SPARKING SCIENTIFIC CREATIVITY VIA LLM-DRIVEN INTERDISCIPLINARY INSPIRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Interdisciplinary research yields disproportionate scientific impact, yet most work remains confined within disciplinary boundaries. Existing AI systems for scientific discovery primarily focus on automating experiment design and execution, often neglecting the exploratory reasoning processes that underlie cross-domain innovation. We propose **IDEA-CATALYST**, a metacognition-driven framework for interdisciplinary research ideation. Given a research problem, the framework decomposes it into core research questions, identifies unresolved conceptual challenges through literature analysis, retrieves corresponding insights from external disciplines, and converts them into candidate ideas ranked by interdisciplinary potential. Across LLM-based and human evaluations, IDEA-CATALYST produces ideas that are **21%** more novel and **16%** more insightful while remaining grounded in the original research problem. These results demonstrate the value of structured cross-domain synthesis for AI-assisted scientific creativity.

1 INTRODUCTION

Scientific breakthroughs rarely emerge from isolated "*eureka*" moments. Instead, research advances through the gradual accumulation and recombination of partial ideas across domains (Sosa, 2019; Gonçalves & Cash, 2021). Early conceptual fragments seed interdisciplinary discussion, critique, and refinement, eventually coalescing into mature research directions. Reinforcement learning, for example, arose from the convergence of behavioral psychology, control theory, and animal learning research rather than from a single field (Sutton et al., 1998). Such boundary-spanning synthesis has repeatedly driven scientific progress.

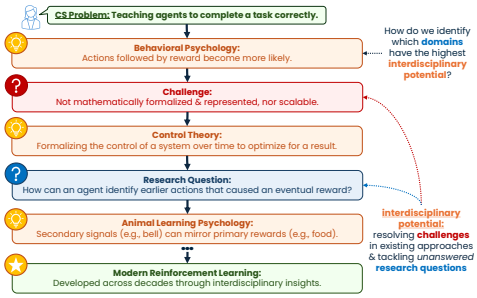

Figure 1: Interdisciplinary process of formalizing RL (Sutton et al., 1998).

Empirical evidence shows that interdisciplinary integration substantially increases long-term impact, with each additional discipline associated with roughly 20% higher citation impact (Van Noorden, 2015; Okamura, 2019). Yet deeply integrative cross-domain collaboration remains rare (Porter & Rafols, 2009; Raasch et al., 2013). A key challenge is how to systematically foster interdisciplinary scientific creativity while avoiding the constraints of disciplinary silos.

Recent work on AI-assisted scientific discovery explores "AI co-scientists" that support ideation, experimentation, and critique (Gottweis et al., 2025; Goel et al., 2025; Si et al., 2026; Jansen et al., 2025). However, prior studies reveal a trade-off (Si et al., 2024): human-generated ideas are typically well grounded but domain-constrained, whereas LLM-generated ideas more readily draw cross-domain inspiration but often lack depth, feasibility, or conceptual grounding (Si et al., 2024; 2026; Gupta & Pruthi, 2025). Attempts to improve feasibility by tightly coupling ideation with automatic experimentation further risk narrowing exploration to incremental refinements (Si et al., 2026; Jansen et al., 2025). Premature evaluation can truncate creative exploration rather than expand it (Bose et al., 2013; Catmull & Wallace, 2023). To prioritize early-stage creative exploration,

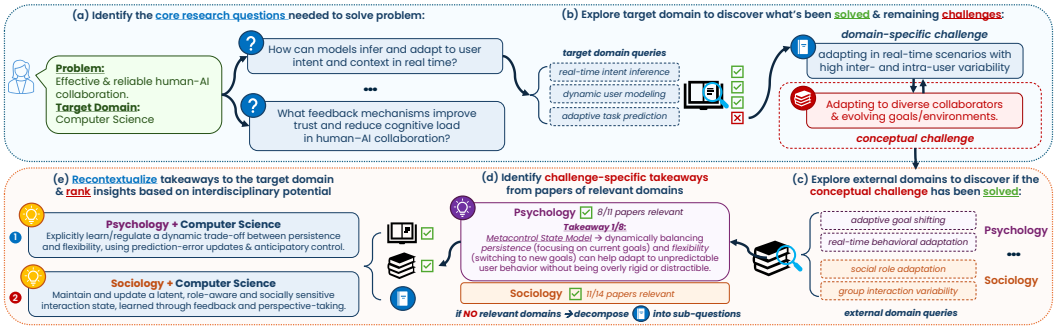

Figure 2: **IDEA-CATALYST**: Critical & creative reasoning framework for interdisciplinary ideation.

we propose **IDEA-CATALYST**, a framework for generating interdisciplinary insights grounded in unresolved conceptual challenges in a target domain's research problem.

Rather than automating end-to-end research execution, IDEA-CATALYST structures exploratory cross-domain synthesis to augment early-stage ideation: ***(a)* analyzes the target domain** by decomposing the problem into core research questions and identifying unresolved conceptual challenges through literature-conditioned analysis, ***(b)* explores external domains** to retrieve conceptually analogous solutions studied under different assumptions or formalisms, and ***(c)* recontextualizes and prioritizes insights** by translating cross-domain findings into candidate idea fragments and ranking them by novelty and relevance. Our main contributions are: ***(i)*** we introduce **IDEA-CATALYST**, a structured framework for interdisciplinary research ideation through problem decomposition, cross-domain retrieval, and strategic prioritization, ***(ii)*** we construct a dataset and evaluation protocol for benchmarking interdisciplinary idea generation across novelty, insightfulness, relevance, and usefulness, and ***(iii)*** based on automatic and human evaluations, **IDEA-CATALYST** produces 21.38% more novel and 16.22% more insightful ideas.

