

Ground the selected direction in the target topic and concrete trajectory. Clearly
state the unresolved problem, explain the proposed method and how it changes the
current approach, and describe the main training and inference procedure when
applicable. The proposal should form one coherent idea whose motivation, method,
contributions, evaluation, and limitations are mutually consistent.

Do not invent empirical results, citations, datasets, or resources that are not
provided in the input. The proposal must stand alone and must not mention retrieval,
reference ranks, prompts, analogies, or these instructions.

The proposal must contain exactly these seven Markdown sections:
1. Proposed Title
2. Background & Motivation
3. Problem Statement
4. Proposed Methodology & Framework
5. Key Contributions
6. Experimental Design
7. Challenges & Limitations

Return only four tag pairs and no text outside them. The third pair contains only
one or two integers separated by a comma.

<ABSTRACT_PROBLEM>
[the unresolved problem addressed by the proposal]
</ABSTRACT_PROBLEM>
<ABSTRACT_SOLUTION>
[the selected methodological direction as grounded in the proposal]
</ABSTRACT_SOLUTION>
<USED_REFERENCE_RANKS>
[one rank, or two comma-separated ranks]
</USED_REFERENCE_RANKS>
<GENERATED_IDEA>
[the complete seven-section Markdown proposal]
</GENERATED_IDEA>
```

User template.

```
[TARGET RESEARCH TOPIC]
{target research topic}

[TARGET CONCRETE HISTORICAL TRAJECTORY]
{formatted target concrete trajectory}

[ABSTRACT REFERENCE DIRECTIONS FROM STEP C]
{formatted retrieved prefix--successor records}

[TASK]
Use the target topic and concrete trajectory to identify an unresolved problem after
the final existing solution. Select the most suitable subsequent abstract direction
from the retrieved records and translate it into a concrete method for the target
setting. Use at most one additional direction if it helps complete the same idea.
Ensure that ABSTRACT_PROBLEM and ABSTRACT_SOLUTION summarize the problem and method
actually presented in GENERATED_IDEA. Follow the required tags and seven section
names exactly.
```

C.3.5 OUTPUT-NORMALIZATION PROMPTS

All four systems are normalized with the same lossless formatter instruction and the same ordered fields. The only source-specific operation is unpacking each system's native serialization into the input payload. The shared instruction is prepended verbatim to each of the four payload templates below.

Shared formatter instruction.

```
You are a lossless scientific-proposal formatter, not an idea generator,
reviewer, or summarizer. Convert the supplied proposal into the exact canonical
template below.

Use every heading and bold field label exactly once, in this order. Do not add
any other heading, subheading, or labeled subsection. Omitting a field is invalid
even when the source provides no information for it; in that case, retain the
field label and use the required missing-information sentence.

# Proposed Title
[one title line]

# Background & Motivation
**Background:** [source-supported background and prior-work context]
**Motivation:** [source-supported motivation and importance]

# Problem Statement
**Existing Approach:** [what the source says is currently done]
**Unresolved Problem:** [all source-stated gaps, limitations, or questions]
**Hypothesis or Objective:** [the source hypothesis or intended objective]

# Proposed Methodology & Framework
**Method Overview:** [overall proposed approach]
**Mechanism and Components:** [all proposed components, operations, signals, and
their relationships]
**Training or Update Procedure:** [all source-stated training, optimization, or
update information]
**Inference or Execution Procedure:** [all source-stated inference, planning, or
execution information]
**Rationale or Causal Hypothesis:** [all source-stated explanations of why the
method should work and any causal hypothesis]

# Key Contributions
**Contributions:** [all contribution claims in the source]

# Experimental Design
**Evaluation Setup:** [all source-stated tasks, datasets, baselines, controls,
evaluation protocols, and implementation settings]
**Evaluation Metrics:** [all source-stated metrics and success criteria]
**Ablations and Analyses:** [all source-stated ablations, robustness checks,
comparative analyses, or diagnostic experiments]

# Challenges & Limitations
**Challenges:** [all source-stated implementation or research challenges]
**Limitations and Failure Conditions:** [all source-stated assumptions, risks,
failure cases, and limitations]
```

```
Preserve every proposal-bearing statement. Do not
summarize away details, merge distinct claims, reduce a list of mechanisms, or
replace specific statements with broader language. Light paraphrasing and
reordering are allowed only to place the same information into the canonical
fields. Treat every source sentence and bullet as an item to be accounted for;
before returning, compare the normalized text against the source line by line and
restore every missing statement or clause. There is no brevity
target or word limit. Retain source-level examples and qualifications. Do not
improve, complete, criticize, or simplify the idea, and do not add scientific
content or external knowledge. Do not create a title when the source has none.

If the source contains no information for a field, write exactly
"Not specified in the source proposal." Return only the normalized proposal.
```

CoI-Pro payload.

```
[TARGET TOPIC]
{target topic}

[SOURCE PROPOSAL]
{generated_idea}
```

CoI Agent payload.

