

From Mechanism Discovery to Proposal Closure: Graph-Grounded Hierarchical Search for Scientific Ideation

Zijian Hu¹ Hanqi Li^{1,4} Zijian Wang¹ Tiancheng Huang¹
Chenrun Wang^{1,2} Ziyue Yang¹ Kai Yu^{1,3,4} Lu Chen^{1,2,3,4*}

¹X-LANCE Lab, School of Computer Science, Shanghai Jiao Tong University, Shanghai, China

²Shanghai Innovation Institution, Shanghai, China

³Jiangsu Key Lab of Language Computing, Suzhou, China

⁴Suzhou Laboratory, Suzhou, China

*Corresponding author

Abstract

Large language models can generate research ideas that appear novel, but novelty alone does not determine whether an idea is worth following up. In practice, researchers also need ideas that are feasible, experimentally testable, and consistent with the evidence they build on. We therefore assess early-stage proposal quality through four dimensions: novelty, feasibility, experimental design, and evidence consistency. We formulate scientific ideation as graph-grounded hierarchical search over a structured scientific evidence graph. Constructed at the method level, the graph encodes core methods, components, limitations, future work, baselines, datasets, and their relations. On top of this representation, we propose a two-stage framework: the first stage explores candidate mechanism directions over evidence-backed topic and problem frames, and the second stage refines promising candidates through evidence-grounded closure, binding ideas to concrete baselines, datasets, and experimental protocols. In controlled evaluation on 40 matched problem frames from 20 computer science topics, we compare against ResearchAgent, Chain-of-Ideas, direct LLM generation, and Text-RAG under the same four-dimensional protocol. The proposed framework achieves the highest mean overall score in both the dual-LLM and expert evaluations, with its clearest gains in feasibility and evidence consistency. ResearchAgent performs better in novelty and experimental design in both evaluations and remains close overall in the dual-LLM evaluation. Ablations further support complementary contributions from structured evidence, hierarchical search, and proposal closure.

1 Introduction

Large language models (LLMs) (Brown et al., 2020) are increasingly being explored as tools for scientific ideation and can generate research ideas that appear novel. However, novelty alone is not enough to make an idea worth pursuing. Researchers also need proposals that are implementable, experimentally testable, and consistent with the evidence they invoke. We therefore study early-stage proposal quality through four dimensions: novelty, feasibility, experimental design, and evidence consistency.

Many existing ideation systems still construct proposals primarily in text space. They retrieve papers, summarize them, and ask an LLM to synthesize a proposal. This can produce fluent ideas, but often leaves key research decisions under-specified: which prior methods form the right comparison set, what evidence supports the central mechanism, and what experiment would directly test the main claim. As a result, generated proposals may sound promising while remaining weak in feasibility, experimental design, or evidence consistency.

We address this limitation with a structured scientific evidence graph built from structured paper analysis. Unlike a standard citation graph, it organizes core method entities, extracted baseline comparisons, and dataset/evaluation mentions while retaining local evidence for methodological links, comparison structure, and evaluation context. On top of this representation, we propose a hierarchical graph-grounded framework spanning topic discovery, problem framing, mechanism exploration, and proposal closure.

We evaluate the framework on 40 matched problem frames from 20 topics. This controlled setting isolates downstream proposal quality after framing. We compare against ResearchAgent, Chain-of-Ideas, direct LLM generation, and Text-RAG. Under this matched-frame setting, our framework achieves the highest mean overall score in both the dual-LLM and expert evaluations, with the clearest gains in feasibility and evidence consistency. ResearchAgent performs better in novelty and experimental design in both evaluations and remains close overall in the dual-LLM evaluation.

2 Related work

Literature-grounded ideation. Recent systems study research ideation over scientific literature rather than from parametric model knowledge alone. Retrieval-augmented approaches ground generation in retrieved external documents, reducing reliance on implicit parametric knowledge (Lewis et al., 2020). Our work follows this literature-grounded paradigm, but differs in representing retrieved literature as structured evidence over methods, baseline comparisons, and evaluation settings, rather than treating papers as the primary reasoning units.

Graph-grounded ideation. Graph-based retrieval and generation methods show that structured relations can improve grounding, controllability, and multi-hop evidence access (Procko & Ochoa, 2024). ResearchAgent generates ideas from a core paper augmented with relevant references and retrieved entities over an academic graph, while Chain of Ideas organizes literature into a chain structure intended to reflect progressive development within a research domain (Baek et al., 2025; Li et al., 2024). Our work builds on this direction, but uses a structured method-evolution evidence graph that explicitly separates core methods, baseline comparisons, and evaluation context, and couples this representation with staged mechanism exploration and proposal closure.

Idea evaluation. Evaluating research ideas remains open-ended: prior work relies on human assessment, pairwise comparison, and increasingly LLM-based judges, each with different trade-offs between scale and evaluation fidelity (Si et al., 2024; Qiao et al., 2026; Chen et al., 2024a). This motivates our matched-frame evaluation, in which normalized proposals are independently assessed by two LLM judges through blinded five-way listwise ranking, together with an expert evaluation on a smaller subset.