2 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.

2.1 PROBLEM FORMULATION

To support early-stage conceptual brainstorming, we assume as input only a short research problem statement p (e.g., 1–2 sentences on effective and reliable human-AI collaboration) situated within a *target domain* $\mathcal{D}_{\text{target}}$ (e.g., Natural Language Processing). Our *objective* is to generate a set of interdisciplinary idea fragments $\mathcal{F}$, where each fragment $f_i \in \mathcal{F}$ is grounded in an external *source domain* $\mathcal{D}_{s_i}$ and comprised of a set of literature-derived insights $t \in T_{s_i}$. Each fragment proposes a candidate interdisciplinary idea $\hat{T}_{s_i}$ by recontextualizing these insights to address p, thereby integrating concepts from $\mathcal{D}_{s_i}$ into the target domain $\mathcal{D}_{\text{target}}$. We define each of these key terms in Appendix A.1, as well as how we retrieve scientific literature in Appendix A.2.1.

2.2 METACOGNITION-DRIVEN IDEATION

Scientific ideation requires balancing novelty and usefulness (Yanai & Lercher, 2019). We ground IDEA-CATALYST in *metacognition*: the monitoring and regulation of reasoning during problem solving (Flavell, 1979). Creative performance depends on accurately assessing what is known, identifying gaps, selecting appropriate strategies, and evaluating progress (Urban & Urban, 2025; Kargupta et al., 2025). Accordingly, our framework is grounded on five metacognitive functions: ***(i)*** assessing progress and open challenges in the target domain (self-awareness), ***(ii)*** recognizing domain assumptions and limits (context awareness), ***(iii)*** selecting suitable external disciplines for exploration (strategy selection), ***(iv)*** decomposing and prioritizing research questions (goal manage-

ment), and ***(v)*** evaluating whether generated insights meaningfully address unresolved challenges (evaluation). Overall, ideation emerges from coordinating these functions during both *critical reasoning* (structured analysis of challenges and opportunities) and *creative reasoning* (cross-domain ideation) (Johnson-Laird, 2010). IDEA-CATALYST balances both by grounding exploration in systematic target-domain analysis while preserving interdisciplinary expansion. We include all details for our framework in Appendix A.2.2-A.2.5.

Critical Reasoning over the Target Domain. We begin with structured analysis of the target domain $\mathcal{D}_{\text{target}}$. Given a problem p, we decompose it into research questions $q_i \in \mathcal{Q}$ and assess how thoroughly each has been addressed through literature retrieval and analysis. Each question is represented in two forms: a domain-specific formulation $q_i^{\mathcal{D}}$ (grounded in target-domain terminology) and a domain-agnostic abstraction q_i' that captures the underlying conceptual issue. This dual representation enables both precise assessment of progress and cross-domain comparability. For each q_i, retrieved literature is analyzed to categorize the question as resolved, partially addressed, or open. We further extract unresolved conceptual challenges that remain weakly addressed. These challenges form the focal points for interdisciplinary exploration.

Creative Reasoning Across Source Domains. For questions or challenges that are open or partially addressed, we transition to cross-domain exploration. Using the domain-agnostic formulation q', we identify external source domains that may contain analogous perspectives and concepts which address q. For each candidate source domain $\mathcal{D}_s$, we retrieve and analyze relevant literature. We extract literature-grounded conceptual takeaways only when a majority of retrieved work meaningfully addresses the underlying challenge. This ensures that insights reflect domain-level perspectives rather than isolated findings. These cross-domain takeaways expand the solution space by introducing mechanisms, abstractions, or theoretical principles absent from the target domain.

Target–Source Interdisciplinary Integration. We define an *idea fragment* as a structured intermediate representation linking: (i) a target-domain challenge and its literature, (ii) source-domain conceptual takeaways and supporting evidence, and (iii) a rationale describing how the takeaway addresses the challenge. Idea fragments are intentionally incomplete, serving as catalysts for further development rather than full solutions. For each eligible question–source-domain pair, we integrate relevant takeaways with target-domain assumptions to produce candidate fragments. Integration evaluates how cross-domain perspectives complement existing approaches and address unresolved conceptual gaps. Because multiple fragments may be generated, we rank them by *interdisciplinary potential*, reflecting depth of integration, expected innovation payoff, and balance between novelty and grounding Porter & Rafols (2009). We perform pairwise comparisons among fragments to obtain a relative ordering, prioritizing those most likely to yield impactful cross-domain advances. Overall, IDEA-CATALYST *structures interdisciplinary ideation through coordinated critical and creative reasoning* while maintaining grounding in the target domain.

3 EXPERIMENTAL DESIGN

We use `Qwen3-14B` (Yang et al., 2025) as the primary model and `gpt-oss-120b` (Agarwal et al., 2025) as the LLM judge. We provide all details of our experimental design in Appendix A.3 and baselines/ablations in Appendix A.3.1.

