SCIDEATOR: Human-LLM Compound System for Scientific Ideation through Facet Recombination and Novelty Evaluation

Marissa Radensky*, Simra Shahid*
marissaradensky@gmail.com,simrashahid@microsoft.com
University of Washington, Microsoft
USA, India
* **Equal Contributors**

Raymond Fok
rayfok@cs.washington.edu
University of Washington
USA

Pao Siangliulue
paos@allenai.org
Allen Institute for AI
USA

Tom Hope†, Daniel S. Weld†
tomh@allenai.org,danw@allenai.org
Allen Institute for AI
USA
† **Equal Advisors**

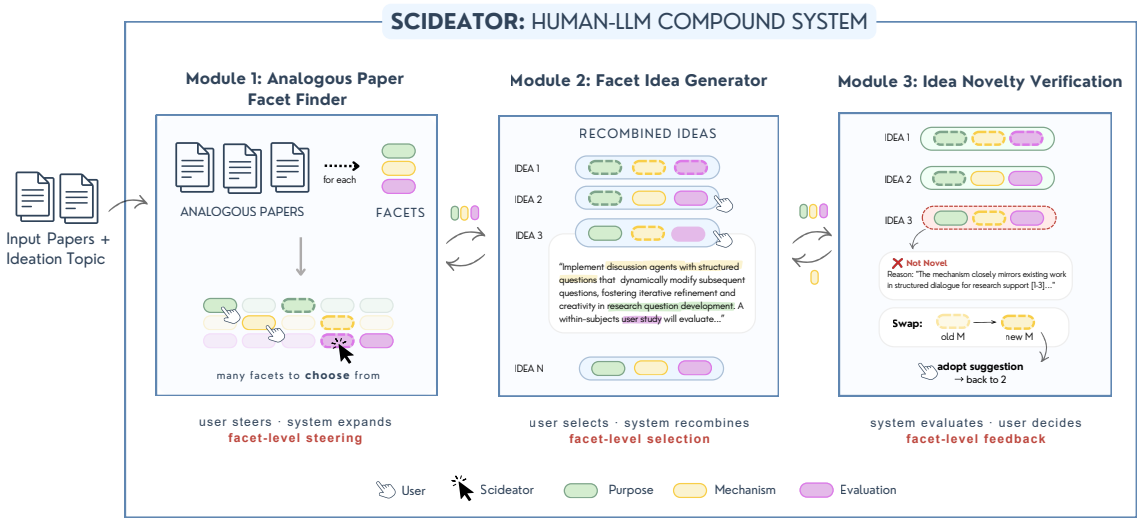

Figure 1: SCIDEATOR is a human-LLM system for scientific ideation. The user and system interact through idea facets (purposes, mechanisms, evaluations): in Module 1, with the user's input papers and an optional topic, the system retrieves analogous papers and extracts facets; the user may select interesting facets or leave it to the system. In Module 2, the system recombines these facets into many ideas via analogy; the user selects ideas to pursue. In Module 3, the system evaluates ideas for novelty against literature retrieved according to facet-based relevance; if "not novel," it suggests a facet swap. The user can adopt the suggestion, return to Module 2 to regenerate, or return to Module 1 to select or add new facets—an iterative ideation loop.

Abstract

The scientific ideation process often involves blending salient aspects of existing papers to create new ideas — a framework known as facet-based ideation. We contribute SCIDEATOR the first human-LLM system for facet-based scientific ideation. Starting from a user-provided set of scientific papers, SCIDEATOR extracts key facets—purposes, mechanisms, and evaluations—from these and related papers, allowing users to explore the idea space by interactively recombining facets to synthesize inventive ideas. SCIDEATOR is driven by three design choices: (1) human-in-the-loop facet recombination, in which users select facets from retrieved papers and the system generates ideas by finding analogies across them via the *Faceted Idea Generator* module; (2) distance-controlled retrieval via the *Analogous Paper Facet Finder* module, which surfaces papers from the same topic to entirely different subareas to provide a spectrum of creative directions; and (3) facet-based novelty verification via the *Idea Novelty Checker* module, a retrieve-then-rerank pipeline that evaluates idea originality using facets. In a user study with computer science researchers, SCIDEATOR provided significantly more creativity support than a baseline using the same backbone LLM without our facet-based modules, particularly in idea exploration and expressiveness. Participants' favorite ideas more often included facets selected by themselves rather than the LLM, and participants used fewer free-text instructions with SCIDEATOR, indicating a preference for facet-level steering over prompting. Finally, re-ranking papers by facet matching rather than general relevance improved novelty classification accuracy from 13.79% to 89.66%.

