Sparking Scientific Creativity via LLM-Driven Interdisciplinary Inspiration

Anonymous ACL submission

Abstract

Interdisciplinary research often yields larger long-term impact, yet most research ideation remains constrained by single-domain assumptions. Recent AI-based scientific discovery systems increasingly automate hypothesis generation, planning, and execution, but often bypass the exploratory reasoning needed to identify meaningful cross-domain inspiration. We present **IDEA-CATALYST**[1], a metacognition-driven framework for early-stage interdisciplinary ideation. Given an abstract research goal, IDEA-CATALYST decomposes it into target-domain research questions, analyzes existing progress and unresolved challenges, reformulates these challenges as domain-agnostic conceptual problems, and retrieves analogous insights from distant source domains. It then recontextualizes these insights into ranked idea fragments that support brainstorming without prematurely anchoring on executable solutions. Across LLM-based and human evaluations, IDEA-CATALYST improves average novelty by **21%** and insightfulness by **16%** while remaining grounded in the original research problem, suggesting a practical path toward AI systems that augment, rather than replace, creative scientific reasoning.

1 Introduction

Scientific breakthroughs rarely arise from a single, isolated "*eureka*" moment. Instead, research ideation unfolds as a complex, iterative formation process in which creative solutions emerge through the gradual synthesis of many fragmentary and partial ideas (Sosa, 2019; Gonçalves and Cash, 2021). These early, tangible conceptual fragments, often originating from multiple domains (Yanai and Lercher, 2019), act as seeds for discussion, critique, and collaboration. They open up more meaningful research questions and address each other's conceptual gaps and challenges, gradually coalescing into more mature research directions (Sosa, 2019; Gonçalves and Cash, 2021). This pattern recurs throughout the history of scientific progress. Reinforcement learning, now a foundational paradigm in machine learning, did not originate within a single field but instead emerged from the convergence of behavioral psychology's reward-driven learning principles, control theory's mathematical formalizations, and animal learning psychology's insights into secondary reinforcement signals (Sutton et al., 1998), as illustrated in Figure 1. The field advanced through the accumulation, recombination, and refinement of ideas contributed by researchers operating under diverse conceptual lenses.

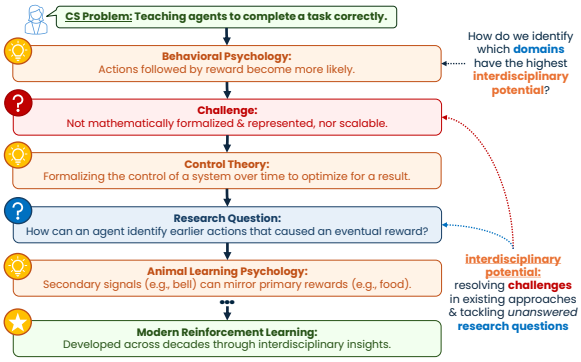

Figure 1: Interdisciplinary process of formalizing RL (Sutton et al., 1998).

Empirical evidence shows that interdisciplinary synthesis yields substantially higher long-term impact, with each additional discipline increasing citation impact by approximately 20% (Van Noorden, 2015; Okamura, 2019). Yet despite this benefit, deeply integrative interdisciplinary research remains rare and fragile: only 5% of cross-domain work involves high-involvement collaboration across non-neighboring fields (Porter and Rafols, 2009; Raasch et al., 2013). A central challenge, then, is how to *foster interdisciplinary sci-*

[1] https://anonymous.4open.science/r/idea_catalyst-D097

entific creativity and help researchers move beyond their academic silos. Recent work on AI-driven scientific discovery has explored the notion of "AI co-scientists," in which large language models (LLMs) support (and in many cases automate) key stages of the research process, including ideation, experimentation, and critique (Gottweis et al., 2025; Goel et al., 2025; Si et al., 2026; Jansen et al., 2025). Prior studies (Si et al., 2024) comparing human- and LLM-generated ideas highlight a complementary but unresolved tension (Si et al., 2024). Human-generated ideas are typically well grounded in existing literature, datasets, and practical constraints, but tend to remain focused on familiar problem formulations within a single domain. LLM-generated ideas, in contrast, exhibit a stronger tendency to draw inspiration from other disciplines, yet often do so in surface-level or stereotyped ways that undermine motivation, feasibility, and practical grounding (Si et al., 2024, 2026; Gupta and Pruthi, 2025).

Attempts to address these limitations by tightly coupling LLM ideation with automatic experiment execution (Si et al., 2026; Jansen et al., 2025) introduce further trade-offs. While grounding ideas through execution improves feasibility, it also drives convergence toward incremental, single-domain refinements, eroding the exploratory strengths and cross-domain potential that initially distinguish LLM-based ideation (Si et al., 2026). Similarly, other works primarily focus on generating execution plans from high-level research ideas, bypassing the exploratory and collaborative processes of ideation (Goel et al., 2025; Huang et al., 2025). This pattern reflects a broader challenge in research ideation: early-stage evaluation, especially when driven by empirical validation, can be counterproductive for creativity. As prior work notes, premature data and evaluation can "cut the conversation" and curtail exploration of the broader space of possibilities (Bose et al., 2013; Catmull and Wallace, 2023).

To avoid prematurely anchoring on end-to-end automated solutions and instead augment the initial creative ideation process itself, we propose **IDEA-CATALYST**, a novel framework for the automatic generation of insightful, interdisciplinary idea fragments. Given a research problem in a target domain (e.g., "*teaching agents to complete a task*" in Computer Science), our goal is to surface conceptual insights from external "source" domains (e.g., *Behavioral Psychology*, *Control Theory*, *Animal Learning Psychology*) that can either help resolve persistent challenges in existing target-domain approaches or address research questions that remain unanswered within the target domain. Our retrieval-augmented, hierarchical ideation framework is guided by three core principles:

1. ***Analyzing the target domain to assess progress and reveal challenges.*** We first decompose the overarching research problem into a structured set of core research questions. By retrieving and analyzing target-domain literature conditioned on each question, the framework identifies what has already been addressed, where progress is uneven, and which challenges remain unresolved. Crucially, this analysis distinguishes between *domain-specific challenges* (e.g., getting an algorithm to learn reliably from limited or noisy feedback) and deeper *conceptual challenges* (e.g., understanding what an agent should aim for when its goals or feedback are unclear or change over time) that persist despite extensive prior work.

2. ***Exploring external source domains to uncover conceptually analogous solutions.*** For each unresolved conceptual challenge, the framework queries & analyzes a diverse set of external domains to determine whether similar problems have been previously studied or solved under different assumptions, formalisms, or empirical settings. This step emphasizes cross-domain awareness, enabling the discovery of alternative perspectives, mechanisms, and abstractions that are absent from the target domain but potentially transferable.

3. ***Recontextualizing and strategically prioritizing interdisciplinary insights.*** Finally, extracted insights from relevant source domains are recontextualized into the language and constraints of the target domain, forming candidate idea fragments. These fragments are then ranked based on their potential to address high-impact challenges, balancing conceptual novelty with relevance to the original research goals. This strategic prioritization supports exploratory ideation while avoiding premature convergence or feasibility-driven pruning.

Overall, **IDEA-CATALYST** provides a pathway for both AI-assisted human research and autonomous scientific discovery systems to engage in the kind of creative, boundary-spanning ideation

process that drives breakthrough innovations. It addresses the critical gap in current automated scientific discovery methodologies by prioritizing the inherently exploratory, collaborative nature of research while providing systematic structure to cross-domain knowledge synthesis. Our main **contributions** are:

- We propose **IDEA-CATALYST**, a metacognition-driven framework that systematically guides interdisciplinary scientific ideation through problem decomposition, cross-domain exploration, and strategic prioritization.
- We introduce a structured dataset and evaluation framework for benchmarking interdisciplinary research ideation across novelty, insightfulness, relevance, and usefulness.
- We show through LLM-based and human evaluations that IDEA-CATALYST produces 21.38% more novel and 16.22% insightful ideas, while remaining grounded in the target research problem.

2 Related Work

Interdisciplinary Research and Scientific Creativity Prior science-of-science work shows that interdisciplinary research, especially across distant fields, can yield larger and longer-term impact (Van Noorden, 2015; Okamura, 2019). Creative-science studies similarly emphasize that breakthroughs often emerge through the gradual recombination of partial ideas rather than isolated insights (Sosa, 2019; Gonçalves and Cash, 2021; Yanai and Lercher, 2019). Yet deeply integrative cross-domain research remains rare (Porter and Rafols, 2009; Raasch et al., 2013), partly because researchers must identify relevant external domains and translate their concepts back into target-domain problems.

Automated Scientific Discovery and Research Ideation LLMs have enabled AI-assisted literature review, hypothesis generation, experimental planning, and code-based scientific agents (Gottweis et al., 2025; Goel et al., 2025; Si et al., 2026; Kargupta et al., 2025a,d; Jansen et al., 2025). Systems such as SCIMON (Wang et al., 2024) and benchmarks such as IdeaBench (Guo et al., 2025) study literature-grounded ideation, while prior evaluations show that LLM ideas can be novel yet insufficiently grounded or feasible (Si et al., 2024). Many systems further couple ideation to execution or early evaluation, which can improve feasibility but bias exploration toward incremental, single-domain refinements (Si et al., 2026). IDEA-CATALYST instead targets the earlier exploratory stage, before ideas are forced into executable plans.

Human-Centered Scientific Knowledge Discovery Human-centered discovery tools use LLMs to support literature exploration, question formulation, and iterative refinement (Kargupta et al., 2025c). IdeaSynth (Pu et al., 2025) and DisciplLink (Zheng et al., 2024), for example, show the value of interactive support for exploratory and interdisciplinary thinking. However, such systems often rely on user-provided or parametric domain suggestions, which can favor nearby or familiar fields. IDEA-CATALYST complements this work by decomposing target problems, abstracting unresolved conceptual challenges, and using those abstractions to guide distant source-domain retrieval. Appendix A provides additional context.

3 Methodology

IDEA-CATALYST aims to augment early-stage scientific ideation by (a) decomposing research problems into core questions, (b) identifying unresolved conceptual challenges in the target domain, (c) extracting insights from external source domains which address these challenges, and (d) integrating them into an interdisciplinary idea fragment. The overall framework is illustrated in Figure 2. Full methodology prompt templates are provided in Appendix H.