Dataset. We evaluate on CHIMERA (Sternlicht & Hope, 2025), a dataset of interdisciplinary arXiv papers annotated with *inspiration* relations between source and target domains. We select 400 instances where source and target belong to distinct coarse-grained scientific fields and the problem context (input p) does not reveal interdisciplinary insights. For evaluation, each instance is converted into our idea-fragment format (Appendix A.6).

Evaluation Metrics. We evaluate early-stage ideation following pairwise preference protocols, where generated outputs are compared against ground-truth interdisciplinary contributions and win rates are reported (Dubois et al., 2024; Zheng et al., 2023; Si et al., 2024). We assess **takeaways** based on interdisciplinary insightfulness and relevance, and **integrated ideas** based on interdisciplinary novelty and usefulness. We provide all metric and prompt details in Appendix A.4 & B.

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

Method (Takeaway)	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†

Method (Idea)	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

4 EXPERIMENTAL RESULTS

Overall Performance. IDEA-CATALYST consistently outperforms baselines on exploratory dimensions (Table 1). At the takeaway level, it improves insightfulness by **16.22%** over *Guided Dual* and **282.21%** over *Free-Form Source*; at the idea level, it increases novelty by **21.38%** and **407.65%**, respectively. These gains indicate that metacognition-driven, target-grounded exploration produces more novel and insightful outputs than retrieval-based approaches. We observe a trade-off: ranking by interdisciplinary potential favors novelty and insightfulness, with relevance and usefulness increasing as k grows. As a whole, LLM judgments tend to weight relevance and usefulness more heavily when assigning an overall winner.

Ablations. Removing target-domain decomposition reduces novelty and insightfulness, suggesting that structured identification of conceptual gaps is critical. Replacing pairwise interdisciplinary-potential ranking with simple relevance heuristics also lowers performance, highlighting the importance of comparative evaluation. Although *Conceptual Rewriting* does not improve quantitative metrics, it enhances clarity and interpretability in qualitative and human assessments.

Analyzing Source Distribution and Target–Source Pairings (Appendix A.5.1). Source-domain distribution analysis shows that *Free-Form Source* collapses toward Computer Science (low diversity), while IDEA-CATALYST achieves broad cross-domain exploration with balanced diversity and stronger novelty/insightfulness, indicating that effective ideation requires selective, conceptually grounded expansion rather than maximal spread. Target–source flow visualizations further reveal intuitive but diverse inspiration patterns (e.g., AI drawing from Psychology and Linguistics), suggesting structured rather than arbitrary interdisciplinarity.

Qualitative and Human Evaluation (Appendix A.5.2, A.5.3). Qualitative comparison shows that, beyond identifying relevant domains, IDEA-CATALYST extracts more problem-aligned and mechanism-specific insights than baselines. In a human study with six PhD researchers, participants rated the framework highly for early-stage ideation, with ideas judged more novel than useful, highlighting both its creative strength and the typical ideation–execution gap.

5 CONCLUSION

We present IDEA-CATALYST, a metacognition-driven framework for interdisciplinary research ideation that guides cross-domain exploration using structured target-domain analysis. The framework generates ideas that are significantly more novel and insightful than baselines while remaining grounded in the original problem. Our findings highlight the value of supporting the *process* of critical & creative reasoning for human–AI scientific collaboration. Future work includes personalized summarization for researchers and collaborator recommendation based on interdisciplinary signals.

REFERENCES

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

Mousumi Bose, Judith Anne Garretson Folse, and Sooyeon Nikki Lee-Wingate. 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, 2013.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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.

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

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

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

Noy Sternlicht and Tom Hope. Chimera: A knowledge base of idea recombination in scientific literature, 2025. URL https://arxiv.org/abs/2505.20779.

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

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

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

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

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

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

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

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

Chengbo Zheng, Yuanhao Zhang, Zeyu Huang, Chuhan Shi, Minrui Xu, and Xiaojuan Ma. 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*, pp. 1–20, 2024.

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

A Appendix

A.1 Problem Formulation Definitions

Definition A.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 established literature, methodologies, problem formulations, and evaluation norms. The goal of our framework is to augment ideation within $\mathcal{D}_{\text{target}}$ by introducing conceptually grounded insights originating outside the domain.

Definition A.2 (***Source Domain***)**.** A *source domain* $\mathcal{D}_s$ is a scientific field distinct from the target domain $\mathcal{D}_{\text{target}}$, characterized by its own literature, conceptual frameworks, and problem-solving traditions. Source domains serve as potential reservoirs of transferable insights that, when appropriately recontextualized, may help address unresolved conceptual challenges in $\mathcal{D}_{\text{target}}$. To promote non-trivial interdisciplinary connections, we restrict source domains to fields that are sufficiently distant from the target domain at a coarse-grained level of similarity (e.g., Computer Science and Psychology), and exclude closely related subfields (e.g., Natural Language Processing and Machine Learning).

Definition A.3 (***Interdisciplinary Insight***)**.** An *interdisciplinary insight* $t \in T_{s_i}$ is a literature-grounded conceptual takeaway extracted from the source domain $\mathcal{D}_{s_i}$. Such insights typically describe mechanisms, principles, or abstractions that are not natively expressed in the target domain $\mathcal{D}_{\text{target}}$ but may become relevant when mapped onto the research problem p.