1 Introduction

Research papers are major sources of scientific inspiration, exposing scientists to concepts that can be recombined into new ideas [9, 31, 47, 54], but discovering inspirations is increasingly challenging due to the ever-expanding literature [7, 29] and cognitive fixation that biases scientists toward familiar thinking [18, 49]. This challenge has spurred interest in automated systems to assist researchers in discovering new ideas [21, 38, 53]. Prior work has highlighted the important role of faceted representations for ideation systems: decomposing ideas into their constituent facets has been used to facilitate creative exploration of design spaces and discover abstract structural linkages across ideas [12, 31, 56]. Example idea facets include purposes, which describe problems addressed, and mechanisms, which describe proposed solutions [25, 27]. In the scientific domain, a small body of work has demonstrated the effectiveness of these facets for finding analogies between research papers for idea inspiration (e.g., papers with similar purposes using different mechanisms) [9, 31, 47]. However, this line of work stopped at surfacing analogous papers as inspirations with no interface for applying the inspirations to synthesize recombinant ideas or for evaluating the generated ideas vis-à-vis existing literature to assess novelty. These important and cognitively taxing tasks were left to the scientists, with no support.

Meanwhile, large language models (LLMs) raise the prospect of quickly synthesizing and evaluating ideas. Recent work has demonstrated their promise in human-AI interfaces for scientific ideation support [35, 37, 48]. However, none of these human-LLM works explored *facet-based* scientific idea generation, despite the fundamental role facets play in ideation literature. Furthermore, none evaluated interfaces for assessing novelty of generated ideas compared to existing papers.

We present Scideator, the first human-LLM compound system for facet-based scientific idea generation and novelty evaluation. Scideator composes three LLM-powered modules — *Analogous Paper Facet Finder* for distance-controlled retrieval of facets, *Faceted Idea Generator* for analogy-based idea generation, and *Idea Novelty Checker* for novelty verification — into a system where the human directs each stage. The system uses a faceted representation (purposes, mechanisms, and evaluations) across all modules: the same representation used to describe retrieved papers is also used to compose ideas, rank related work for novelty assessment, and suggest modifications for non-novel ideas.

The user actions are also at the facet level — selecting a purpose, swapping a mechanism, adopting a novelty suggestion — which propagates to each module, giving a structured signal rather than long textual feedback for models to interpret. With our novel system design and user study, we fill an important knowledge gap regarding the potential of faceted representations for human-AI compound systems for creative tasks such as scientific ideation.

To use Scideator (Fig 2), scientists provide input papers and an optional ideation topic. Scideator extracts key facets (purpose, mechanism, and evaluation) from the input papers; the evaluation facet describes the paper's method to determine if the mechanism successfully addressed the purpose. Scideator retrieves papers with purpose-mechanism pairs analogous to the input papers' overarching purpose and mechanism. Scientists work with Scideator to select candidate facets from retrieved and input papers for recombination. The tool generates analogies involving candidate facets and produces ideas based on the most promising ones.

Scideator also provides idea novelty classifications with explanations, using relevant retrieved literature and carefully constructed in-context examples labeled by experts. Drawing from an expert annotation study, we observed that novelty judgments can be highly subjective, so we grounded our definition in the facet representation used throughout the system: an idea is novel if it differs from retrieved papers in at least one core facet, uniquely combines these facets, or applies them to a new domain. We use this definition to compute a structured, facet-based comparison to existing literature. The system also provides suggestions for improving the novelty of ideas, by replacing one of the initial idea's facets to make it novel.