3.1 Preliminaries

Given a short research problem p in a *target domain* $\mathcal{D}_{\text{target}}$ (e.g., Natural Language Processing), our objective is to produce a ranked set of interdisciplinary idea fragments $\mathcal{F}$. Each fragment f_i links an unresolved target-domain challenge to literature-grounded insights from an external *source domain* $\mathcal{D}_{s_i}$ and recontextualizes those insights as an early-stage research direction. We restrict source domains to coarse fields sufficiently distant from the target domain to avoid trivial subfield transfers; formal definitions are in Appendix B.

To ground ideation in existing work, we retrieve short passages with the Semantic Scholar

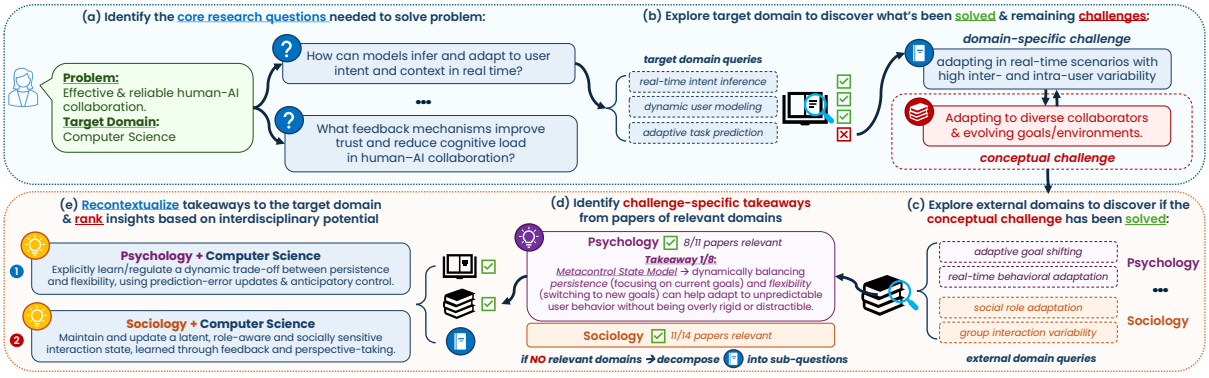

Figure 2: **IDEA-CATALYST** is a metacognition-driven framework that *(a)* analyzes target-domain progress, *(b)* identifies unresolved challenges, *(c)* explores source domains for analogous insights, and *(d)* integrates them into interdisciplinary idea fragments. The figure is illustrated with a real case study (summarized).

Snippets API[2] (Kinney et al., 2023). For each natural-language query and coarse field of study (Appendix I), we retrieve the top-k papers, aggregate available snippets, and fall back to abstracts when snippets are unavailable or degenerate. This black-box retrieval design keeps the contribution focused on structuring interdisciplinary reasoning rather than optimizing search infrastructure.

3.2 Metacognition-Driven Ideation

Scientific research ideation is inherently creative, requiring ideas to be both novel and useful (Yanai and Lercher, 2019). We model it as a metacognitive process: monitoring what is known, recognizing domain assumptions, selecting appropriate search strategies, managing intermediate goals, and evaluating whether emerging insights address unresolved challenges (Flavell, 1979; Yeung and Summerfield, 2012; Kitchner, 1983; Urban and Urban, 2025; Kargupta et al., 2025b). Under this view, IDEA-CATALYST coordinates *critical reasoning* over the target-domain literature with *creative reasoning* over distant source domains (Johnson-Laird, 2010; de Chantal and Markovits, 2022; Dwyer et al., 2025; Wechsler et al., 2018; Halpern, 2007). Appendix C gives concrete intermediate outputs.

3.3 Critical Reasoning over the Target Domain

IDEA-CATALYST first decomposes p into research questions $\mathcal{Q}$ and represents each question in both target-domain language ($q_i^{\mathcal{D}}$) and a domain-agnostic form (q_i'). Target-domain retrieval then estimates whether each question is largely resolved, partially addressed, or open, and extracts unresolved conceptual challenges for later cross-domain search. This dual representation preserves problem-specific constraints while making conceptual analogies easier to find; examples are shown in Appendix C.

3.4 Creative Reasoning Across Source Domains

IDEA-CATALYST next explores source domains using the domain-agnostic forms of open questions and unresolved challenges. For each selected challenge q', the model proposes one to three candidate source domains based on analogy, shared mechanisms, or transferable principles, excluding domains that are too proximal to $\mathcal{D}_{\text{target}}$. It then generates domain-specific search queries, retrieves snippets, and extracts takeaways only when the majority of retrieved papers are relevant. If no source domain passes this threshold, the system exposes the relevancy scores and, for benchmark consistency, falls back to the most relevant domain; this occurred in only ∼1.75% of samples. Appendix C provides an example takeaway.

3.5 Target–Source Interdisciplinary Integration

Finally, IDEA-CATALYST integrates target-domain challenges, source-domain takeaways, and retrieved evidence into ranked idea fragments. Each fragment states the target challenge, the source-domain mechanism, a synthesis rationale, and a concrete but intentionally incomplete realization intended to seed further brainstorming rather than prescribe an executable study (format in Appendix F). We rank fragments via pairwise comparisons of interdisciplinary potential, prioritizing depth of integration, multi-stage disciplinary engagement, in-

[2] https://api.semanticscholar.org/api-docs/snippets

novation payoff, and the balance between novelty and plausibility (Okamura, 2019; Porter and Rafols, 2009).

4 Experimental Design

We choose `Qwen3-14B` (Yang et al., 2025) as our primary model for experiments (`no-thinking` for efficiency, temperature = 0.7), and gpt-oss-120b (temperature = 0.0) (Agarwal et al., 2025) for our LLM judge. We retrieve a maximum of 20 papers per round of retrieval and prune source domains where the majority (50%) of papers are irrelevant to the research problem.

4.1 Dataset

We evaluate IDEA-CATALYST using the CHIMERA dataset (Sternlicht and Hope, 2025), a collection of interdisciplinary research papers drawn from arXiv with annotated *inspiration relations* between source and target domains. Each instance links a target-domain contribution (`target_text`) to a distinct source-domain inspiration (`source_text`), making it well suited for studying cross-domain knowledge transfer and interdisciplinary ideation. We select 400 instances where the source and target domains belong to different coarse-grained scientific fields, the annotated relation is *inspiration*, both domains are explicitly specified, and the annotated problem context (which serves as our input p) does not leak the source insights. Although the benchmark is naturally skewed toward CS/AI, 16% of selected instances are non-CS and 40% fall outside AI/ML/NLP/CV target areas, providing broader coverage than an AI-only evaluation. We further prevent knowledge leakage by restricting retrieval (Section 3.1) to papers published strictly before the arXiv posting year of each instance.

To enable direct comparison between generated idea fragments and CHIMERA references, we preprocess each sample into a structured representation aligned with our framework's output format (Appendix F) using `Qwen3-14B` (`no-thinking`, `temperature`= 0). The model is strictly constrained to reorganize only information already present in the abstract and annotated context while ignoring experimental results, yielding a controlled insight-level reference rather than a polished-paper target. For the human study, participants provide a brief description of a real research problem and target domain.

4.2 Baselines

We compare IDEA-CATALYST against two baselines that reflect common LLM-driven approaches to ideation with increasing degrees of retrieval structure: (a) ***Free-Form Source Retrieval*** (Zheng et al., 2024) prompts the model to directly identify source domains without constraints on distance to the target domain, retrieve related literature, and synthesize ideas, without explicit target-domain analysis or problem decomposition; (b) ***Guided Dual-Retrieval*** first retrieves representative target-domain literature and then conditions cross-domain retrieval (with explicit distance constraints) and ideation on this context, but does not explicitly identify unresolved conceptual challenges, construct domain-agnostic abstractions, or strategically guide source-domain selection. Both effectively serve as retrieve-then-ideate pipelines without metacognitive control.

Ablations We evaluate three ablations to assess the contribution of individual components of IDEA-CATALYST: $\times$ *Decomposition*, which removes explicit problem decomposition and target-domain retrieval; $\times$ *Interdisciplinary Ranking*, which replaces the potential-based ranking with the proportion of relevant source papers; and $+$ *Conceptual Rewriting*, which rewrites final outputs for further clarity without altering technical content or structure. Together, these ablations isolate the effects of structured target-domain analysis, comparative interdisciplinary potential-based ranking, and articulation quality. We include further details in Appendix G.

4.3 Evaluation Metrics

Although evaluation for creative ideation in scientific discovery is inherently subjective and requires high levels of domain expertise (Si et al., 2025), LLMs have shown strong capabilities in evaluating along various dimensions, while aligning with human judgements and capturing nuanced feedback signals (Lee et al., 2023; Madaan et al., 2023; Mehri et al., 2025; Mehri and Shwartz, 2023). In particular, LLM judges have proven effective in evaluating creativity, demonstrating alignment with human judgments across diverse creativity dimensions and research ideation tasks (Si et al., 2024; Hou et al., 2025; Zhao et al., 2025).

We follow established practices in preference-based evaluation (Dubois et al., 2024; Zheng et al., 2023) and compare each generated output against

Table 1: Takeaway-level average win rates at top-k ($k \in \{1, 2, 3\}$). Bold indicates best; † denotes second-highest.

Method	Insightfulness			Relevance			Overall		
	@1	@2	@3	@1	@2	@3	@1	@2	@3
Free-Form Source	18.25	23.06	27.35	47.75	51.45	51.57	44.25	48.06	50.14
Guided Dual	73.00	74.19	72.36	**62.25**	59.68	60.54	**66.75**	63.23	64.39
IDEA-CATALYST	**85.50**	**85.16**	84.47†	60.25	**62.42**	61.25	63.75	66.45†	65.67
× Decompose	84.00†	84.88†	83.80	60.75†	61.63	**65.32**	65.25†	**66.82**	**70.13**
× Potential Ranking	84.71	84.28	**84.83**	59.40	59.97	61.13	64.16	64.18	65.03
+ Conceptual Rewriting	82.00	83.71	83.19	59.75	61.94†	62.82†	66.25	67.10	66.95†

Table 2: Idea-level average win rates at top-k ($k \in \{1, 2, 3\}$). Bold indicates best; † denotes second-highest.