```
[TARGET TOPIC]
{target topic}

[SOURCE PROPOSAL: TITLE]
Not provided in the source.

[SOURCE PROPOSAL: MOTIVATION]
{generated_idea.motivation}

[SOURCE PROPOSAL: NOVELTY]
{generated_idea.novelty}

[SOURCE PROPOSAL: METHOD]
{generated_idea.method}
```

AI-Researcher payload.

```
[TARGET TOPIC]
{target topic}

[SOURCE PROPOSAL]
{generated_idea}
```

AI-Scientist-v2 payload.

```
[TARGET TOPIC]
{target topic}

[SOURCE PROPOSAL: TITLE]
{generated_idea.Title}

[SOURCE PROPOSAL: SHORT HYPOTHESIS]
{generated_idea.Short Hypothesis}

[SOURCE PROPOSAL: RELATED WORK]
{generated_idea.Related Work}

[SOURCE PROPOSAL: ABSTRACT]
{generated_idea.Abstract}

[SOURCE PROPOSAL: RISK FACTORS AND LIMITATIONS]
{generated_idea.Risk Factors and Limitations}
```

C.3.6 IDEA ARENA JUDGE PROMPT

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

The idea0 is: {idea 0}

The idea1 is: {idea 1}

The topic is: {target topic}

Which idea do you think is better? Please write a short paragraph to explain your choice.

Here are your evaluation criteria:
1. Novelty: Are the problems or approaches new? Is this a novel combination of familiar
↪  techniques? Is it clear how this work differs from previous contributions? Is related
↪  work adequately referenced?
2. Significance: Are the idea important? Are other people (practitioners or researchers)
↪  likely to use these ideas or build on them? Does the idea address a difficult problem
↪  in a better way than previous research? Does it provide a unique theoretical or
↪  pragmatic approach?
3. Feasibility: Can the idea be realized with existing technology or methods? Are there
↪  any technical difficulties or bottlenecks? Is the idea clear and logical? Is there any
↪  obvious error or unreasonable part in the idea, and can the experiment be designed
↪  normally according to this idea.
4. Clarity: Is the paper clearly written? Is it well-organized? Does it adequately inform
↪  the reader?
5. Effectiveness: How likely the proposed idea is going to work well (e.g., better than
↪  existing baselines).

Note:
Avoid any position biases and ensure that the order in which the responses were presented
↪  does not influence your decision. DO NOT allow the LENGTH of the responses to
↪  influence your evaluation, choose the one that is straight-to-the-point instead of
↪  unnecessarily verbose. Be as objective as possible. (very important!!!)

If you think idea0 is better than idea1, you should output 0. If you think idea1 is better
↪  than idea0, you should output 1. If you think idea0 and idea1 are equally good, you
↪  should output 2.

Your output should be strictly in following format:
Your thinking process:
...

Your choice:
<novelty>{ Your choice for novelty }</novelty>
<significance>{ Your choice for significance }</significance>
<feasibility>{ Your choice for feasibility }</feasibility>
<clarity>{ Your choice for clarity }</clarity>
<effectiveness>{ Your choice for effectiveness }</effectiveness>
```

D IDEA ARENA AND STATISTICAL DETAILS

D.1 EVALUATION PROTOCOL

For each topic and each unordered pair of methods, Idea Arena evaluates both presentation orders. With four systems, this produces $4 \times 3 = 12$ ordered comparisons per topic and 600 ordered comparisons per judge over 50 topics. Every ordered comparison returns five labels—novelty, significance, clarity, feasibility, and expected effectiveness—each taking value 0 (the first idea wins), 1 (the second idea wins), or 2 (tie). Reversing the presentation order yields 100 decisions for every unordered method pair and dimension.

The judge prompt is identical across judge families. A response is accepted only if all five XML-style fields contain one of $\{0, 1, 2\}$. Invalid or incomplete responses are retried with exponential backoff for up to five attempts. The final archives contain 600 valid, non-empty textual judge responses for each judge; failed parses are not converted into synthetic ties.

D.2 Preference Score

For a fixed contrast, labels are first mapped into wins, ties, and losses from the perspective of the named method. We use

$$\mathrm{Pref} = \frac{W + 0.5T}{W + T + L}.$$

The aggregate score pools the five dimensions and all opponents. Because both presentation orders are retained, a method must remain preferred when it appears in either position to obtain a strong aggregate score.

D.3 Elo Calculation

We compute an order-independent Elo fit from all dimension-level outcomes in an arena. For methods i and j with ratings R_i and R_j, the fitted win probability is

$$P(i \succ j) = \frac{1}{1 + 10^{(R_j - R_i)/400}}.$$

A win contributes target value 1, a loss 0, and a tie 0.5. We maximize the resulting Bernoulli log-likelihood over all ordered comparisons and five dimensions, then center the four fitted ratings to have mean 1,000. This Bradley–Terry form of Elo avoids dependence on an arbitrary sequential match order.