We investigate the impact of facet-based interaction on scientific ideation through a within-subjects user study with 22 computer-science researchers. Participants completed ideation sessions with two tools: Scideator and a baseline that uses the same backbone LLM without our facet-based modules. We analyze how Scideator impacts aspects of the user's creative experience, such as the idea exploration process and the perceived ability to express oneself. Participants experienced significantly more creativity support with Scideator, particularly in exploring different ideas, which they considered the most important factor for creativity support. Importantly, when participants discussed finding new concepts with Scideator, they cited the tool's facets or ideas as the source, whereas with the baseline, they most often cited the input papers as the source. This shows that Scideator extends users' thinking beyond initial perspectives (represented by input papers), while a similar interface without a facet-based interaction tends to provide ideas only within those confines.

Results also suggest Scideator's novelty checker helps to filter unoriginal ideas. Participants tended to lower their idea novelty assessments when Scideator classified the idea as 'not novel,' which they could verify using provided related papers and explanations. Ablation experiments demonstrate that re-ranking papers by specific facet overlap rather than general relevance improved novelty classification accuracy from 13.79% to 89.66%. Together, these results provide the first systematic evaluation of novelty assessment within a human-AI ideation system—examining both the retrieval pipeline and how researchers use novelty judgments in practice.

In summary, we make the following contributions:

- Scideator, the first human-LLM compound system for scientific ideation that maintains a single faceted representation (purposes, mechanisms, and evaluations) across retrieval, generation, novelty evaluation and suggestion, where users can select, modify, or build on the system's outputs at each stage.
- A within-subjects user study (N=22) showing that Scideator provides significantly more creativity support than a baseline using the same LLM without faceted interaction, with evidence that participants discovered new concepts through the system's facets and preferred facet-level steering over prompting.
- Evaluation of our facet-based novelty checker, with ablation experiments showing that facet-based re-ranking improves paper selection and novelty classification. Combined with user study

evidence of how researchers interpret and act on novelty assessments, this provides the first systematic evaluation of novelty assessment within a human-AI ideation system.

The remainder of the paper is organized as follows. We describe Scideator's building blocks and shared faceted representation (Section 2), its end-to-end workflow (Section 3), and experiments evaluating the novelty checker and the user study (Section 4), followed by related work (Section 5) and discussion (Section 6). We encourage readers to consult the appendix for the full user study design along with UI screenshots, implementation details, LLM prompts used across all modules, and sample ideas generated by the system.

2 Building Blocks of Scideator

In this section we describe the individual modules that make up Scideator. We then describe the end-to-end workflow in Section 3, illustrating how these components of facets, analogous facet generation and paper retrieval, idea generation, and novelty assessment work together in a continuous, human-directed ideation loop.

Shared Representation: Paper Facets

Scideator represents ideas and papers using three facets: the purpose (the problem being addressed), the mechanism (the proposed solution), and the evaluation (the method for determining whether the solution works). We add the evaluation facet to the purpose and mechanism facets used in prior work [25, 27] because evaluation is an important part of a research idea. Facets are extracted by prompting an LLM to produce short phrases (no more than 7 words) based on a paper's title and abstract.

This faceted representation is maintained across every stage of the pipeline. The purpose and mechanism facets drive retrieval — the system finds analogous papers by matching these two facets — while all three facets are used to describe retrieved papers, compose ideas, assess novelty, and suggest improvements. For the user, this means working with a single representation throughout; rather than switching between free-text prompts and paper abstracts, they select, combine, and revise the same structured facets at each stage. For the system, these user interactions at the facet level become precise signals that directly shape what the next module produces.

Next we describe the three modules that leverage this shared representation for different stages of the ideation pipeline.