Method	Novelty			Usefulness			Overall		
	@1	@2	@3	@1	@2	@3	@1	@2	@3
Free-Form Source	13.50	17.26	19.80	35.50	39.84	40.31	30.50	34.52	35.19
Guided Dual	68.00	70.32	67.95	**70.25**	65.48†	64.39	**72.25**	68.39	66.10
IDEA-CATALYST	**83.25**	**84.03**	**83.05**	65.75†	**66.13**	**66.38**	71.25†	**70.65**	**70.09**
× Decompose	82.00†	81.94	78.48	61.00	63.43	65.06†	63.75	65.91	67.85†
× Potential Ranking	83.21†	83.31†	81.79†	62.41	65.15†	64.16	68.17	69.04†	67.92
+ Conceptual Rewriting	82.00	82.90	80.91	62.75	62.90	61.54	66.75	66.29	64.96

a controlled reference derived from the CHIMERA paper abstract, not against the full published paper. This reference isolates the paper's interdisciplinary insight while filtering downstream experimental results using the same base model and schema, making the comparison better aligned with early-stage brainstorming. Because creative ideation involves a known novelty/usefulness trade-off, we intentionally disentangle novelty, insightfulness, relevance, and usefulness; "overall" is reported as a diagnostic joint preference rather than the sole success criterion. Prompt details, human-LLM agreement, robustness checks, and cost analysis are in Appendix E.

To address the subjectivity of LLM-as-judge evaluation, we further conduct a blind human–LLM agreement study over 90 randomly sampled judgments (15 problems, 98 takeaways, 30 ideas). Agreement is strongest on the creativity-oriented dimensions: 93.3% for takeaway insightfulness ($\kappa = 0.867$), 93.3% for idea novelty ($\kappa = 0.867$), and 86.7% for idea usefulness ($\kappa = 0.732$). Relevance and overall judgments show lower but still meaningful agreement, consistent with our observation that judges trade off novelty and immediate applicability differently. We evaluate each sample along two dimensions:

(1) Takeaways. We assess the quality of source-domain insights extracted by our approach. Specifically, we evaluate takeaways based on (1) interdisciplinary insightfulness, which measures whether the takeaways introduce specific, non-obvious concepts or frameworks from the source domain that are intellectually interesting to researchers in the target domain, and (2) interdisciplinary relevance, which assesses whether the takeaways have strong potential to inspire new approaches or address gaps in the target domain.

(2) Idea. We assess the quality of the final generated idea that integrates interdisciplinary source-domain insights for the target domain. Each idea is evaluated based on (1) interdisciplinary novelty, which measures whether the novelty of the idea, and (2) interdisciplinary usefulness, which measures which idea has greater potential for addressing the research problem in the target domain.

The complete evaluation prompts are provided in Appendix K. This evaluation approach allows us to assess the practical utility of IDEA-CATALYST for supporting early-stage research ideation and provides a methodology that can be adopted for evaluating creativity in idea generation in future work.

5 Experimental Results

Overall Performance & Analysis. Across both evaluation levels, IDEA-CATALYST consistently outperforms baseline approaches on dimensions associated with exploratory and creative ideation. At the takeaway level (Table 1), it achieves the highest insightfulness scores across all top-k settings, corresponding to an average relative improvement of **16.22%** over *Guided Dual* and **282.21%** over *Free-Form Source* when averaged across $k \in \{1, 2, 3\}$.

Similarly, at the idea level (Table 2), IDEA-CATALYST attains the strongest novelty scores, yielding an average relative gain of **21.38%** over *Guided Dual* and **407.65%** over *Free-Form Source*. This suggests that metacognition-driven ideation–strategically guided creative exploration grounded in target-domain critical reasoning–enables the model to surface ideas and takeaways that are more novel and insightful than purely retrieval-based baselines. We also observe a trade-off between exploratory creativity and grounded problem applicability: the interdisciplinary potential-based ranking generally favors novelty and insightfulness, with relevance and usefulness increasing as k grows. When these criteria conflict, the overall LLM-judged score tends to prioritize relevance and usefulness, explaining why baselines can remain competitive on overall metrics despite substantially lower novelty or insightfulness.

Robustness and Novelty Checks. In direct method-to-method comparison on a 50% subset, IDEA-CATALYST is preferred over *Free-Form Source* for novelty (95.6%) and insightfulness (95.6%), and over *Guided Dual* for novelty (81.53%) and insightfulness (83.2%). A stronger Qwen3.5-27B judge preserves the main trend at idea-level @1, with IDEA-CATALYST scoring 84.8 novelty versus 13.0 and 61.8 for the two baselines. To test whether source-domain "novelty" merely rephrases concepts already present in the target domain, we retrieve target-domain papers for each problem plus top-ranked source insight and judge conceptual overlap independent of terminology. IDEA-CATALYST achieves the highest no-overlap rate (85.0%, vs. 66.0% and 72.5%), with a human audit showing substantial agreement ($\kappa = 0.8$). Appendix E provides full details, including Llama-3.1-8B backbone results and cost analysis.

Ablation Studies. Ablation results further highlight the contribution of each component in IDEA-CATALYST. Removing target-domain decomposition reduces novelty and insightfulness, even when some overall scores improve because overall judgments tend to reward immediate relevance/usefulness over exploratory novelty. Replacing potential-based ranking with relevance-based heuristics similarly lowers the creativity-oriented metrics, emphasizing that source-domain relevance alone is insufficient for surfacing high-impact interdisciplinary ideas. Conceptual rewriting improves clarity but not quantitative scores, suggesting that the gains stem from deeper interdisciplinary integration rather than presentation alone.

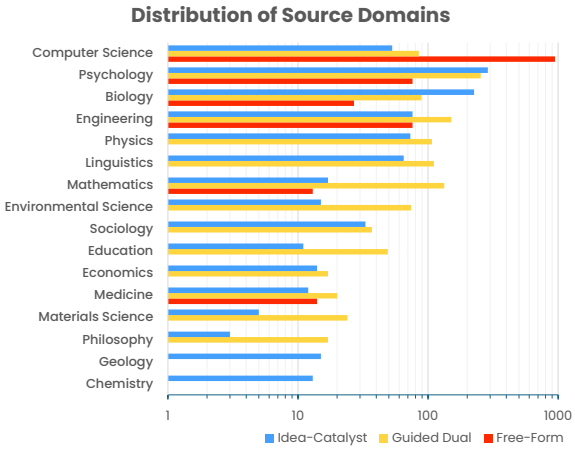

Figure 3: Source-domain distributions (log-scale) for each method's top three ideas.

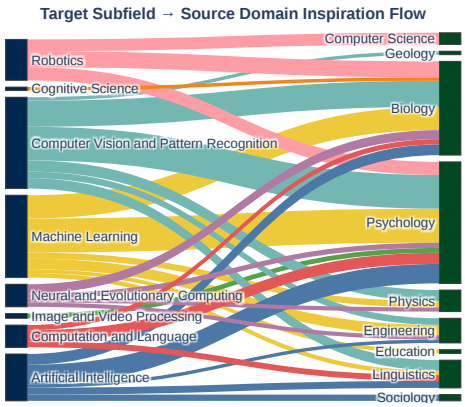

Figure 4: Target-source flow of interdisciplinary inspiration.

5.1 Source Domain Distribution Analysis

We analyze the distribution of source domains selected by each method by aggregating the top-3 ideas per problem and filtering out domains with fewer than 10 occurrences (Figure 3). The *Free-Form Source Retrieval* baseline exhibits a severe skew toward *Computer Science* (947 occurrences), resulting in very low domain diversity (normalized entropy $H_{\text{norm}} = 0.326$), indicating that unconstrained LLM-driven ideation tends to remain within the target domain's immediate conceptual neighborhood even when encouraged to explore externally. In contrast, *Guided Dual-Retrieval* achieves the highest overall spread across domains such as Psychology, Engineering, Mathematics, Linguistics, and Physics ($H_{\text{norm}} = 0.812$), although it also favors closer domains, such as Computer Science and Engineering (19.67% of its

source domains vs. IDEA-CATALYST's 10.75%). IDEA-CATALYST exhibits broad cross-domain exploration spanning Psychology, Biology, Physics, Linguistics, Engineering, and additional scientific fields ($H_{\text{norm}} = 0.682$), while consistently achieving competitive or higher relevance/usefulness alongside substantially stronger novelty/insightfulness (Table 1 & 2). Overall, this suggests that effective interdisciplinary ideation depends not only on increasing domain diversity, but on critically selecting source domains which yield meaningful conceptual insights.

Target–Source Flow of Inspiration Figure 4 visualizes the flow of interdisciplinary inspiration between target subfields and source domains, restricted to target–source pairs occurring at least 10 times (top-10 sources per target) to highlight stable patterns. *Psychology* emerges as the most prevalent source across many AI-related targets, reflecting its foundational role in cognition, decision-making, and human–AI interaction. We also observe intuitive alignments, such as *Neural and Evolutionary Computing* drawing from *Biology*, and *Artificial Intelligence* sourcing from both *Psychology* and *Linguistics*. Overall, the diagram shows that IDEA-CATALYST surfaces diverse yet intuitive cross-domain influences, with different Computer Science subfields drawing from complementary external disciplines rather than a single dominant source.

5.2 Qualitative Comparison of Takeaways

Table 6 in Appendix D presents a qualitative comparison of the top-ranked source-domain takeaways selected by *Guided Dual* and IDEA-CATALYST for the problem of human–AI collaboration in open-ended tasks. While both methods identify *Psychology* as the most relevant source domain, *Guided Dual* selects a broad *Theory of Mind* framing, whereas IDEA-CATALYST surfaces a more targeted takeaway centered on *reciprocal information flow* and *dynamic role distribution*. This contrast suggests that metacognitive guidance improves not only source-domain selection, but also the precision of extracted interdisciplinary insights.

5.3 Human Study

We conducted a complementary human study with six PhD researchers in Machine Learning, Natural Language Processing, and Electrical Engineering, each using a real problem from their own work (details in Appendix J). Participants rated generated research questions as relevant (4.00/5) and retrieved papers as aligned with their interests (3.50/5). Takeaways and final ideas were more moderate but still useful for exploration: takeaways averaged 3.13/5 relevance and 3.16/5 insightfulness, while ideas were rated more novel (3.22/5) than immediately useful (3.00/5). This supports our creativity-first framing: IDEA-CATALYST is most helpful for surfacing new conceptual directions, while concrete recontextualization and concision remain bottlenecks for future human-in-the-loop refinement.