Definition A.4 (***Interdisciplinary Potential***)**.** *Interdisciplinary potential* denotes the expected value of an idea fragment $f_i \in \mathcal{F}$ for advancing the research problem p through cross-domain integration. It reflects a fragment's ability to (i) address unresolved conceptual challenges in $\mathcal{D}_{\text{target}}$, (ii) introduce non-trivial perspectives from $\mathcal{D}_{s_i}$, and (iii) plausibly inspire novel research directions when recontextualized within the target domain. Idea fragments in $\mathcal{F}$ are ranked according to this potential.

A.2 Methodology Details

A.2.1 Scientific Literature Snippet Retrieval

To ground interdisciplinary ideation in existing scientific knowledge, we retrieve literature snippets using the Semantic Scholar *Snippets* API[1]. Given a natural-language query and a coarse-grained scientific domain (chosen from the options below), the Snippets API returns short, relevance-ranked text passages extracted from papers, along with their associated metadata. Given that Semantic Scholar performs the underlying document parsing, indexing, and query-snippet relevance matching internally, this allows us to treat snippet retrieval as a black-box operation and focus on structuring

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

the research ideation process rather than optimizing retrieval models. For each query, we retrieve the top-k papers within a specified domain and aggregate multiple snippets per paper when available. When snippet text is unavailable or degenerate (e.g., identical to the paper title), we fall back to retrieving the paper abstract to ensure minimal contextual grounding. Retrieved snippets are used as lightweight, fine-grained evidence for downstream analysis, enabling the identification of unresolved challenges and transferable conceptual insights across domains while maintaining scalability across diverse scientific fields.

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.

A.2.2 METACOGNITION-DRIVEN IDEATION

Scientific research ideation is inherently creative, requiring ideas to be both novel and useful (Yanai & Lercher, 2019). A central component of this process is *metacognition*: the ability to monitor, evaluate, and regulate one's own reasoning during problem solving Flavell (1979); Yeung & Summerfield (2012); Kitchner (1983). Prior work shows that creative performance depends on the metacognitive awareness of which strategies are appropriate, when to apply them, and how to assess progress. Inaccurate metacognitive monitoring can disrupt how individuals guide and adjust their creative reasoning, whereas stronger metacognition is consistently associated with more effective creative problem solving (Urban & Urban, 2025). Thus, we align our framework to the following metacognitive behaviors (Kargupta et al., 2025):

- ***Self-awareness:*** Assessing what is known, what remains uncertain, and which aspects of a research problem are both challenging and actionable. In our framework, this corresponds to evaluating how thoroughly different facets of problem p have been addressed in the target-domain literature.
- ***Context awareness:*** Recognizing the assumptions, constraints, and norms that shape a problem. For our task, this includes recognizing target-domain limitations and identifying external source domains that may offer complementary perspectives.
- ***Strategy selection:*** Choosing reasoning strategies aligned with the nature of the problem. In practice, this involves selectively exploring disciplines that are well suited to particular challenges and open research questions of p (e.g., *control theory* for formalization, *psychology* for learning behavior).
- ***Goal management:*** Maintaining and adapting intermediate objectives. This manifests as decomposing p into research questions, prioritizing those with the greatest potential for conceptual advancement, and assessing progress made, post-ideation.
- ***Evaluation:*** Monitoring the quality & promise of the reasoning process. Rather than prematurely enforcing feasibility, the ideation process should assess whether insights meaningfully address unresolved conceptual challenges.

Under this view, creative ideation emerges from the coordination of two complementary reasoning modes (Johnson-Laird, 2010): ***critical reasoning***, which emphasizes structured evaluation and analytical rigor, and ***creative reasoning***, which supports the generative synthesis of novel and valuable ideas (de Chantal & Markovits, 2022; Dwyer et al., 2025). Our framework balances these modes by grounding ideation in a systematic analysis of the target domain while preserving space

for exploratory, interdisciplinary reasoning that expands the solution space (Wechsler et al., 2018; Halpern, 2007).

A.2.3 CRITICAL REASONING OVER THE TARGET DOMAIN

IDEA-CATALYST initiates the creative ideation process with a systematic analysis of $\mathcal{D}_{\text{target}}$, grounding the model's self-awareness and context awareness in the current state of the literature (e.g., state-of-the-art approaches, technical/conceptual limitations). This analysis enables a structured, critical assessment of what has already been addressed, where progress is uneven, and which conceptual challenges remain unresolved.

To steer ideation toward insights that are both novel and useful, we analyze $\mathcal{D}_{\text{target}}$ to identify aspects of p that are weakly addressed and therefore offer the greatest potential for impact. We first decompose p into a structured set of research questions $q_i \in \mathcal{Q}$, allowing us to examine the problem from multiple complementary perspectives that may exhibit varying levels of maturity in the existing literature. Each question q_i is represented in two forms: a ***domain-specific*** formulation $q_i^{\mathcal{D}}$, expressed in the language and assumptions of $\mathcal{D}_{\text{target}}$, and a corresponding ***domain-agnostic*** formulation q_i' that abstracts away academic jargon and implementation details. This dual representation enables precise assessment of progress within $\mathcal{D}_{\text{target}}$ while facilitating conceptual cross-domain comparison. For example, given the problem of "*effective and reliable human–AI collaboration*" (Figure 2), the resulting research questions include:

Domain-Specific Question ($q_i^{\mathcal{D}}$)	**Domain-Agnostic Question** (q_1')
How can models be trained to dynamically infer and adapt to user intent and task context in real-time collaborative scenarios?	*How can understanding of intent and context be updated through continuous interaction?*
How should a system decide when to take initiative vs. defer to human to maintain well-calibrated autonomy & control across contexts?	*When should control be exercised versus withheld?*

Table 2: Output examples of domain-specific & agnostic research question pairs for "human-AI collaboration".