3 Method

Our system performs graph-grounded scientific ideation over a structured method-evolution evidence graph.

3.1 Structured method-evolution evidence graph

We build a *structured method-evolution evidence graph* rather than a paper-level citation graph. Figure 1 illustrates the graph construction pipeline. Its basic unit is a core method entity grounded in a paper, together with baseline and dataset entities extracted from local comparison and evaluation contexts.

Formally, we represent the graph as

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{a}_V, \mathbf{a}_E),$$

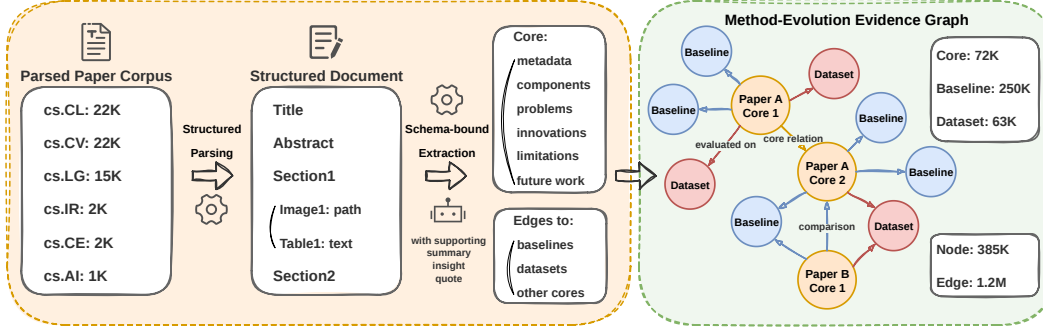


Figure 1: Construction of the structured method-evolution evidence graph from structured paper analysis. Papers are parsed into structured content and converted through schema-bound extraction into core-method, baseline, and dataset entities connected by evidence-bearing relations.

where node attributes $\mathbf{a}_V$ store structured scientific information, including method components, problems, innovations, limitations, future work, baselines, and datasets. Edge attributes $\mathbf{a}_E$ retain local relational evidence such as comparison context, summaries, supporting passages, metrics, and evaluation roles when available.

The graph is constructed from structured paper analysis rather than raw citation expansion. We first parse each paper using the MinerU document parsing pipeline (Wang et al., 2024), and then apply schema-constrained extraction to identify core methods, baseline comparisons, datasets, and the evidence-bearing relations among them.

3.2 Topic family discovery and problem framing

The pipeline groups related core methods into method-family topics and organizes their local evidence into five functional categories: problem gaps, method and component evidence, baseline anchors, experimental evidence, and novelty neighbors. These categories provide different evidence views for subsequent framing and proposal generation rather than treating the graph neighborhood as an undifferentiated retrieval corpus.

For each topic, the pipeline constructs a set of problem frames from the associated problems, limitations, future directions, and comparison evidence. The frames are fixed before proposal search. For a selected frame f , the search receives a frozen evidence package

$$P_f = (p_f, M_f, X_f, B_f, N_f),$$

where p_f contains the problem frame and its direct supporting evidence, M_f contains relevant method and component evidence, X_f contains experimental evidence, B_f contains matched baseline anchors, and N_f contains nearby methods and prior work relevant to novelty assessment. Search is therefore conducted within a fixed problem frame and a bounded evidence context.

3.3 Proposal representation and hierarchical search

Given P_f , the system performs a two-stage graph-grounded search, as illustrated in Figure 2. The upper level explores alternative mechanism directions, while the lower level converts promising mechanisms into complete proposals through evidence binding and experimental specification.

For exposition, we summarize a proposal candidate as

$$\mathbf{z} = (m, i, c, b, e),$$

where m is the proposed mechanism, i is the proposal rationale and its distinction from nearby work, c is the central claim, b records the supporting evidence and baseline con-