Practical Efficiency. The pipeline averages about 30 LLM calls per problem (roughly 140K input tokens and 120K maximum output tokens), corresponding to approximately $0.25 per problem using GPT-4.1-mini pricing. Runtime scales linearly in decomposed questions, unresolved challenges, and candidate source domains, with only the final pairwise idea ranking quadratic in the number of generated fragments.

6 Conclusion

In this work, we introduced IDEA-CATALYST, a metacognition-driven framework for interdisciplinary research ideation that supports targeted, creative exploration across scientific domains. By explicitly analyzing target-domain challenges and strategically guiding cross-domain inspiration, IDEA-CATALYST surfaces ideas and insights that are significantly more novel and insightful than existing baselines, while remaining grounded in the original research problem. Our results suggest that supporting the *process* of interdisciplinary ideation is a promising direction for human–AI collaboration in scientific research. Future work includes developing personalized summarization strategies tailored to researchers' backgrounds, as well as leveraging interdisciplinary signals to recommend potential collaborators across domains.

7 Limitations

While IDEA-CATALYST is designed to support early-stage interdisciplinary ideation, it remains limited by the quality and scope of retrieved literature and the granularity of coarse field categories. Our retrieval pipeline depends on Semantic Scholar snippets and abstracts, which can omit important details or introduce bias toward papers with stronger snippet coverage. The generated idea fragments are intended as conceptual starting points

rather than fully validated research plans, so they require expert refinement, feasibility assessment, and task-specific grounding before being pursued.

Our evaluation emphasizes novelty, insightfulness, and relevance, but does not directly measure downstream research impact or practical implementation success. The human study is small and focused on a handful of PhD researchers, so further validation with larger, more diverse researcher populations and domain-specific experts would be needed to generalize the findings. Finally, the cost and latency of LLM-based retrieval and reasoning currently limit the framework's scalability for very large problem collections, suggesting future work should explore more efficient retrieval and summarization strategies.

References

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, and 1 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Mousumi Bose, Judith Anne Garretson Folse, and Sooyeon Nikki Lee-Wingate. 2013. Enhancing the influence of distal primes on creativity: the role of contextual and personal variables. *Journal of Marketing Theory and Practice*, 21(4):351–370.

Sofía Castro, Marcin Bukowski, Juan Lupiáñez, and Zofia Wodniecka. 2021. Fast or accurate? the change of goals modulates the efficiency of executive control. *Polish Psychological Bulletin*, 52(1):49–66.

Ed Catmull and Amy Wallace. 2023. *Creativity, Inc.(The Expanded Edition): Overcoming the unseen forces that stand in the way of true inspiration*. Random House.

Pier-Luc de Chantal and Henry Markovits. 2022. Reasoning outside the box: Divergent thinking is related to logical reasoning. *Cognition*, 224:105064.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.

Christopher P Dwyer, Deaglán Campbell, and Niall Seery. 2025. An evaluation of the relationship between critical thinking and creative thinking: Complementary metacognitive processes or strange bedfellows? *Journal of Intelligence*, 13(2):23.

John H Flavell. 1979. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist*, 34(10):906.

Shashwat Goel, Rishi Hazra, Dulhan Jayalath, Timon Willi, Parag Jain, William F Shen, Ilias Leontiadis, Francesco Barbieri, Yoram Bachrach, Jonas Geiping, and 1 others. 2025. Training ai co-scientists using rubric rewards. *arXiv preprint arXiv:2512.23707*.

Milene Gonçalves and Philip Cash. 2021. The life cycle of creative ideas: Towards a dual-process theory of ideation. *Design Studies*, 72:100988.

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

Ivan Grahek, Xiamin Leng, S Musslick, and A Shenhav. 2023. The cost of adjusting cognitive control: A dynamical systems approach. In *Conference on Cognitive Computational Neuroscience*.

Sikun Guo, Amir Hassan Shariatmadari, Guangzhi Xiong, Albert Huang, Myles Kim, Corey M Williams, Stefan Bekiranov, and Aidong Zhang. 2025. Ideabench: Benchmarking large language models for research idea generation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 5888–5899.

Tarun Gupta and Danish Pruthi. 2025. All that glitters is not novel: Plagiarism in ai generated research. *arXiv preprint arXiv:2502.16487*.

Diane F Halpern. 2007. The nature and nurture of critical thinking. *Critical thinking in psychology*, (1):1–14.

Zhaoyi Joey Hou, Bowei Alvin Zhang, Yining Lu, Bhiman Kumar Baghel, Anneliese Brei, Ximing Lu, Meng Jiang, Faeze Brahman, Snigdha Chaturvedi, Haw-Shiuan Chang, and 1 others. 2025. Creativityprism: A holistic benchmark for large language model creativity. *arXiv preprint arXiv:2510.20091*.

Jin Huang, Silviu Cucerzan, Sujay Kumar Jauhar, and Ryen W White. 2025. Idea2plan: Exploring ai-powered research planning. *arXiv preprint arXiv:2510.24891*.

Peter Jansen, Oyvind Tafjord, Marissa Radensky, Pao Siangliulue, Tom Hope, Bhavana Dalvi, Bodhisattwa Prasad Majumder, Daniel S Weld, and Peter Clark. 2025. Codescientist: End-to-end semi-automated scientific discovery with code-based experimentation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13370–13467.

Philip N Johnson-Laird. 2010. Mental models and human reasoning. *Proceedings of the National Academy of Sciences*, 107(43):18243–18250.

Priyanka Kargupta, Ishika Agarwal, Tal August, and Jiawei Han. 2025a. Tree-of-debate: Multi-persona debate trees elicit critical thinking for scientific comparative analysis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 29378–29403, Vienna, Austria. Association for Computational Linguistics.

Priyanka Kargupta, Shuyue Stella Li, Haocheng Wang, Jinu Lee, Shan Chen, Orevaoghene Ahia, Dean Light, Thomas L Griffiths, Max Kleiman-Weiner, Jiawei Han, and 1 others. 2025b. Cognitive foundations for reasoning and their manifestation in llms. *arXiv preprint arXiv:2511.16660*.

Priyanka Kargupta, Runchu Tian, and Jiawei Han. 2025c. Beyond true or false: Retrieval-augmented hierarchical analysis of nuanced claims. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29664–29679, Vienna, Austria. Association for Computational Linguistics.

Priyanka Kargupta, Nan Zhang, Yunyi Zhang, Rui Zhang, Prasenjit Mitra, and Jiawei Han. 2025d. TaxoAdapt: Aligning LLM-based multidimensional taxonomy construction to evolving research corpora. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29834–29850, Vienna, Austria. Association for Computational Linguistics.

Rodney Michael Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, Miles Crawford, Doug Downey, Jason Dunkelberger, Oren Etzioni, Rob Evans, Sergey Feldman, Joseph Gorney, David W. Graham, F.Q. Hu, and 29 others. 2023. The semantic scholar open data platform. *ArXiv*, abs/2301.10140.

Karen Strohm Kitchner. 1983. Cognition, metacognition, and epistemic cognition: A three-level model of cognitive processing. *Human development*, 26(4):222–232.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Ren Lu, Thomas Mesnard, Johan Ferret, Colton Bishop, Ethan Hall, Victor Carbune, and Abhinav Rastogi. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *International Conference on Learning Representations*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.

Shuhaib Mehri, Xiusi Chen, Heng Ji, and Dilek Hakkani-Tür. 2025. Beyond sample-level feedback: Using reference-level feedback to guide data synthesis. *arXiv preprint arXiv:2502.04511*.

Shuhaib Mehri and Vered Shwartz. 2023. Automatic evaluation of generative models with instruction tuning. *arXiv preprint arXiv:2310.20072*.

Ciarán O'Driscoll, Aneesha Singh, Iya Chichua, Joachim Clodic, Anjali Desai, Dara Nikolova, Alex Jie Yap, Irene Zhou, and Stephen Pilling. 2024. An ecological mobile momentary intervention to support dynamic goal pursuit: Feasibility and acceptability study. *JMIR Formative Research*, 8:e49857.

Keisuke Okamura. 2019. Interdisciplinarity revisited: evidence for research impact and dynamism. *Palgrave Communications*, 5(1).

Alan Porter and Ismael Rafols. 2009. Is science becoming more interdisciplinary? measuring and mapping six research fields over time. *scientometrics*, 81(3):719–745.

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

Christina Raasch, Viktor Lee, Sebastian Spaeth, and Cornelius Herstatt. 2013. The rise and fall of interdisciplinary research: The case of open source innovation. *Research policy*, 42(5):1138–1151.

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

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

Chenglei Si, Zitong Yang, Yejin Choi, Emmanuel Candès, Diyi Yang, and Tatsunori Hashimoto. 2026. Towards execution-grounded automated ai research. *arXiv preprint arXiv:2601.14525*.

Ricardo Sosa. 2019. Accretion theory of ideation: Evaluation regimes for ideation stages. *Design Science*, 5:e23.

Noy Sternlicht and Tom Hope. 2025. Chimera: A knowledge base of idea recombination in scientific literature. *Preprint*, arXiv:2505.20779.

Richard S Sutton, Andrew G Barto, and 1 others. 1998. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.

Marek Urban and Kamila Urban. 2025. Do we need metacognition for creativity? a necessary condition analysis of creative metacognition. *Psychology of Aesthetics, Creativity, and the Arts*, 19(6):1467.

Richard Van Noorden. 2015. Interdisciplinary research by the numbers. *Nature*, 525(7569):306–307.

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

Solange Muglia Wechsler, Carlos Saiz, Silvia F Rivas, Claudete Maria Medeiros Vendramini, Leandro S Almeida, Maria Celia Mundim, and Amanda Franco. 2018. Creative and critical thinking: Independent or overlapping components? *Thinking skills and creativity*, 27:114–122.

Itai Yanai and Martin Lercher. 2019. Night science. *Genome Biology*, 20(1):179.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Nick Yeung and Christopher Summerfield. 2012. Metacognition in human decision-making: confidence and error monitoring. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 367(1594):1310–1321.

Yunpu Zhao, Rui Zhang, Wenyi Li, and Ling Li. 2025. Assessing and understanding creativity in large language models. *Machine Intelligence Research*, 22(3):417–436.