Module 1: Analogous Paper Facet Finder

This module takes a set of papers and retrieves analogous papers based on their purpose-mechanism pairs at controlled conceptual distances, then extracts all three facets (including evaluation) from the retrieved papers.

Consider the example in Figure 2: the input paper has purpose support research question development and mechanism large language model co-creation. The module first extracts this purpose and mechanism, then generates analogous purpose-mechanism pairs at three distance levels:

- **Near** (same topic): pairs that address a similar problem with a different approach within the same research topic. For instance, the pair (purpose: support crime story hypothesis generation and the mechanism: discussion agents with structured questions, both involve guiding a user through structured dialogue to develop ideas, but in a different application context.
- **Far** (same subarea, different topic): pairs from a related area of computer science that share a more abstract structural parallel with the input. In Figure 2 the analogous paper has purpose enhance human creative expression and mechanism ai-assisted storytelling tools, still within human-AI collaboration but addressing a different topic (creative expression rather than research question development).
- **Very far** (different subarea entirely): pairs from a different area of computer science, connected only by a high-level analogy to the input.

The module generates four facet-pairs at each distance level, giving the user a range of options to explore at each level of similarity. Each of these 12 pairs (3 distances x 4 facet pairs) is then grounded in the literature. For each pair, the LLM generates a search query, and the module retrieves matching papers from Semantic Scholar, shortening the query iteratively if no results are found. The first retrieved paper becomes the representative for that pair and the remaining three provide additional context. Separately, four **very near** papers (most similar to the input) are retrieved directly via Semantic Scholar's paper similarity endpoint [32]. The LLM then extracts purpose, mechanism, and evaluation facets from all retrieved papers (16 total across four distance levels)[1]. The module also produces a summary of relevant works from the input and very-near papers, which the idea generator (Section 2) uses to differentiate new ideas from existing work.

Module 2: Faceted Idea Generator

This module takes a pool of facets from input and retrieved papers and generates ideas by recombining them via analogy, in three steps. First, for a pair of papers from different distance groups (e.g., one input and one far), the LLM generates six candidate analogies between their purpose-mechanism pairs. Continuing the earlier example, one analogy between the input paper purpose support research question development and the near paper mechanism discussion agents with structured questions might be: "both involve using structured dialogue to iteratively guide a user toward generating new ideas." Second, the LLM selects the two strongest analogies based on idea quality (understandability, relevance, feasibility, specificity, and novelty). Third, each selected analogy is converted into an idea: one combines the input paper's purpose with the analogous paper's mechanism, and the other does the reverse (Figure 2, steps 3-4).

The module is designed to combine papers from different distances to produce a range of ideas. When combining with near papers, ideas tend to stay within the same research area but apply a different approach. When combining with far or very far papers, ideas cross domain boundaries. The module adapts to user selections: (i) when **no facets are selected**, the user is likely exploring,

[1] A small-scale annotation study assessing the consistency of the system's distance labels is reported in Appendix C.