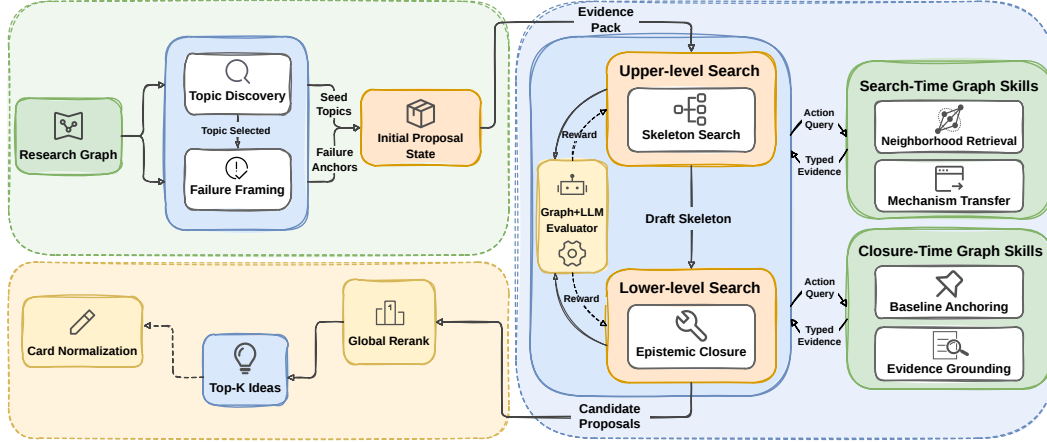


Figure 2: Two-stage graph-grounded proposal search. A frozen problem frame and its evidence package initialize the process. The upper level explores alternative mechanism directions through graph-conditioned operators, and the lower level completes promising candidates through proposal refinement, experimental specification, and evidence binding.

nections, and e is the experimental plan. The frame f and evidence package P_f are shared search context rather than mutable candidate fields.

At the upper level, m and i contain the primary mechanism-level decisions, while c , b , and e may remain provisional or only partially specified. The proposal-closure stage subsequently completes, revises, and re-evaluates these fields.

3.4 Upper-level search: mechanism exploration

The upper level explores mechanism hypotheses under the fixed problem frame. Its operator families capture several forms of scientific change, including responding to recurring failure patterns, introducing new interfaces between supported components, extending local method trajectories, transferring mechanisms across related families, and revising existing representations or assumptions. Each operator exposes a corresponding slice of graph evidence and constrains the perspective from which the candidate is generated.

Let $\mathcal{A}$ denote the operator set and $R_a(P_f)$ the evidence selected for operator a . A mechanism candidate is generated as

$$z_a \sim p_\theta(z \mid f, R_a(P_f), a), \quad a \in \mathcal{A}.$$

The search first covers the available operator families and then uses critic feedback to refine promising directions. This coverage-first strategy preserves diversity across mechanism families while still allocating additional refinement to stronger candidates.

At this stage, candidates are assessed using novelty and feasibility:

$$s_1(z) = \frac{r_{\text{nov}}(z) + r_{\text{fea}}(z)}{2}.$$

The resulting scores and reviewer-style feedback guide the selection of mechanism candidates passed to proposal closure.

3.5 Lower-level search: evidence binding and experimental grounding

A mechanism hypothesis alone does not yet specify a complete research proposal. The lower level therefore performs proposal closure by refining the proposal rationale, defining a falsifiable central claim, selecting appropriate baselines and evaluation settings, and binding each major component to the available evidence.

We summarize this process as

$$z' = E(F(R(z))),$$

where R develops the proposal rationale and its relation to prior work, F constructs the claim and falsification-oriented experimental plan, and E binds the resulting proposal to graph-backed problem, method, baseline, and experimental evidence. Critic feedback may be used to revisit the proposal rationale and evidence bindings while preserving the central experimental structure.

This stage is designed to turn a plausible mechanism direction into a proposal whose claim, comparison setting, and experiment are mutually aligned.

3.6 Multi-objective critic and progressive selection

Each completed candidate is assessed along four dimensions:

$$r(z) = (r_{\text{nov}}(z), r_{\text{fea}}(z), r_{\text{exp}}(z), r_{\text{ec}}(z)),$$

corresponding to novelty, feasibility, experimental design, and evidence consistency. Their arithmetic mean provides a summary score for search-time comparison:

$$s_2(z) = \frac{r_{\text{nov}}(z) + r_{\text{fea}}(z) + r_{\text{exp}}(z) + r_{\text{ec}}(z)}{4}.$$

The critic combines an LLM-based assessment with reviewer-style checks, lightweight heuristics, and graph-side evidence diagnostics. The mean score guides candidate refinement, while final selection additionally considers evidence-related risks and removes near-duplicate mechanisms. The search-time critic is instantiated with GPT-5.2 and is kept separate from the blinded evaluation pipeline, with independent prompts and contexts used for the final judges.

4 Experiments

4.1 Setup

We evaluate our framework on a structured method-evolution evidence graph constructed from nearly 50k computer science papers indexed through Semantic Scholar (Kinney et al., 2023). After structured parsing and schema-constrained extraction, the full graph contains about 72k core method nodes, about 250k baseline-comparison mentions, and about 63k dataset/evaluation mentions, totaling about 385k nodes and 1.15M evidence-bearing edges. The dominant core domains are cs.CL, cs.CV, cs.LG, and cs.AI, with a smaller tail from neighboring CS areas.

Our experiments operate on topic-local regions rather than on the full graph at once. Topic families are first discovered and post-processed into locked topic packets, after which evidence packets, clean problem frames, and frame-matched baseline locks are constructed for downstream search. We select 20 topics and 40 corresponding system-derived frames for the main evaluation, with a 10-frame subset used for expert evaluation. We use matched problem frames as a controlled comparison setting to isolate downstream proposal quality after framing rather than end-to-end topic-to-frame discovery.