Chengbo Zheng, Yuanhao Zhang, Zeyu Huang, Chuhan Shi, Minrui Xu, and Xiaojuan Ma. 2024. Disciplink: unfolding interdisciplinary information seeking process via human-ai co-exploration. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, pages 1–20.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

A Extended Related Work

Interdisciplinary Research and Scientific Creativity Prior work in the science-of-science literature has shown that interdisciplinary research is a key driver of scientific innovation and long-term impact, with ideas that integrate concepts across distant fields often achieving substantially higher influence (Van Noorden, 2015; Okamura, 2019). Studies of creative processes further emphasize that scientific breakthroughs typically emerge through the gradual accumulation and recombination of partial ideas rather than isolated insights (Sosa, 2019; Gonçalves and Cash, 2021), often drawing from multiple conceptual domains (Yanai and Lercher, 2019). Despite these benefits, deeply integrative interdisciplinary research remains rare and fragile (Porter and Rafols, 2009; Raasch et al., 2013), in part because identifying which external domains are meaningfully relevant—and how their ideas can be translated into a target domain—poses a substantial cognitive and practical challenge for individual researchers.

Automated Scientific Discovery and Research Ideation Recent advances in large language models have spurred work on AI-assisted scientific discovery, including automated literature review, hypothesis generation, research ideation, and experimental planning (Gottweis et al., 2025; Goel et al., 2025; Si et al., 2026; Kargupta et al., 2025a,d; Jansen et al., 2025). Systems such as SCIMON (Wang et al., 2024) and benchmarks like IdeaBench (Guo et al., 2025) explore literature-grounded idea generation and evaluation, finding that while LLMs tend to produce highly novel ideas, they often lack technical depth, feasibility, or strong grounding in concrete research challenges (Si et al., 2024). Moreover, many automated approaches tightly couple ideation with execution or early evaluation, which can bias exploration toward incremental, single-domain refinements and reduce cross-domain creativity (Si et al., 2026). In contrast, IDEA-CATALYST explicitly avoids premature execution and instead focuses on supporting early-stage exploratory ideation through structured target-domain analysis and strategically guided interdisciplinary retrieval.

Human-Centered Scientific Knowledge Discovery Complementary to fully automated methods, human-centered systems leverage LLMs to support researchers in literature exploration, question formulation, and iterative refinement (Kargupta et al., 2025c). Tools such as IdeaSynth (Pu et al., 2025) and DiscipLink (Zheng et al., 2024) demonstrate that interactive, human-in-the-loop approaches can effectively support exploratory thinking and interdisciplinary information seeking. However, these systems typically rely on either user input or the LLM's parametric knowledge to suggest relevant domains, which can favor nearby or familiar fields and overlook deeper conceptual analogies across distant disciplines. IDEA-CATALYST complements this line of work by explicitly decomposing target-domain problems, abstracting persistent conceptual challenges, and strategically guiding cross-domain exploration.

B Formal Definitions

Definition B.1 (*Target Domain*). The *target domain* $\mathcal{D}_{\text{target}}$ denotes the primary scientific field in which the research problem p is situated. It is characterized by its literature, methodologies, problem formulations, and evaluation norms.

Definition B.2 (*Source Domain*). A *source domain* $\mathcal{D}_s$ is a scientific field distinct from $\mathcal{D}_{\text{target}}$ that provides potentially transferable conceptual frameworks, mechanisms, or abstractions. We exclude closely related subfields to promote non-trivial interdisciplinary transfer.

Definition B.3 (*Interdisciplinary Insight*). An *interdisciplinary insight* $t \in T_{s_i}$ is a literature-grounded conceptual takeaway extracted from a source domain $\mathcal{D}_{s_i}$ that may become relevant when mapped back onto the research problem p.

Definition B.4 (*Interdisciplinary Potential*). *Interdisciplinary potential* is the expected value of an idea fragment for advancing p through cross-domain integration. It reflects whether the fragment addresses unresolved target-domain challenges, introduces non-trivial source-domain perspectives, and plausibly inspires new research directions.

C Pipeline Examples

Domain-Specific Question	Domain-Agnostic Question
How can models dynamically infer and adapt to user intent and task context in real-time collaboration?	*How can understanding of intent and context be updated through continuous interaction?*
How should a system decide when to take initiative versus defer to the human?	*When should control be exercised versus withheld?*

Table 3: Example question abstractions for human–AI collaboration.

Domain-Specific Challenge	Domain-Agnostic Challenge
How can a system adapt in real time to high inter- and intra-user variability?	*How can behavior adapt to diverse collaborators and evolving goals or environments?*

Table 4: Example unresolved challenge extracted from target-domain analysis.

D Source-Domain and Qualitative Analysis

Table 6 provides the exact outputs discussed in the main qualitative comparison.

Source-Domain Concept	Transferable Logic
Metacontrol State Model: Goal-directed behavior balances persistence and flexibility.	Dynamic regulation between persistence and flexibility enables behavior to remain focused while adapting to changing goals and environments (Castro et al., 2021).
Cognitive Control: Control is prospectively adjusted based on expected goal switches.	Anticipatory control can reduce switching costs and cognitive load under frequent change (Grahek et al., 2023).
Dynamic Goal Pursuit: Just-in-time interventions monitor behavior and provide context-sensitive support.	Real-time adaptive feedback can align support with evolving goals and situational demands (O'Driscoll et al., 2024).

Table 5: Example Psychology takeaways for the challenge in Table 4.

E Robustness Analysis

Human–LLM Agreement We conducted a blind agreement analysis over 90 randomly sampled judgments covering 15 research problems, 98 takeaways, and 30 ideas. For takeaways, agreement was 93.3% for insightfulness ($\kappa = 0.867$), 73.3% for relevance ($\kappa = 0.455$), and 80.0% overall ($\kappa = 0.595$). For ideas, agreement was 93.3% for novelty ($\kappa = 0.867$), 86.7% for usefulness ($\kappa = 0.732$), and 80.0% overall ($\kappa = 0.587$). This indicates strong alignment on novelty, insightfulness, and usefulness, with relevance and overall judgments reflecting the expected creativity/usefulness trade-off.

Conceptual Overlap Verification To test whether "novel" source-domain concepts had already appeared in the target domain under different terminology, we retrieved the top 10 target-domain papers for each problem plus top-ranked source insight and asked an LLM judge to assess conceptual overlap. The no-overlap rates were 66.0% for *Free-Form Source*, 72.5% for *Guided Dual*, and 85.0% for IDEA-CATALYST. A human audit of 20 sampled overlap/non-overlap cases achieved 90% accuracy and Cohen's $\kappa = 0.8$, suggesting substantial agreement.

Additional Judges, Backbones, and Direct Comparisons Using a stronger Qwen3.5-27B judge, IDEA-CATALYST remains best at idea-level @1: novelty 84.8 versus 13.0 and 61.8 for *Free-Form Source* and *Guided Dual*; usefulness 80.5 versus 13.5 and 58.0; and overall 80.5 versus 13.5 and 58.0. With Llama-3.1-8B-Instruct as the backbone on a random subset of 50 examples, relative gains in novelty and insightfulness persist, although absolute scores vary with model capacity. In di-

Table 6: **Qualitative Comparison of Source-Domain Takeaways.** Exact outputs taken from *Guided Dual* and IDEA-CATALYST.

Method	Source Domain Framing	Applicability to Target Domain
Guided Dual	Theory of Mind (ToM) and its role in predicting others' mental states	By equipping AI with ToM capabilities, the system can better understand and predict user intentions, leading to more natural and effective collaboration. This reduces cognitive load by aligning AI behaviors with user expectations and improving task efficiency.
IDEA-CATALYST	Reciprocal information flow and role distribution enhance joint action coordination by allowing individuals to dynamically assign and shift roles based on task demands and the predictability of others' actions.	When individuals engage in reciprocal information flow, they can dynamically assign roles (e.g., leader–follower) and adjust their strategies in real time based on the actions and predictability of their partner. This enables them to adapt to complex and undefined tasks without predefined boundaries.

rect method-to-method comparison on a 50% subset, IDEA-CATALYST is preferred over *Free-Form Source* for novelty (95.6%), usefulness (66.4%), and insightfulness (95.6%), and over *Guided Dual* for novelty (81.53%) and insightfulness (83.2%).

Cost and Scaling Let Q be the number of decomposed questions, U the number of unresolved questions/challenges selected for cross-domain search, D the candidate source domains per challenge, and I the number of generated idea fragments. The LLM-call count scales as $C = 1 + Q + U + U \cdot D + I + I(I-1)/2$, where only the final pairwise ranking term is quadratic in I. Empirically, the pipeline averages approximately 30 calls per problem, about 140K input tokens and 120K maximum output tokens, corresponding to roughly $0.25 per problem using GPT-4.1-mini pricing.

F Idea Fragment Output Format

We represent each idea as an *idea fragment* with the following schema:

Idea Fragment Format

```
"idea_fragment": {
  "title": "Brief, descriptive title (max 15 words)",
  "core_insight": "2-3 sentence summary of the integration",
  "integration_mechanism": {
    "target_domain_elements": [
      "Target-domain concept or method",
      "Another target-domain concept or method"
    ],
    "selected_takeaways": [
      {
        "takeaway_id": "t1",
        "source_domain_formulation":
          "Conceptual insight using source-domain framing",
        "mechanism_explanation":
          "Explanation of the underlying conceptual logic",
        "selection_rationale":
          "Why this takeaway is relevant for integration"
      }
    ],
    "synthesis_approach":
      "Description of how elements are combined"
    },
    "challenge_resolution": {
      "addresses_target_challenge":
        "How the integration addresses the challenge",
      "addresses_source_limitations":
        "How integration mitigates limitations of the insight"
      "addresses_research_problem":
        "How this contributes to the overall research problem"
    },
    "concrete_realization": {
      "proposed_approach":
        "Specific algorithm, or technical realization",
      "key_innovations": [
        "Novel aspect enabled by integration",
        "Additional emergent innovation"
      ]
    }
  }
}
```

G Baselines

We compare IDEA-CATALYST against two baseline methods that reflect common LLM-driven approaches to interdisciplinary ideation with increasing degrees of retrieval structure:

- ***Free-Form Source Retrieval*** (Zheng et al., 2024) prompts the model to directly identify potentially relevant source domains (with no restriction on distance to the target domain) for a given research problem, generate search queries, retrieve papers from those domains, and synthesize research ideas, without explicit analysis of the target domain or decomposition of the problem. This baseline captures an intuition-driven approach that relies primarily on the model's parametric knowledge of the target domain rather than systematic reasoning about research gaps.
- ***Guided Dual-Retrieval*** introduces additional structure by first retrieving representative literature from the target domain, then conditioning cross-domain exploration and ideation

on this retrieved context. While this baseline incorporates retrieval from both target and source domains, it does not explicitly identify unresolved conceptual challenges, construct domain-agnostic abstractions, or strategically guide source-domain selection, effectively serving as a retrieve-then-ideate pipeline without metacognitive control.