For each research question q_i, we generate a set of natural-language search queries that capture its domain-specific formulation $q_i^{\mathcal{D}}$ (e.g., *real-time intent inference*, *dynamic user modeling*). These queries are used to retrieve a set of relevant papers and associated literature snippets $\{d_1, \ldots, d_k\} \subset \mathcal{D}_{\text{target}}$ (Section A.2.1). Based on the retrieved papers, we assess their relevance to q_i and evaluate the extent to which the question has been addressed in the target domain (e.g., *largely resolved*: $\mathcal{Q}_{resolved} \subseteq \mathcal{Q}$, *partially addressed*: $\mathcal{Q}_{partial}$, or *largely unexplored*: $\mathcal{Q}_{open}$). Crucially, this analysis surfaces remaining critical, non-incremental challenges q_j^i that are explicitly stated or implicitly suggested by the literature and are not resolved by existing approaches. Each remaining challenge q_j^i inherits the same dual representation as its parent question q_i, consisting of both a domain-specific and a domain-agnostic formulation, as shown in Table 3.

Domain-Specific Challenge ($q_1^{1,\mathcal{D}}$)	**Domain-Agnostic Challenge** ($q_1'^1$)
How can a system adapt in real-time to high inter/intra-user variability?	*How can behavior adapt to diverse collaborators & evolving goals/environments?*

Table 3: Outputted dual representation of a remaining challenge q_j^i in addressing q_i, derived from analysis of $\mathcal{D}_{\text{target}}$.

A.2.4 CREATIVE REASONING ACROSS SOURCE DOMAINS

Having identified weakly addressed research questions and unresolved conceptual challenges in the target domain $\mathcal{D}_{\text{target}}$, IDEA-CATALYST transitions from critical reasoning to creative exploration. Rather than expanding the solution space indiscriminately, we use insights from the target-domain analysis to strategically guide cross-domain exploration toward directions with the greatest potential for non-incremental interdisciplinary insight. Specifically, we prioritize research questions $q_i \in \mathcal{Q}_{open}$ that remain largely unexplored in $\mathcal{D}_{\text{target}}$, as well as conceptual challenges $q_j^i \in \mathcal{Q}_{partial}$, where alternative perspectives have the most potential to contribute.

This exploration is driven by the domain-agnostic form of each selected question or challenge. By abstracting away target-domain terminology and implementation details, these formulations isolate the underlying conceptual gaps that remain unresolved (Table 3). Such abstractions are more likely to correspond to theoretical constructs, explanatory frameworks, or empirical phenomena studied in external fields, even when surface-level applications differ. Consequently, domain-agnostic questions form the basis for selecting candidate source domains and structuring cross-domain search.

Cross-Domain Retrieval. For each domain-agnostic question or challenge q', we first identify a small set of external source domains that are plausibly relevant through analogy (e.g., how groups coordinate in Sociology vs. how teams coordinate in human–AI systems), shared mechanisms (e.g., adaptation through feedback in Psychology vs. Control Theory), or transferable principles (e.g., reasoning about uncertainty in Cognitive Science vs. machine learning), while explicitly excluding domains that are overly proximal to $\mathcal{D}_{\text{target}}$. This selection reflects the intuition that *more distant* domains are likelier to contribute novel perspectives rather than incremental variations of existing approaches. For each selected source domain $\mathcal{D}_s$, we then generate a small set of search queries which reflect the domain-specific vocabulary of $\mathcal{D}_s$ (e.g., specific terminologies, frameworks $\rightarrow$ "*cognitive load theory*" or "*social role adaptation*" in Sociology).

Using these queries, we retrieve papers and their respective snippets (Section A.2.1) from each source domain and analyze them to determine whether they provide meaningful conceptual insight into the challenge. For each source domain, we assess the relevance of retrieved papers, where we only extract literature-grounded conceptual takeaways $t_i \in T_{s_i}^{q'}$ from domains where the *majority* of retrieved papers are relevant to conceptual question/challenge q'. This ensures that *the overall insight t_i of the source domain has sufficient grounding in its literature* and is not an isolated finding. Table 4 showcases a real example of an insight from Psychology on the "*human-AI collaboration* problem. Each insight is structured into a set of takeaways, where each contains the specific source domain concept and its underlying logic/perspective in understanding the question/challenge ${q'}_j^i$.