4.2 Compared methods

We compare five systems under the same system-derived topic and locked problem-frame input. GPT-5.2 is used as the common generation backbone where an LLM backend is required.

ResearchAgent. A reproduced multi-agent scientific-ideation baseline that retains its core proposal-development and validation workflow. We adapt its input/output interface to the matched problem frames and common proposal schema, and replace its retrieval and

LLM channel with the shared retrieval and LLM backend. We therefore treat it as a strong reproduction under the common protocol rather than an exact re-run of the original system at its reported best configuration.

Chain-of-Ideas. A reproduced iterative ideation baseline that retains its idea-chain evolution and review loop. As with ResearchAgent, the adaptation is limited to the matched-frame interface, common output schema, and shared retrieval/LLM channel. Its results should consequently be interpreted as performance under our controlled evaluation setting rather than as a direct reproduction of every implementation choice in the original system.

Direct LLM. A one-pass baseline that directly produces a structured proposal from the topic and frame. Its output follows the same proposal schema as our method, except for graph-specific fields that cannot be instantiated fairly without graph access.

Text-RAG. A retrieval-augmented baseline in the spirit of RAG that retrieves topic- and frame-relevant textual evidence and then generates a proposal under the same schema as Direct LLM and our method. The evidence budget is matched as closely as possible to that used by our method, but the retrieved evidence remains unstructured text rather than structured graph evidence.

4.3 Evaluation protocol

All raw outputs are passed through one shared normalizer using the same script, prompt, and GPT-5.2 model. The normalizer rewrites each proposal into the same four prose fields, with each field constrained to 150–300 words, while being instructed not to add, remove, or judge concrete proposal content. This step reduces confounds from writing style, output length, and field granularity so that judgments can focus on proposal substance under the same problem frame.

We evaluate each proposal along four dimensions:

- **Novelty:** whether the proposal introduces a genuinely new idea rather than rephrasing prior work, making an incremental update, or relying primarily on combinatorial innovation;
- **Feasibility:** whether the method is clearly described, concretely implementable, and capable of addressing the stated problem;
- **Experimental Design:** whether the proposed experiment can verify or falsify the central claim;
- **Evidence Consistency:** whether the cited evidence coherently supports the problem formulation, proposed mechanism, and experimental design.

Together, these four dimensions assess complementary aspects of proposal quality under a fixed problem frame: the originality of the proposed idea, the clarity and implementability of the method, the validity of the experimental test, and the coherence of the supporting evidence.

We use two complementary evaluation modes. For the main LLM-based evaluation, the five normalized proposals for each of 40 frames from 20 topics are anonymized and independently ranked by GPT-5.2 and Claude Sonnet 4.5. For each dimension, a judge assigns all five systems a strict ordering from best to worst, with each rank used exactly once. We repeat the evaluation under three independent proposal-label shuffles at low temperature. Ranks are converted to scores from 5 (best) to 1 (worst), and the reported dual-LLM results are averaged across the three shuffles and the two judges; the overall score is the mean of the four dimensions.

For expert evaluation, four PhD annotators independently apply the same normalized cards, rubric, and listwise ranking procedure to a 10-frame subset of the main evaluation set. Each annotator ranks all five systems on every dimension for each evaluated frame.

System	Novelty	Feasibility	Exp. Design	Evidence Cons.	Overall
OURS	3.01	4.17	3.25	3.63	3.51
RESEARCHAGENT	4.02	1.86	4.17	3.60	3.41
CHAIN-OF-IDEAS	3.27	2.93	3.48	2.85	3.13
DIRECT.LLM	2.80	2.99	2.59	2.98	2.84
TEXT.RAG	1.92	3.06	1.53	1.96	2.12

Table 1: Dual-LLM evaluation on 40 frames from 20 topics with three repeated evaluations. Each entry reports the arithmetic mean of the scores assigned by GPT-5.2 and Claude Sonnet 4.5.

System	Novelty	Feasibility	Exp. Design	Evidence Cons.	Overall
OURS	3.40	4.12	3.85	4.53	3.97
RESEARCHAGENT	4.17	1.18	4.10	2.88	3.08
CHAIN-OF-IDEAS	2.70	3.17	2.88	2.30	2.76
DIRECT.LLM	2.38	3.30	2.53	2.90	2.77
TEXT.RAG	2.35	3.23	1.65	2.40	2.41

Table 2: Blinded human expert evaluation on 10 frames. Each entry reports the mean score aggregated across four expert annotators under the same listwise evaluation protocol.