Ablations We conduct the following ablation studies to isolate the contributions of key components of IDEA-CATALYST:

- ***No Decomposition*** removes the target-domain decomposition stage, relying instead on the model's parametric knowledge to assess which aspects of the research problem have been addressed and which challenges remain, without explicitly decomposing the problem into research questions or retrieving target-domain literature conditioned on them. This evaluates the importance of structured, retrieval-grounded self- and context-awareness for identifying meaningful research gaps.

- ***No Interdisciplinary Ranking*** removes the interdisciplinary potential-based ranking stage, replacing pairwise LLM-based comparisons with a heuristic ranking based solely on the proportion of retrieved source-domain papers deemed relevant to the target challenge. This tests whether explicit comparative evaluation is necessary to surface high-impact interdisciplinary ideas beyond relevance alone.

- ***Conceptual Rewriting*** retains the full pipeline but replaces the final idea fragment with a rewritten version that improves conceptual clarity and accessibility while preserving structure, technical content, and domain grounding. This assesses whether observed gains stem from deeper interdisciplinary integration rather than improved articulation or presentation.

H Methodology Prompt Templates

We include the core prompt templates used to implement the key stages of the IDEA-CATALYST pipeline. These templates were used to decompose target-domain problems, analyze target-domain literature, identify cross-domain search queries, extract source-domain insights, integrate interdisciplinary ideas, and rank final fragments.

H.1 Initial Decomposition Prompt

Initial Decomposition Prompt

```
You are an expert research strategist. Your
↪  task is to:
(1) identify the most relevant
↪  coarse-grained domain that encompasses
↪  the input fine-grained domain,
(2) decompose the problem into fundamental
↪  bottleneck questions grounded in that
↪  domain,
and (3) produce two parallel formulations
↪  for each question:
   - a domain-specific version (optimized
   ↪  for searching within the identified
   ↪  subfield)
   - a domain-agnostic version (optimized
   ↪  for discovering relevant work in
   ↪  external domains)

# RESEARCH PROBLEM
{problem_statement}

# FINE-GRAINED SUBFIELD
{fine_grained_domain}

# STEP 1: SUBFIELD IDENTIFICATION
↪  (MANDATORY)
First, determine the most relevant
↪  coarse-grained domain (the target
↪  domain) that the problem best fits out of
↪  the provided list. Then, all subsequent
↪  questions and queries MUST be aligned to
↪  the domain and subfield's typical
↪  concepts, bottlenecks, and vocabulary.

# STEP 2: CORE CHALLENGE
Write a 2-3 sentence core_challenge that
↪  frames the fundamental difficulty as
↪  researchers in the fine_grained_domain
↪  would.

# STEP 3: QUESTION DECOMPOSITION
↪  (SUBFIELD-GROUNDED)
Produce 2-4 underlying bottleneck questions.
↪  These should be:
- atomic
- fine-grained enough to be useful in the
↪  identified subfield
- critical to progress
- mechanistic
- not solution-prescriptive

# STEP 4: QUESTION PAIRING (TWO VERSIONS OF
↪  THE SAME BOTTLENECK)
For EACH research question, produce:
- domain_specific_question: uses
↪  fine-grained domain terminology
- domain_agnostic_question: same bottleneck
↪  phrased in cross-disciplinary
↪  mechanistic language

# STEP 5: SEARCH QUERIES (SUBFIELD-ALIGNED)
```

Table 7: Participant backgrounds and research problems used in the human study. All participants are PhD researchers ranging from 3 to 5 years of research experience.

Primary Research Area	Target Domain	Research Problem Description
Natural Language Processing	Multilingual NLP	Are there tasks where LLM performance varies systematically across languages, particularly for culture-specific queries, and how can such performance disparities be mitigated?
Electrical Engineering	In-Memory Computing	How can the accuracy of in-memory computing systems be improved while preserving high energy efficiency, especially for Edge-AI applications?
Natural Language Processing	Multilingual Semantics	Why do language models change their answers across languages for the same query, even in high-resource settings, and does this reflect knowledge or semantic misalignment?
Natural Language Processing	Persuasion and Safety	How can we characterize and mitigate the susceptibility of LLMs to persuasion, including harmful or adversarial influence, while preserving beneficial adaptability?
Machine Learning	Model Interpretability	How can influence functions be made dynamic, such that the importance of data points adapts across models and training contexts rather than remaining static?
Natural Language Processing	User Simulation	How can LLM-based user simulators better maintain consistent personas, reflect diverse user behaviors, and be evaluated for realism in multi-turn interactions?

```
For EACH research question, provide 3-5
↪  target_domain_queries (max 5 words each)
↪  that reflect fine-grained domain
↪  vocabulary.

# OUTPUT FORMAT
Return a JSON object containing
↪  coarse_grained_domain, core_challenge,
↪  and research_questions with both
↪  question formulations and
↪  target_domain_queries.
```

H.2 Target-Domain Analysis Prompt

Target-Domain Analysis Prompt

```
You are an expert research analyst. Analyze
↪  retrieved papers from a specific
↪  subfield of the target domain to assess
↪  progress on a research question.

# RESEARCH PROBLEM
{research_problem}

# TARGET DOMAIN
{target_domain}

# FINE-GRAINED DOMAIN
{fine_grained_domain}

# RESEARCH QUESTION (PAIRED FORMULATIONS)
- Domain-specific question:
{domain_specific_question}
- Domain-agnostic question:
{domain_agnostic_question}

# RETRIEVED PAPERS AND SNIPPETS
{papers_text}

# YOUR TASK

1. Assess paper relevance to the
↪  domain-specific question.
2. Identify which aspects are addressed.
3. Identify remaining challenges in both
↪  domain-specific and domain-agnostic
↪  forms.
4. Choose one overall assessment: largely
↪  unaddressed, partially addressed, or
↪  substantially addressed.

# OUTPUT FORMAT
Return a JSON object containing paper
↪  relevance annotations, addressed
↪  aspects, remaining challenges, and
↪  overall assessment.
```

H.3 Cross-Domain Query Generation Prompt

Cross-Domain Query Generation Prompt

```
You are an expert at identifying
↪  cross-disciplinary research connections.
Your task is to identify external domains
↪  and search queries that could address a
↪  research challenge by analogy, shared
↪  mechanisms, or transferable principles.

# RESEARCH PROBLEM
{research_problem}

# RESEARCH QUESTION
- Domain-specific version:
{domain_specific_question}
- Domain-agnostic version:
{domain_agnostic_question}

# ORIGINAL DOMAIN CONTEXT
- Broad domain: {target_domain}
- Fine-grained subfield:
↪  {fine_grained_domain}
```

```
# YOUR TASK
Identify 1-3 external domains from the
↪ allowed list (excluding the target
↪ domain) that are likely to contain
↪ relevant insights for the
↪ domain-agnostic question.
For each domain, explain why it is a good
↪ match and provide 2-4 domain-appropriate
↪ search queries (max 5 words each).

# OUTPUT FORMAT
Return a JSON object with
↪ cross_domain_searches containing domain,
↪ rationale, and queries.
```

H.4 Cross-Domain Analysis Prompt

Cross-Domain Analysis Prompt

```
You are an expert at identifying
↪ cross-disciplinary conceptual insights.
Analyze papers from an external domain to
↪ determine whether they provide
↪ meaningful conceptual understanding of a
↪ challenge.

# RESEARCH PROBLEM
{research_problem}

# TARGET DOMAIN CONTEXT
- Broad Domain: {target_domain}
- Specific Subfield: {fine_grained_domain}

# CONCEPTUAL CHALLENGE
{domain_agnostic_question}

# WHAT MAKES THIS CHALLENGING
{question_challenge}

# SOURCE DOMAIN
{source_domain}

# RETRIEVED PAPERS AND SNIPPETS
{papers_text}

# YOUR TASK
1. Assess paper relevance to the challenge.
2. Extract conceptual takeaways grounded in
↪ source-domain framing.
3. Evaluate whether the source domain
↪ provides sufficiently rich conceptual
↪ grounding.

# OUTPUT FORMAT
Return a JSON object with paper relevance,
↪ solution takeaways, and a challenge
↪ sufficiency assessment.
```

H.5 Target-Domain Integration Prompt

Target-Domain Integration Prompt

```
You are an expert research synthesizer.
Your task is to integrate target-domain
↪ challenges, source-domain takeaways, and
↪ retrieved evidence into an
↪ interdisciplinary idea fragment.

# RESEARCH PROBLEM
{research_problem}

# TARGET QUESTION
- Domain-specific:
↪ {domain_specific_question}
- Domain-agnostic:
↪ {domain_agnostic_question}

# SELECTED CHALLENGE
{question_challenge}

# TARGET DOMAIN PAPERS
{target_domain_papers}

# SOURCE DOMAIN
{source_domain}

# SOURCE DOMAIN PAPERS
{source_domain_papers}

# SOURCE DOMAIN TAKEAWAYS
{source_domain_takeaways}

# YOUR TASK
Produce an idea fragment containing: core
↪ insight, integration mechanism,
↪ challenge resolution, and a concrete
↪ realization.

# OUTPUT FORMAT
Return a JSON object describing the
↪ integrated idea fragment with selected
↪ takeaways, synthesis rationale, and a
↪ proposed realization.
```

H.6 Interdisciplinary Comparison Prompt

Interdisciplinary Comparison Prompt

```
You are an expert evaluator of
↪ interdisciplinary ideas.
Your task is to compare two candidate idea
↪ fragments and determine which has
↪ stronger interdisciplinary potential.

# RESEARCH PROBLEM
{research_problem}

# TARGET DOMAIN
{target_domain}

# CANDIDATE FRAGMENTS
{integrated_ideas}

# EVALUATION CRITERIA
Compare candidates based on:
```

```
- depth of integration
- source-domain relevance
- innovation payoff
- plausibility and potential impact

# OUTPUT FORMAT
Return a JSON object ranking the ideas or selecting the fragment with stronger interdisciplinary potential.
```

I Domains Supported by Semantic Scholar for Retrieval

Our literature retrieval pipeline relies on the Semantic Scholar *Snippets* API, which supports search queries over a fixed set of coarse-grained scientific domains. Specifically, Semantic Scholar indexes papers under the following fields of study: *Computer Science, Medicine, Chemistry, Biology, Materials Science, Physics, Geology, Psychology, Art, History, Geography, Sociology, Business, Political Science, Economics, Philosophy, Mathematics, Engineering, Environmental Science, Agricultural and Food Sciences, Education, Law*, and *Linguistics*.