Source Domain Concept	**How Does it Work?**
Metacontrol State Model: Goal-directed behavior reflects a balance between persistence (maintaining current goals) and flexibility (switching goals when conditions change).	Dynamic regulation between persistence and flexibility allows adaptive behavior that remains focused while responding efficiently to changing goals and environments Castro et al. (2021).
Cognitive Control: Control processes should be prospectively adjusted based on the expected frequency of goal switches.	Anticipatory adjustment of cognitive control reduces the cost of goal switching, enabling smoother transitions and lower cognitive load under frequent change Grahek et al. (2023).
Dynamic Goal Pursuit: Goal pursuit can be supported through just-in-time adaptive interventions that monitor behavior and provide context-sensitive support.	Real-time monitoring and adaptive feedback help sustain goal pursuit under variability by aligning support with evolving goals and situational demands O'Driscoll et al. (2024).

Table 4: Summarized conceptual takeaways $t_1, t_2 \in T_{\text{psychology}}$ for challenge ${q'}_1^1$ in Table 3 and their top identified paper.

By grounding source domain exploration in domain-agnostic challenges and emphasizing conceptual relevance over methodological similarity, **IDEA-CATALYST** enables the systematic discovery of interdisciplinary perspectives that meaningfully expand the ideation space. The resulting conceptual takeaways form the foundation for the subsequent recontextualization and integration stage, where insights from multiple source domains are mapped back into $\mathcal{D}_{\text{target}}$ to generate candidate interdisciplinary idea fragments.

A.2.5 TARGET–SOURCE INTERDISCIPLINARY INTEGRATION

While interdisciplinary insights from source domains may already inspire researchers to explore new perspectives or refine existing ideas, we further examine whether such insights can be meaningfully integrated with the target domain $\mathcal{D}_{\text{target}}$. This integration step allows us to assess which insights extend beyond inspiration to support concrete, cross-domain synthesis, and thus exhibit strong interdisciplinary potential.

Idea Fragments. We define an *idea fragment* as a structured, intermediate representation capturing a conceptual mechanism, principle, or strategy extracted from a source domain and recontextualized for the target domain. Formally, an idea fragment links: (i) a target-domain research challenge

with its relevant retrieved papers, (ii) source-domain conceptual takeaways & corresponding literature, and (iii) a rationale describing how the takeaway could address the target-domain challenge. Idea fragments are intentionally incomplete: they are designed to support creative ideation rather than prescribe full solutions, aiming to capture promising directions for synthesis by articulating how concepts from $\mathcal{D}_s$ could be combined with existing approaches in $\mathcal{D}_{\text{target}}$.

Generating Integrated Fragments. For each eligible question–source-domain pair $(q_i, \mathcal{D}_s)$, we integrate the most relevant source-domain takeaways and their corresponding papers with the relevant literature retrieved from $\mathcal{D}_{\text{target}}$. Integration is guided by three considerations: ***(i)*** how target-domain methods and assumptions can be complemented by source-domain perspectives, ***(ii)*** how the combined view addresses the specific challenge underlying q_i, and ***(iii)*** how limitations of either domain are mitigated through synthesis. The outcome is an idea fragment f_i that proposes a concrete pathway for interdisciplinary integration.

Ranking Interdisciplinary Potential. Because multiple idea fragments $\mathcal{F} = \{f_i\}$ may be generated for a given research problem p (from different q_i and $\mathcal{D}_s$, we introduce the notion of *interdisciplinary potential* to prioritize fragments that are most likely to yield impactful cross-domain advances. Interdisciplinary potential reflects a fragment's expected value along dimensions such as depth of integration, degree of multi-stage disciplinary engagement, innovation payoff, and balance between novelty and feasibility Okamura (2019); Porter & Rafols (2009).

Rather than assigning absolute scores, we conduct pairwise comparisons between idea fragments. Given two fragments $f_a, f_b \in \mathcal{F}$, we assess which fragment exhibits stronger interdisciplinary potential based on the above criteria. Aggregating preferences across all pairwise comparisons yields a ranked ordering of $\mathcal{F}$, from strongest to weakest interdisciplinary potential. This relative evaluation avoids the need for a single scalar metric, while still prioritizing the most promising integrated ideas. By structuring interdisciplinary exploration through a metacognition-driven framework, **IDEA-CATALYST** supports early-stage research ideation while preserving both novelty and rigor.

A.3 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 = 0.7) (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.

A.3.1 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.

A.3.2 DATASET

We evaluate IDEA-CATALYST using the CHIMERA dataset (Sternlicht & 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. We further prevent knowledge leakage by restricting retrieval (Section A.2.1) to papers published strictly before the arXiv posting year of each instance.

To enable direct comparison between generated idea fragments and ground-truth interdisciplinary contributions, we pre-process each sample into a structured representation aligned with our framework's output format (Appendix A.6) using `Qwen3-14B` (`no-thinking, temperature=0`). The model is strictly constrained to extract and reorganize information already present in the abstract and its annotated context, and ignoring any experimental results, thereby preserving the original interdisciplinary intent while enabling fair, structure-aligned evaluation. For the human study, we ask each participant to provide a brief (1–2 sentence) description of a research problem they are currently working on or have worked on previously, along with the corresponding target domain.

A.4 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 & 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 conduct a comparative evaluation against the ground truth. For each sample, we present both the generated output and the ground truth to an LLM judge, which determines which output better satisfies the evaluation criteria. We report the overall win rate against the ground truth across all samples, where a win indicates that the generated output is preferred over the ground truth.

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 B. 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.