4.4 Main results

Tables 1 and 2 summarize the results of the dual-LLM and expert evaluations. On the 40-frame main evaluation, OURS achieves the highest mean overall score (3.51). Its clearest advantage is in feasibility, where it scores substantially higher than all baselines. It also obtains the highest mean evidence-consistency score, although the difference from ResearchAgent is small. ResearchAgent performs best in novelty and experimental design and achieves an overall score close to that of OURS. Compared with Chain-of-Ideas, Direct LLM, and Text-RAG, OURS shows a more consistent advantage in overall proposal quality. These results suggest that the proposed framework is effective at producing proposals that are concrete, implementable, and coherently supported by the retrieved evidence.

The expert evaluation shows a similar dimension-wise pattern on the 10-frame subset. OURS achieves the highest mean overall score (3.97) and leads in feasibility and evidence consistency, whereas ResearchAgent leads in novelty and experimental design. Although the expert evaluation covers fewer frames, it shows the same division of strengths as the main evaluation, with a larger overall gap in favor of OURS on this subset.

4.5 Ablation study

To examine the contributions of structured graph evidence, upper-level mechanism exploration, and lower-level proposal closure, we compare the full system against the following ablations:

- **A: no graph, same search.** Remove structured graph evidence while retaining the same two-stage search procedure.
- **B: graph, no hierarchical search.** Retain the graph-grounded evidence package but replace the two-stage search process with one-shot graph-conditioned generation.
- **C: no lower-level closure.** Retain graph grounding and upper-level mechanism exploration but remove the lower-level proposal-closure stage.

The ablation study uses a separate blind pairwise evaluation on 10 frames rather than the listwise protocol used in the main evaluation. For each ablation, the two proposals produced by OURS are fully crossed with the two proposals produced by the ablated system, yielding four pairwise comparisons per frame and 40 comparisons per ablation. Each pairing is evaluated once, with left-right presentation order balanced across the comparison set using two presentation seeds. For each comparison, Claude Sonnet 4.6 assigns a preference

Ablation	Novelty	Feasibility	Exp. Design	Evidence Cons.
A	+0.325 (24/12/4)	+0.300 (23/17/0)	+0.325 (24/16/0)	+0.225 (21/16/3)
B	+1.725 (37/0/3)	+0.175 (23/15/2)	+1.250 (34/6/0)	+0.875 (27/11/2)
C	+0.100 (21/18/1)	+0.175 (21/19/0)	+0.575 (27/13/0)	+0.150 (20/18/2)

Table 3: Blind pairwise ablation comparison with Claude Sonnet 4.6. Each cell reports the mean signed margin of OURS against the ablation, with win/loss/tie counts in parentheses. A: no graph. B: no hierarchical search. C: no lower-level closure.

score on a seven-point scale from -3 (strong preference for the ablation) to $+3$ (strong preference for OURS), with 0 indicating no preference. We report the mean signed margin, computed by averaging these preference scores across all pairwise comparisons, together with the corresponding win, loss, and tie counts. Because the ablation study differs from the main evaluation in comparison format, judge, and evaluated frames, its margins should be interpreted only within this study and are not directly comparable to the scores in Tables 1 and 2.

Table 3 shows that the full system achieves positive mean margins over all three ablations across all four dimensions, indicating a consistent overall preference for the full system. Variant **B** exhibits the largest degradation, particularly in novelty, experimental design, and evidence consistency. This pattern suggests that graph conditioning alone is insufficient in this setting and that hierarchical search helps convert graph evidence into more fully specified proposals and experiments.

Variants **A** and **C** show smaller but consistently positive margins in favor of the full system. The comparison with **A** is consistent with hierarchical search not fully replacing structured graph evidence, while the comparison with **C** is consistent with lower-level closure contributing to evidence binding and experimental specification. Given the limited size of the pairwise study, these results support the complementary roles of the three components but do not establish precise differences in their relative contributions.

5 Case study

5.1 Dynamic prompt adaptation in vision-language tracking

In the vision-language tracking subgraph, the key failure anchors come from ATCTrack (Feng et al., 2025) and ReasoningTrack (Wang et al., 2025b). ATCTrack explicitly identifies the mismatch between initial target-context cues and evolving target states, while ReasoningTrack likewise motivates updating language descriptions as target appearance changes over long sequences. These anchors are paired with matched baseline-comparison edges to earlier VLT-family trackers, including JointNLT (Zhou et al., 2023), which are identified in the graph as the affected comparison family. The resulting frame therefore focuses on the narrower bottleneck of adapting language cues to dynamic target states. This framing is also consistent with related VLT work that identifies limitations of static references or initial prompts under target-state changes (Feng et al., 2024; Li et al., 2025).

Conditioned on this frame, upper-level search does not freely invent a new tracker. It consumes two graph-backed evidence types. First, the problem anchors from ATCTrack and ReasoningTrack establish the failure mode: static prompt conditioning becomes unreliable as the target changes. Second, the method-side precedents are narrower than the generated proposal. ATCTrack provides evidence for temporal visual target-context modeling and context-word calibration, while ViPT (Zhu et al., 2023) provides interface-level evidence that prompt-based adaptation of a frozen tracking backbone is feasible. Search operates over this constrained prompt-update interface and proposes maintaining multiple target-state prompt hypotheses, with their contributions selected online as the tracked target evolves.