To ensure compatibility with this constraint, all fine-grained target-domain subfields provided as input to IDEA-CATALYST (e.g., *Natural Language Processing*, *Reinforcement Learning*, *Cognitive Science*) are mapped to their corresponding coarse-grained Semantic Scholar domains (e.g., *Computer Science*, *Psychology*) prior to retrieval. This mapping is used *solely for literature retrieval and filtering* and does not affect the conceptual formulation of research questions, domain-agnostic abstractions, or subsequent interdisciplinary reasoning stages.

J Human Study Details & Participant Backgrounds

We conducted a human study with six PhD researchers working in Machine Learning, Natural Language Processing, and Electrical Engineering, each of whom provided a real research problem drawn from their own work (as shown in Table 7). We note that the study was declared exempt after being reviewed by our Institutional Review Board (IRB). We provide screenshots of our evaluation interface in Figures 5, 6, and 7.

K Evaluation Prompts

Takeaway Evaluation Prompt

```
You are an expert evaluator assessing the quality of cross-domain research takeaways.
Your task is to compare takeaways from two different methods that attempt to address the
same research problem by drawing insights from domains outside the target domain.

--------------------------------------------------------------------------------
RESEARCH PROBLEM
{research_problem}

TARGET DOMAIN
{target_domain}

--------------------------------------------------------------------------------
METHOD 1 TAKEAWAYS
{method_1_text}

--------------------------------------------------------------------------------
METHOD 2 TAKEAWAYS
{method_2_text}

--------------------------------------------------------------------------------
EVALUATION CRITERIA

When evaluating Method 1 and Method 2, explicitly ground your judgment in the relevant
fields of each takeaway, as described below.

### 1. INTERDISCIPLINARY INSIGHTFULNESS
Assess whether the method's takeaways provide insightful perspectives on the research
problem.
  - Perspectives should introduce specific concepts/frameworks from their respective
    source domain
  - Insightful perspectives should be intellectually interesting, non-obvious, and
    thought-provoking to researchers in the target domain ({target_domain})
    - Non-obvious perspectives typically come from source domains that are meaningfully
      distinct from the target domain ({target_domain})

### 2. INTERDISCIPLINARY RELEVANCE
Assess whether the method's takeaways are relevant to the research problem and have
strong potential for integration in the target domain ({target_domain}).
  - Ideal takeaways should:
```

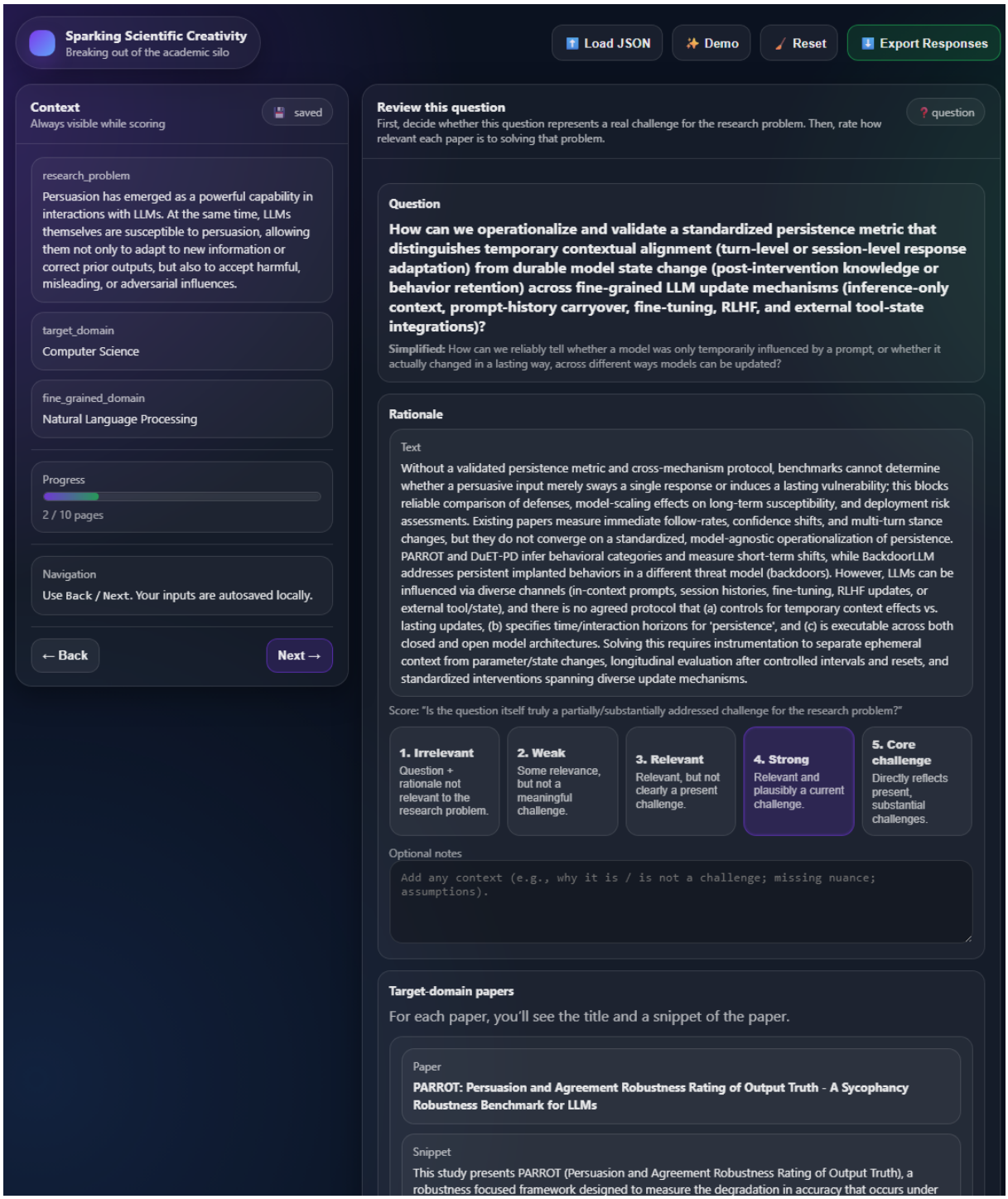

Figure 5: Screenshot of evaluation interface (reviewing questions/challenges).

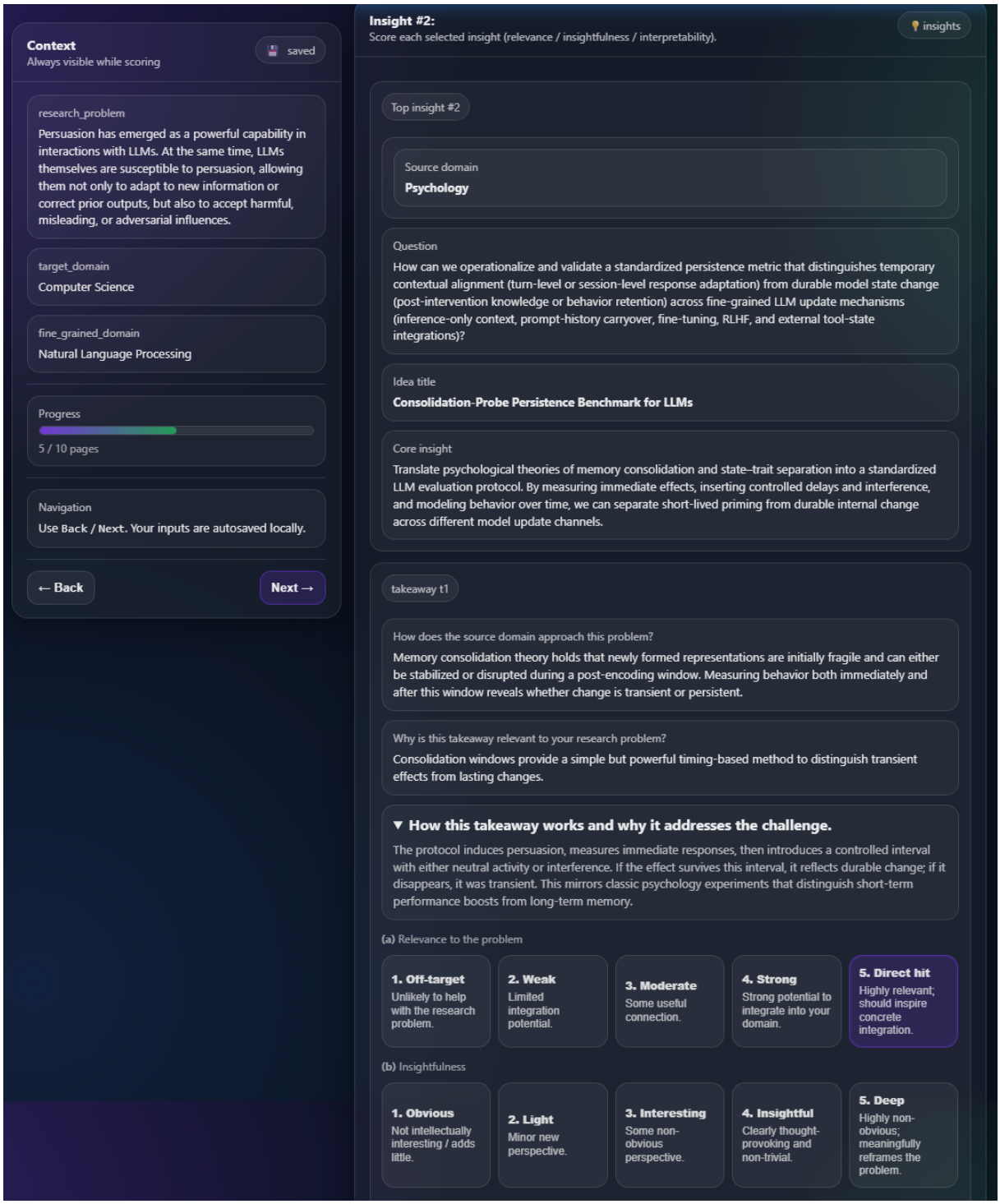

Figure 6: Screenshot of evaluation interface (reviewing takeaways).