D.4 Confidence Intervals

The 95% confidence intervals in the main paper are percentile bootstrap intervals. We sample the 50 topics with replacement 10,000 times and recompute the reported statistic for every resample. All opponents, dimensions, and presentation orders associated with a sampled topic are retained together, making the topic rather than an individual judge label the resampling unit.

E Complete Main-Comparison Results

Tables 8–10 report all six unordered contrasts in each four-system arena. Every dimensional entry is $W/T/L$ from the perspective of the first method and sums to 100. The preference column pools the five dimensions.

Table 8: Complete round-robin results under DeepSeek-V4-Pro.

Contrast	Nov.	Sig.	Cla.	Feas.	Eff.	Overall pref.
CoI-Pro vs. AI-Researcher	51/1/48	40/0/60	58/4/38	68/0/32	55/0/45	54.9%
CoI-Pro vs. CoI Agent	57/0/43	57/0/43	48/2/50	73/0/27	67/0/33	60.6%
CoI-Pro vs. AI-Scientist-v2	67/2/31	72/4/24	77/1/22	74/0/26	77/0/23	74.1%
AI-Researcher vs. CoI Agent	45/2/53	47/8/45	47/2/51	50/10/40	53/0/47	50.6%
AI-Researcher vs. AI-Scientist-v2	54/0/46	72/5/23	75/1/24	60/3/37	79/0/21	68.9%
CoI Agent vs. AI-Scientist-v2	61/1/38	69/0/31	80/1/19	41/7/52	65/0/35	64.1%

Table 9: Complete round-robin results under GPT-5.

Contrast	Nov.	Sig.	Cla.	Feas.	Eff.	Overall pref.
CoI-Pro vs. AI-Researcher	65/3/32	53/3/44	58/3/39	58/1/41	55/0/45	58.8%
CoI-Pro vs. CoI Agent	61/0/39	48/0/52	46/3/51	67/2/31	57/0/43	56.3%
CoI-Pro vs. AI-Scientist-v2	68/1/31	78/1/21	85/3/12	65/2/33	80/0/20	75.9%
AI-Researcher vs. CoI Agent	41/1/58	49/4/47	52/1/47	68/1/31	57/0/43	54.1%
AI-Researcher vs. AI-Scientist-v2	51/2/47	76/2/22	74/1/25	53/3/44	80/0/20	67.6%
CoI Agent vs. AI-Scientist-v2	62/0/38	81/1/18	77/3/20	43/1/56	71/0/29	67.3%

Table 10: Complete round-robin results under Claude-Opus-5.

Contrast	Nov.	Sig.	Cla.	Feas.	Eff.	Overall pref.
CoI-Pro vs. AI-Researcher	58/14/28	51/10/39	61/4/35	48/8/44	59/3/38	59.3%
CoI-Pro vs. CoI Agent	48/21/31	48/10/42	50/5/45	73/5/22	56/6/38	59.7%
CoI-Pro vs. AI-Scientist-v2	61/15/24	72/13/15	79/6/15	59/11/30	79/2/19	74.7%
AI-Researcher vs. CoI Agent	31/17/52	43/15/42	43/3/54	55/7/38	44/4/52	47.8%
AI-Researcher vs. AI-Scientist-v2	50/4/46	69/7/24	86/3/11	45/8/47	72/3/25	66.9%
CoI Agent vs. AI-Scientist-v2	66/5/29	81/9/10	84/2/14	43/8/49	76/6/18	73.0%

Table 11: Complete fitted Elo ratings for the main arenas. Each column is fitted and centered independently.

Method	DeepSeek-V4-Pro	GPT-5	Claude-Opus-5
CoI-Pro	1073	1076	1081
AI-Researcher	1027	1024	1011
CoI Agent	1006	1013	1029
AI-Scientist-v2	895	887	879

F Complete Ablation Round-Robin Results

Table 12 completes the GPT-5 component arena by reporting all six pairwise contrasts. As in the main evaluation, every dimensional entry contains 100 order-controlled decisions and is written from the perspective of the first variant. The corresponding fitted Elo ratings are 1,056 for $A1R1$, 988 for $A0R0$, 986 for $A0R1$, and 970 for $A1R0$, yielding the same ordering as the overall preference scores in Table 2.

Table 12: Complete round-robin component results under GPT-5.

Contrast	Nov.	Sig.	Cla.	Feas.	Eff.	Overall pref.
$A1R1$ vs. $A0R0$	63/1/36	59/1/40	49/0/51	62/1/37	60/0/40	58.9%
$A1R1$ vs. $A0R1$	60/4/36	60/3/37	46/0/54	63/3/34	63/0/37	59.4%
$A1R1$ vs. $A1R0$	55/5/40	60/2/38	52/0/48	81/0/19	66/0/34	63.5%
$A0R0$ vs. $A0R1$	49/3/48	48/4/48	45/0/55	47/2/51	52/0/48	49.1%
$A0R0$ vs. $A1R0$	42/5/53	44/6/50	47/0/53	68/1/31	57/2/41	53.0%
$A0R1$ vs. $A1R0$	38/5/57	43/5/52	48/0/52	66/2/32	52/0/48	50.6%