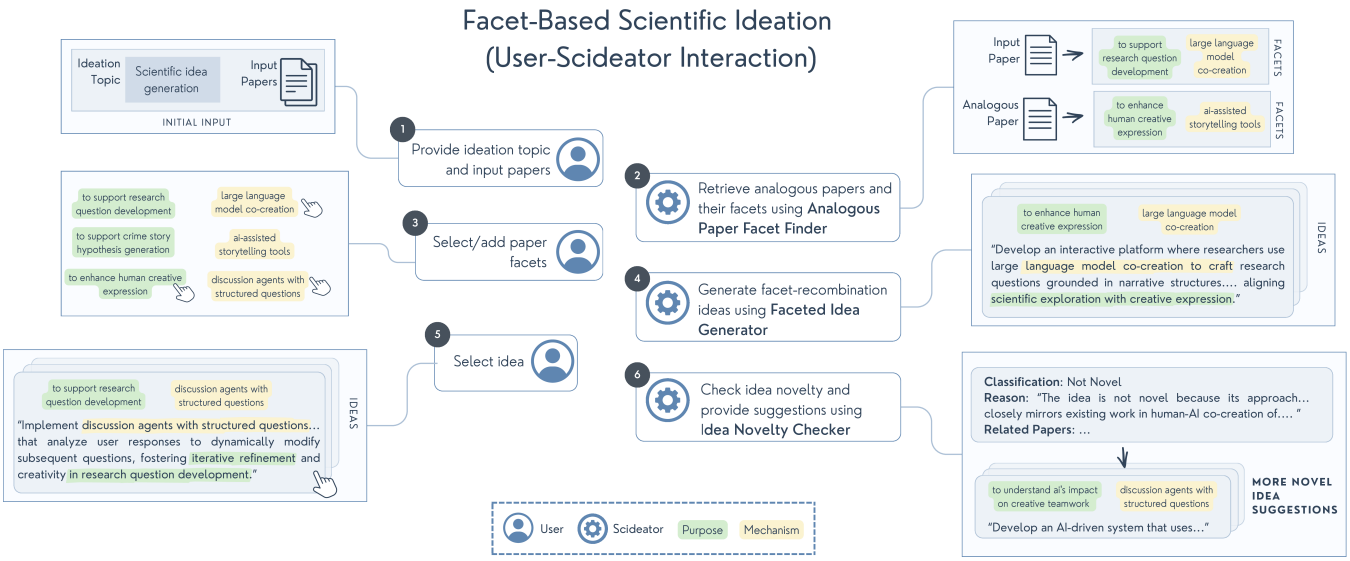

Figure 2: Scideator is a human-LLM compound system for facet-based scientific ideation. 1) The user provides an ideation topic and input papers. 2) Scideator retrieves analogous papers and extracts facets from both input and analogous papers. The same faceted representation — purpose and mechanism — flows across all stages. The evaluation facets are omitted in the figure for clarity. 3) The user selects facets or adds their own; if no selection is made, the system selects automatically. 4) Scideator generates ideas by recombining selected facets through analogy. 5) The user selects an idea to evaluate. 6) Scideator classifies the idea as "novel" or "not novel" with a rationale referencing specific papers. The user reviews the classification and can adjust it. If "not novel," Scideator suggests alternatives by swapping one facet. The user can adopt a suggestion and return to step 3, creating an iterative loop.

so the module combines facets from input and very-near papers with facets from near, far, and very-far papers to maximize diversity; (ii) when only a **purpose or mechanism is selected**, the user has a direction in mind but wants the system to fill in the rest, so the module pairs the selected facet with complementary facets from papers at different distances; and (iii) when **both are selected**, the user has a specific combination in mind, and the module combines them directly. In all cases, the module differentiates ideas from existing work described in the relevant-works summary and self-critiques against quality criteria before finalizing each idea.

Module 3: Idea Novelty Verification

This module takes an idea and determines whether it is novel relative to existing literature. It operates through a four-step retrieve-then-rerank pipeline (Figure 12 in Appendix): 1) retrieve candidate relevant pa- pers, 2) select most relevant papers, 3) evaluate idea novelty, and 4) suggest more novel ideas.

Step 1: Retrieve candidate relevant papers. The module assembles a broad collection of papers that might overlap with the idea. This includes all papers from prior modules and their related papers via the Semantic Scholar API [32]. However, simple retrieval methods often overlook contextual aspects of ideas such as their purpose, mechanism and evaluation facets [42, 43, 61] when assessing similarity between ideas-papers. To expand the coverage of papers, the module generates keyword-based search queries from the idea (LLM-extracted keywords and potential titles) and retrieves matching papers, a popularly used query-based retrieval method also used in [38, 53]. Because keyword searches can introduce irrelevant results, the module complements them with Semantic Scholar's snippet-text search[2], which matches the full idea text against passages from papers to identify contextually relevant work that keyword queries may miss.