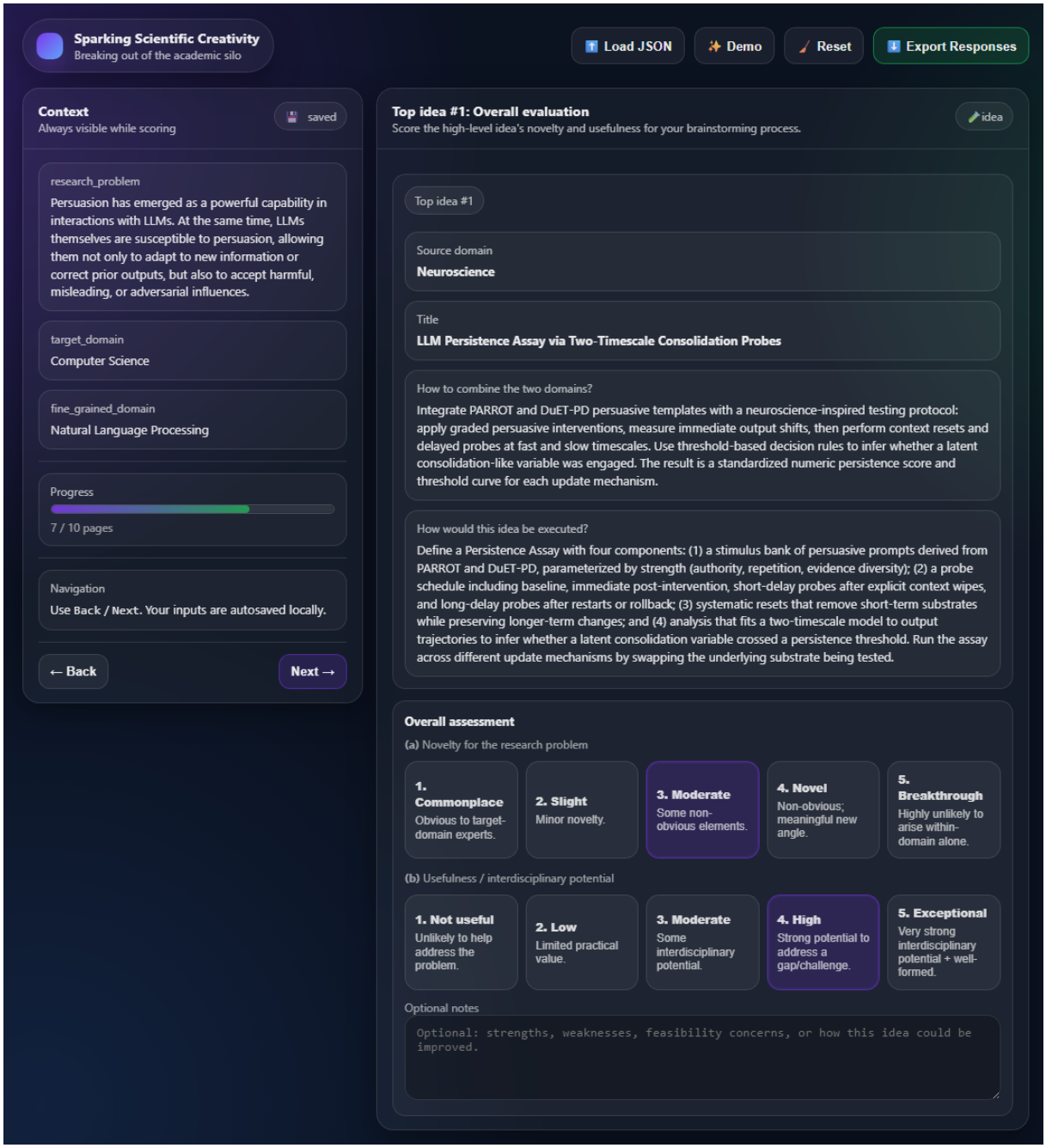

Figure 7: Screenshot of evaluation interface (reviewing ideas).

```
    - Inspire new approaches/solutions to
    ↪  the research problem in the target
    ↪  domain
    ({target_domain})
    - Address a gap/challenge for the
    ↪  research problem in the target
    ↪  domain
    ({target_domain})
  - The complexity, simplicity, or
  ↪  practicality of the takeaway should
  ↪  not factor into
  your decision (e.g., a "clear, immediately
  ↪  applicable" solution does not mean
  ↪  more
  relevant). Relevance is defined based on
  ↪  the potential impact of the source
  ↪  domain
  being introduced to the target domain for
  ↪  the research problem.
  - Keep in mind that if the distance between
  ↪  the source and target domain is larger
  (e.g., Computer Science & Engineering are
  ↪  closer than Computer Science &
  ↪  Philosophy),
  the idea may inherently be less practical.
  ↪  This does not mean that it is less
  ↪  relevant.
  Focus on the degree of the potential impact
  ↪  to the research problem instead.

IGNORE:
- Length of explanations
- Narrative polish
- Missing implementation details

CONSIDER:
- **Consistency**: Are the method's
↪  takeaways consistently
meaningful, or uneven?
- **Groundedness**: Are claims supported by
↪  real conceptual
alignment?
- **Scope appropriateness**: Are takeaways
↪  neither trivial
nor wildly speculative?

--------------------------------------------⌋
↪  -------
OUTPUT FORMAT

Return a JSON object:

{{
  "takeaway_comparison": {{
    "interdisciplinary_insightfulness": {{
      "preferred_method": 1 | 2,
      "reasoning": "1-2 sentences explaining
      ↪  your reasoning for the preferred
      ↪  method"
    }},
    "interdisciplinary_relevance": {{
      "preferred_method": 1 | 2,
      "reasoning": "1-2 sentences explaining
      ↪  your reasoning for the preferred
      ↪  method based
      on the evaluation criteria"
    }},
  }},
  "overall_assessment": {{
    "preferred_method": 1 | 2,
    "summary": "2-3 sentences explaining
    ↪  which method's takeaways are higher
    ↪  quality in
   terms of interdisciplinary insightfulness
    ↪  and interdisciplinary relevance"
  }}
}}
```

Idea Evaluation Prompt

```
  You are an expert evaluator assessing the
  ↪  quality of cross-domain RESEARCH
  ↪  IDEAS.
Your task is to compare two proposed ideas
↪  that integrate insights from an external
↪  domain
to address the same research problem.

--------------------------------------------⌋
↪  -------
RESEARCH PROBLEM
{research_problem}

TARGET DOMAIN
{target_domain}

--------------------------------------------⌋
↪  -------
METHOD 1 IDEA

Source Domain:
{method_1_idea.get("source_domain", "N/A")}

Proposed Approach:
{method_1_idea.get("idea",
↪  {}).get("proposed_approach", "N/A")}

Key Innovations:
{method_1_idea.get("idea",
↪  {}).get("key_innovations", [])}

Supporting Takeaways:
{method_1_text}

--------------------------------------------⌋
↪  -------
METHOD 2 IDEA

Source Domain:
{method_2_idea.get("source_domain", "N/A")}

Proposed Approach:
{method_2_idea.get("idea",
↪  {}).get("proposed_approach", "N/A")}

Key Innovations:
{method_2_idea.get("idea",
↪  {}).get("key_innovations", [])}

Supporting Takeaways:
{method_2_text}

--------------------------------------------⌋
↪  -------
EVALUATION CRITERIA
```

```
### 1.INTERDISCIPLINARY NOVELTY
Which idea is more novel?
  - The **source domain** chosen and its
  ↪  conceptual distance from the target
  ↪  domain
  - The **proposed_approach**: Is the idea
  ↪  non-obvious to target-domain experts?
  - The **key_innovations**: Do they reflect
  ↪  insights unlikely to arise within the
  target domain alone?
  - Whether the supporting takeaways draw on
  ↪  **less common or underexplored
  ↪  external
  insights**

Higher novelty means:
  - The idea is surprising but still credible
  - The cross-domain move feels inventive
  ↪  rather than expected

### 2.INTERDISCIPLINARY USEFULNESS
Which idea has greater interdisciplinary
↪  potential for addressing the research
↪  problem in
the target domain ({target_domain})?
  - Ideas with greater interdisciplinary
  ↪  potential should:
    - Present new approaches/solutions to
    ↪  the research problem in the target
    ↪  domain
    ({target_domain})
    - Address a gap/challenge for the
    ↪  research problem in the target
    ↪  domain
    ({target_domain})
    - The idea integrates the concepts from
    ↪  both the target domain and source
    ↪  domain
    into a well-formed idea that addresses
    ↪  the research problem
  - The complexity, simplicity, or
  ↪  practicality of the proposed idea
  ↪  should not factor
  into your decision (e.g., a more "clear,
  ↪  immediately
  ↪  applicable"/"direct"/"concrete"
  solution does not make it more
  ↪  useful).Usefulness is defined based on
  ↪  the potential
  impact of the source domain being
  ↪  introduced to the target domain.
  ↪  Specifically, a more
  useful interdisciplinary idea integrates
  ↪  the source and target domains in a way
  ↪  that
  allows for a more significant
  ↪  problem/challenge to be solved or a
  ↪  significant gap in
  existing ideas to be addressed.
  - Keep in mind that if the distance between
  ↪  the source and target domain is larger
  (e.g., Computer Science & Engineering are
  ↪  closer than Computer Science &
  ↪  Philosophy),
  the idea may inherently be less practical.
  ↪  This does not mean that it is less
  ↪  useful.
```

```
  Focus on the degree of the potential
  ↪  impact instead.

-------------------------------------------┘
↪  -------
OUTPUT FORMAT

Return a JSON object:

{{
  "idea_comparison": {{
    "interdisciplinary_novelty": {{
      "preferred_method": 1 | 2,
      "reasoning": "1-2 sentences explaining
      ↪  which idea is more novel"
    }},
    "interdisciplinary_usefulness": {{
      "preferred_method": 1 | 2,
      "reasoning": "1-2 sentences explaining
      ↪  why the preferred idea is more
      ↪  useful than
      the other idea for the research problem
      ↪  based on the evaluation criteria"
    }}
  }},
  "overall_assessment": {{
    "preferred_method": 1 | 2,
    "summary": "2-3 sentences summarizing
    ↪  which idea is overall more
    ↪  interdisciplinary
    novel, interdisciplinary useful, and
    ↪  integrates the two domains better"
  }}
}}
```