The general primitives of multiple prompts and learned routing are not themselves new. MoCoOp routes image instances among multiple soft-prompt experts, MoPD uses a gating network to select prompts for distillation, and pFedMoAP combines multiple prompt

experts through input-dependent gating (Du et al., 2025; Chen et al., 2024b; Luo et al., 2025). Within tracking, VPTracker uses a location-aware visual prompt derived from the previous target position to prioritize regional recognition while retaining global search capability (Wang et al., 2025a). Relative to these precedents, the generated proposal is more narrowly characterized as a temporal adaptation of prompt routing to VLT: the prompt components represent competing target-state hypotheses, and their weights are updated from sequence-level tracking state rather than selected only from an individual input or fixed client context. The graph supports the need for temporal adaptation, but it does not directly establish that this particular temporal mixture is effective or novel; the mechanism therefore remains a hypothesis requiring both literature comparison and empirical validation.

Lower-level closure turns this hypothesis into a diagnostic evaluation plan. Using MGIT (Hu et al., 2023) as the direct graph-supported experiment anchor and extending the evaluation to TNLLT (Wang et al., 2025b), the system constructs four principal conditions: (1) a single continuously updated prompt; (2) multiple prompts combined without learned routing; (3) an instance-conditioned prompt router that selects prompts from the current frame without maintaining temporal state; and (4) the proposed sequence-conditioned router over target-state prompt hypotheses. The comparison between the first two conditions tests whether additional prompt capacity alone is sufficient, while the comparison between the second and third tests the effect of generic prompt routing. Most importantly, the comparison between the third and fourth conditions isolates whether conditioning the routing decision on evolving tracking state provides an advantage beyond existing instance-level mixture-of-prompts mechanisms. A VPTracker-style location-aware prompting baseline can additionally be included where architectural compatibility permits. Under this design, improvements cannot be attributed merely to using more prompts or introducing a generic gating module.

6 Limitations

Our conclusions remain limited in several ways. The graph is built through structured extraction, entity alignment, and relation linking, so extraction errors or incomplete evidence can propagate into later ideation stages. Filtering, critic-based rejection, and evidence-consistency checks reduce the influence of noisy or unsupported graph paths but do not eliminate it. Although we provide an intrinsic sanity check of baseline and dataset extraction on a stratified sample of 80 papers, this analysis does not comprehensively evaluate all extracted node categories, relation links, or their downstream effects. A broader evaluation of graph construction quality, together with a controlled analysis of how graph errors affect generated proposals, remains future work.

More broadly, our conclusions concern ideation-stage proposal quality under matched problem frames and the current evaluation protocol; they do not directly establish downstream research success, unconstrained topic-to-frame discovery, or universal scientific quality. The main evaluation spans 40 frames from 20 topics, whereas the expert evaluation covers a 10-frame subset with four PhD annotators. The expert study therefore provides complementary human evidence rather than a comprehensive validation across all evaluated settings. The larger-scale evaluation uses two independent LLM judges, repeated proposal-label shuffles, and agreement analyses, but evaluator variance remains a source of uncertainty, particularly when differences between systems are small.

Comparing systems with different native interfaces also requires some adaptation. We provide every system with the same problem frame and require a shared proposal schema while retaining its central generation, retrieval, or search procedure. All raw outputs are subsequently processed by the same normalizer and evaluated as anonymized proposal cards under a common rubric. These controls improve comparability by reducing differences in response length, organization, and field granularity. However, they do not fully reproduce every baseline in its original input setting, output format, or end-to-end workflow. The results should therefore be interpreted as comparisons of the systems’ core ideation strategies under a controlled common setting.

7 Conclusion

We propose a graph-grounded hierarchical framework for scientific ideation over structured scientific evidence. The framework combines problem framing, upper-level mechanism exploration, and lower-level evidence-grounded closure to construct concrete and testable proposals within a constrained evidence space. Under matched problem frames, OURS achieves the highest mean overall score in both the dual-LLM and expert evaluations, with the clearest gains in feasibility and evidence consistency; ResearchAgent remains close overall in the dual-LLM evaluation and performs better in novelty and experimental design. These results support the effectiveness of organizing structured scientific evidence into a bounded hierarchical search process for proposal generation.

Acknowledgments

This work was supported by the National Science and Technology Major Project (2023ZD0120703), the China NSFC Projects (62576212, 62120106006, 92370206, U23B2057), the Shanghai Municipal Science and Technology Project (25X010202846), and the Frontier Technologies R&D Program of Jiangsu (BF2025029).

Ethics Statement

This work studies automated support for scientific ideation. Such systems may be misused to generate superficially plausible but weakly justified proposals, to overstate novelty, or to accelerate low-quality idea production at scale. They may also encourage over-reliance on LLM-generated reasoning when the underlying scientific evidence is incomplete, noisy, or incorrectly extracted.