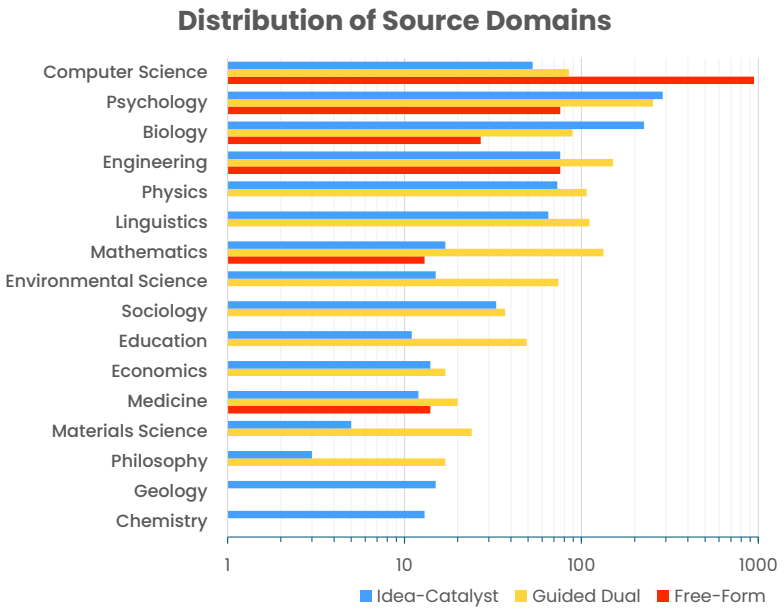

Figure 3: Source-domain distributions (log-scale) for each method's top three ideas.

A.5 SUPPLEMENTAL ANALYSIS

A.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). 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.

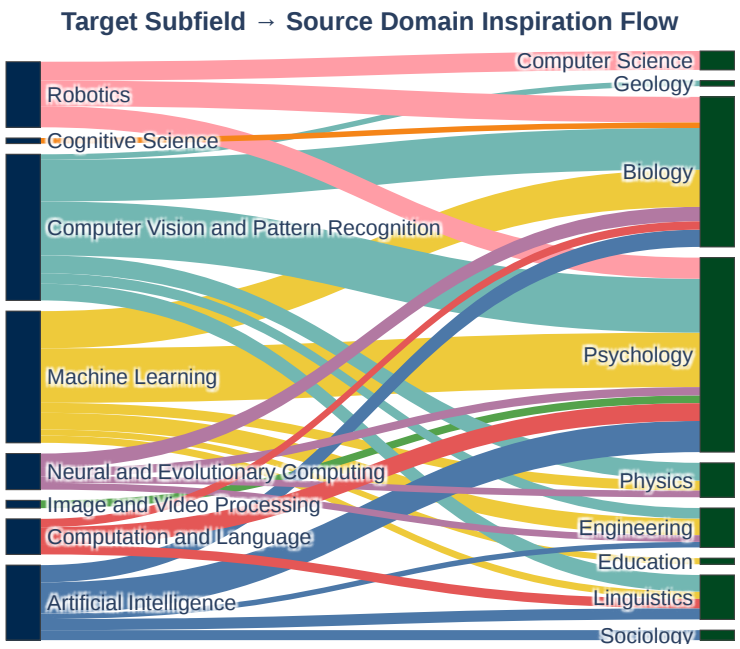

Figure 4: Target-source flow of interdisciplinary inspiration.

Table 5: **Qualitative Comparison of Source-Domain Takeaways.** Exact source-domain formulations and mechanism explanations for *Guided Dual* and IDEA-CATALYST.

Method	Source-Domain Formulation	Mechanism Explanation
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.

A.5.2 QUALITATIVE COMPARISON OF TAKEAWAYS

Table 5 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, the nature of the extracted takeaways differs substantially. *Guided Dual* selects a broadly applicable *Theory of Mind* formulation that reflects well-established, high-level psychological concepts, but remains relatively generic and loosely tied to the specific challenges of open-ended, co-creative collaboration. In contrast, IDEA-CATALYST surfaces a more targeted and problem-aligned takeaway centered on *reciprocal information flow* and *dynamic role distribution*, directly addressing coordination, role adaptation,

Table 6: 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?

and learning dynamics central to the research problem. This contrast suggests that beyond identifying relevant source domains, IDEA-CATALYST's metacognitive guidance enables more precise extraction of interdisciplinary insights meaningful for the target domain.

A.5.3 HUMAN STUDY & 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. Overall, participants found IDEA-CATALYST to be a useful ideation aid, particularly in identifying meaningful research questions and surfacing interdisciplinary perspectives. On average, researchers rated the relevance of the generated research questions highly (4.00/5), indicating that the system effectively captured core challenges in their problem formulations. Retrieved papers were also rated favorably (3.50/5), suggesting that retrieval was generally aligned with the researchers' interests and problem context.

At the level of source-domain reasoning, takeaways were rated as moderately relevant (3.13/5) and insightful (3.16/5), with especially positive feedback from researchers working on problems that are naturally interdisciplinary (e.g., persuasion susceptibility of LLMs). Several participants reported that the takeaways introduced concepts they were motivated to further explore independently. Interpretability received an intermediate score (2.78/5), indicating that rationales were generally clear or mostly clear, with only minor ambiguity on average. Despite this, post-study interviews revealed that participants still perceived the takeaways and ideas as verbose, even after one round of conceptual rewriting. This highlights a central challenge in interdisciplinary ideation: balancing accessibility and brevity with the need to preserve critical technical and conceptual detail when translating ideas across domains. Future work could explore personalization strategies that adapt the level of abstraction and explanation to a user's background and target domain.