We mitigate these risks through the design of the framework. Our approach emphasizes explicit evidence grounding rather than unconstrained proposal generation, and the proposal-closure stage connects candidate mechanisms with relevant evidence, baselines, datasets, and experimental protocols. Nevertheless, the framework cannot guarantee factual correctness, genuine novelty, or downstream scientific validity. Generated proposals should therefore be treated as hypotheses for further examination rather than as validated scientific conclusions. We do not claim that the system replaces human scientific judgment, literature review, peer review, or experimental validation.

Large language models are used as research components in this work. GPT-5.2 is used as the common generation backbone for the LLM-based systems, the shared proposal normalizer, and the search-time critic in our framework. In the main evaluation, GPT-5.2 and Claude Sonnet 4.5 serve as independent judges under the blinded listwise protocol. Claude Sonnet 4.6 is used to evaluate the ablation comparisons, and GPT-5 is used as the auditor in the graph-extraction sanity check. Search-time critic scores are not reused in any reported evaluation. The expert evaluation additionally uses four PhD annotators and is reported separately from the LLM-based evaluation.

The intended use of the system is to assist researchers in exploring and refining candidate research ideas under explicit evidence and experimental constraints. Its outputs are not intended to establish scientific validity or to substitute for expert verification of evidence, novelty, feasibility, or experimental conclusions.

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, 2025.

- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. Humans or llms as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8301–8327, 2024a.
- Yang Chen, Shuai Fu, and Yu Zhang. Mopd: Mixture-of-prompts distillation for vision-language models. *arXiv preprint arXiv:2412.19087*, 2024b.
- Yu Du, Tong Niu, and Rong Zhao. Mixture of prompts learning for vision-language models. *Frontiers in Artificial Intelligence*, 8:1580973, 2025.
- Xiaokun Feng, Xuchen Li, Shiyu Hu, Dailing Zhang, Meiqi Wu, Jing Zhang, Xiaotang Chen, and Kaiqi Huang. Memvlt: Vision-language tracking with adaptive memory-based prompts. *Advances in Neural Information Processing Systems*, 37:14903–14933, 2024.
- Xiaokun Feng, Shiyu Hu, Xuchen Li, Dailing Zhang, Meiqi Wu, Jing Zhang, Xiaotang Chen, and Kaiqi Huang. Atctrack: Aligning target-context cues with dynamic target states for robust vision-language tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19850–19861, 2025.
- Shiyu Hu, Dailing Zhang, Xiaokun Feng, Xuchen Li, Xin Zhao, Kaiqi Huang, et al. A multi-modal global instance tracking benchmark (mgit): Better locating target in complex spatio-temporal and causal relationship. *Advances in neural information processing systems*, 36:25007–25030, 2023.
- Rodney Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, et al. The semantic scholar open data platform. *arXiv preprint arXiv:2301.10140*, 2023.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.
- Long Li, Weiwen Xu, Jiayan Guo, Ruochen Zhao, Xingxuan Li, Yuqian Yuan, Boqiang Zhang, Yuming Jiang, Yifei Xin, Ronghao Dang, et al. Chain of ideas: Revolutionizing research via novel idea development with llm agents. *arXiv preprint arXiv:2410.13185*, 2024.
- Xiaohai Li, Bineng Zhong, Qihua Liang, Zhiyi Mo, Jian Nong, and Shuxiang Song. Dynamic updates for language adaptation in visual-language tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19165–19174, 2025.
- Jun Luo, Chen Chen, and Shandong Wu. Mixture of experts made personalized: Federated prompt learning for vision-language models. In *International Conference on Learning Representations*, volume 2025, pp. 58061–58080, 2025.
- Tyler Thomas Procko and Omar Ochoa. Graph retrieval-augmented generation for large language models: A survey. In *2024 Conference on AI, science, engineering, and technology (AIxSET)*, pp. 166–169. IEEE, 2024.
- Shuofei Qiao, Yunxiang Wei, Xuehai Wang, Bin Wu, Boyang Xue, Ningyu Zhang, Hossein A Rahmani, Yanshan Wang, Qiang Zhang, Keyan Ding, et al. Innoeval: On research idea evaluation as a knowledge-grounded, multi-perspective reasoning problem. *arXiv preprint arXiv:2602.14367*, 2026.
- Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv preprint arXiv:2409.04109*, 2024.

- Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, et al. Mineru: An open-source solution for precise document content extraction. *arXiv preprint arXiv:2409.18839*, 2024.
- Jingchao Wang, Kaiwen Zhou, Zhijian Wu, Kunhua Ji, Dingjiang Huang, and Yefeng Zheng. Vptracker: Global vision-language tracking via visual prompt. *arXiv preprint arXiv:2512.22799*, 2025a.
- Xiao Wang, Liye Jin, Xufeng Lou, Shiao Wang, Lan Chen, Bo Jiang, and Zhipeng Zhang. Reasoningtrack: Chain-of-thought reasoning for long-term vision-language tracking. *arXiv preprint arXiv:2508.05221*, 2025b.
- Li Zhou, Zikun Zhou, Kaige Mao, and Zhenyu He. Joint visual grounding and tracking with natural language specification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 23151–23160, 2023.
- Jiawen Zhu, Simiao Lai, Xin Chen, Dong Wang, and Huchuan Lu. Visual prompt multi-modal tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9516–9526, 2023.

Type	Micro precision	Macro precision	Micro recall	Macro recall
Baseline	0.902	0.870	0.695	0.702
Dataset	0.966	0.947	0.730	0.771

Table 4: Extraction sanity check on 80 papers. Precision and recall follow the baseline and dataset definitions used by the evidence graph.

A Evidence-graph extraction sanity check

Because proposal generation depends on automatically extracted graph evidence, we conduct a targeted sanity check of baseline and dataset extraction. We sample 80 papers using stratified sampling. For each paper, a GPT-5 auditor receives the full paper text, comparison-table HTML, and the extracted baseline and dataset lists. It labels each extracted item as valid or invalid and identifies valid items present in the paper but missed by the extraction pipeline.

A valid baseline is any named external method, model, or algorithm that is either compared against experimentally or discussed as prior work. The paper’s own method and generic terms such as “the baseline” are excluded. A valid dataset is any explicitly named dataset appearing in the paper.

The baseline audit contains 1,046 true positives, 114 false positives, and 458 false negatives; the dataset audit contains 336 true positives, 12 false positives, and 124 false negatives. Only one extracted baseline and no extracted datasets are judged absent from the source paper. Most false positives instead correspond to boundary cases such as the paper’s own method, generic terms, or duplicate mentions.

This analysis is intended as a sanity check rather than a complete validation of graph quality. It covers a limited sample and relies on a model-based auditor, while some extraction boundaries remain subjective.

B Reviewer-card normalization and evaluation protocol

All five systems are evaluated through a common normalized representation. ResearchAgent and Chain-of-Ideas retain their central proposal-generation procedures but are adapted to the matched problem-frame input and shared output interface. Direct LLM and Text-RAG produce the shared proposal schema directly.

Every raw proposal is subsequently processed by the same normalizer using an identical script, prompt, and GPT-5.2 model. The normalizer is instructed not to add, remove, infer, or judge concrete proposal content. It rewrites each proposal into four prose fields:

problem_definition, method, experiment_design, and evidence.

Each field is constrained to 150–300 words. The normalizer must preserve named methods, datasets, baselines, metrics, ablations, numerical values, components, and claims while removing system-specific formatting and internal evidence identifiers.

The five normalized proposals for each problem frame are anonymized as P1–P5. For each evaluation dimension, a judge assigns the five proposals a strict ordering from best to worst, using every rank exactly once. The four dimensions are:

- **Novelty:** whether the proposal introduces a genuinely new idea rather than rephrasing prior work, making an incremental update, or relying primarily on combinatorial innovation;
- **Feasibility:** whether the method is clearly described, concretely implementable, and capable of addressing the stated problem;
- **Experimental Design:** whether the proposed experiment can verify or falsify the central claim;

Evaluators	Novelty	Feasibility	Exp. Design	Evid. Cons.	Overall
Four humans	0.629	0.766	0.562	0.613	0.579
Two LLM judges	0.828	0.781	0.925	0.796	0.900
All six	0.645	0.736	0.637	0.567	0.620

Table 5: Kendall’s W agreement across evaluator groups on the 10-frame expert-evaluation subset.

- **Evidence Consistency:** whether the cited evidence coherently supports the problem formulation, proposed mechanism, and experimental design.

The main evaluation uses GPT-5.2 and Claude Sonnet 4.5 as independent judges. Each problem frame is evaluated under three independent proposal-label shuffles at low temperature. Ranks are converted to scores from 5 (best) to 1 (worst), and the reported dual-LLM results are averaged across the three shuffles and the two judges. Four PhD annotators independently use the same normalized cards, rubric, and five-way listwise procedure on the expert-evaluation subset.

The ablation study is evaluated separately using Claude Sonnet 4.6. For each frame and each ablation, the full system and the ablated variant are compared under balanced presentation order. The ablation evaluation is therefore separate from the five-way listwise protocol used for the main results.

The search-time critic and final evaluation are separate procedures. Search-time scores are used only to guide proposal refinement and are not reused in any reported evaluation.

C Evaluator agreement

Because each evaluator produces a ranking of the five systems, we use rank-based measures to examine agreement across evaluators. Kendall’s W measures concordance among multiple evaluators. As shown in Table 5, agreement is generally higher between the two LLM judges than among the four human annotators, while the combined six-evaluator agreement remains moderate across dimensions.