Finally, participants rated generated ideas as more novel (3.22/5) than useful (3.00/5), echoing prior findings in Si et al. (2025), namely that creative research ideas do not always translate directly into immediately actionable solutions. Taken together, these results suggest that IDEA-CATALYST is

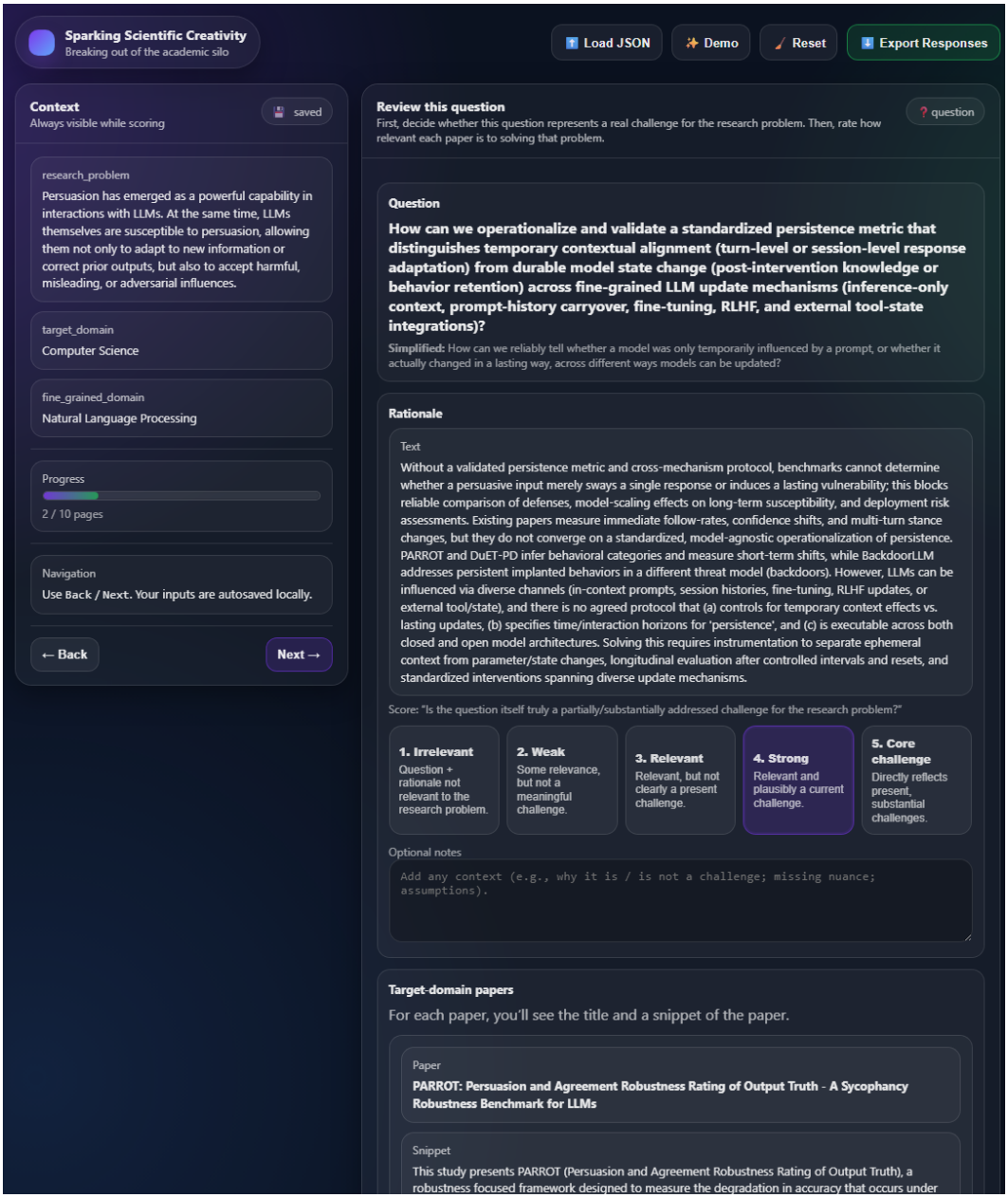

Figure 5: Screenshot of evaluation interface (reviewing questions/challenges).

effective at supporting early-stage exploratory thinking and conceptual reframing, while also revealing opportunities to further improve conciseness, grounding, and user-adaptive explanation in future iterations.

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 Figure 5.

A.6 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"
    ]
  }
}
}
```

B 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:
    - 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": "12 sentences explaining your reasoning for the
      ↪ preferred method"
    }},
    "interdisciplinary_relevance": {{
      "preferred_method": 1 | 2,
      "reasoning": "12 sentences explaining your reasoning for the
      ↪ preferred method based
      on the evaluation criteria"
    }},
  }},
  "overall_assessment": {{
    "preferred_method": 1 | 2,
    "summary": "23 sentences explaining which methods 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"
  }}
}}
```