Step 2: Select most relevant papers. To surface papers most likely to overlap with the idea, the module applies a two-stage re-ranking process following established retrieve-then-rerank practices [1, 5, 20, 40, 45, 58]. The first stage uses **embedding-based filtering**: SPECTER embeddings [14] compute semantic similarity between the idea and each candidate paper, selecting the top N (default 100). This embedding-based ranking efficiently narrows down the paper collection but, compared to LLMs [51], fails to capture more contextual relationships between different facets of the idea and related papers. The second stage applies **facet-based LLM re-ranking** using RankGPT [59]. Unlike general relevance re-ranking, this step compares each paper against the idea's specific application domain, purpose, mechanism, and evaluation. This facet-based ranking reflects our novelty definition (Step 3): an idea is not novel if its core facets overlap with existing work, so papers with more facet overlap are most relevant for assessment. Papers are ranked by decreasing facet overlap: (1) papers matching all key facets, (2) papers matching application domain and purpose, (3)

[2] api.semanticscholar.org/api-docs/#tag/Snippet-Text

papers matching purpose, mechanism, or evaluation, and (4) papers with partially matched or related facets. The top k papers (default 10) proceed to novelty assessment.

STEP 3: EVALUATE IDEA NOVELTY. Using the top-k papers, the module prompts an LLM to classify the idea as "novel" or "not novel," accompanied by reasoning that references specific papers. The prompt incorporates expert-labeled in-context examples, each comprising an idea, the top-k papers, the novelty label, and a classification reason (see Appendix C.2 for examples). The examples encode our novelty definition: an idea is novel if it (1) differs from all retrieved papers in at least one core facet, (2) uniquely combines these facets, or (3) applies them to a new domain.

STEP 4: SUGGEST MORE NOVEL IDEAS. When an idea is classified as "not novel," the module generates three suggestions (one per facet type) for more novel alternatives. Each suggestion replaces a different facet in the original idea (purpose, mechanism, or evaluation) with another available facet, aiming to increase novelty relative to the retrieved papers.

Expert Annotation Study. The in-context examples and novelty definition above emerged from an annotation study in which the first two authors assessed the novelty of 51 ideas (46 generated by SCIDEATOR, 5 adapted from OpenReview) based on top-10 retrieved papers. An initial round using three categories (novel, moderately novel, not novel) achieved moderate agreement (Cohen's $\kappa = 0.64$), with disagreements arising partly because annotators drew on background knowledge rather than the retrieved papers. This led us to simplify the scheme to binary (novel / not novel), require judgments based solely on retrieved papers, and adopt the facet-based definition used throughout the novelty checker. The revised round yielded higher agreement (Cohen's $\kappa = 0.68$). From both rounds, we collected 67 consensus-labeled examples (39 novel, 28 not novel), from which we sampled in-context examples[3] and held out the remainder for the retrieval ablation (Section 4.1).
Implementation details for each module including model choices, API parameters, and default settings are in Appendix E and LLM prompts in G.

3 End-to-End Workflow

The modules above compose into SCIDEATOR's ideation loop in which the system proposes and the scientist steers (Figure 2). A scientist begins by providing an ideation topic and one or more input papers. SCIDEATOR runs the retrieval module, extracting facets from the input papers and retrieving analogous papers at four conceptual distances. The system presents all extracted facets organized by distance (Figure 9 in Appendix).

Because facets are short phrases (up to 7 words), their meaning may not always be immediately clear, especially for facets from distant domains. To support interpretation, each facet is linked to its source paper, so users can trace where a purpose or mechanism originated. Hovering over a facet reveals a longer description. Users can also type in their own facets or request additional facets with an optional guiding query, which triggers a new retrieval in the direction specified by the query.

[3]We tested with 10, 15, 20 examples per class. The module uses 20 examples per class to assess novelty.

The scientist then selects facets and triggers idea generation—or generates without selecting, in which case the system chooses facets automatically. The generator adapts to these selections as described in Section 2: the specific combination of selected facets determines which papers and distances the module draws from. Each generated idea is displayed with its constituent facets color-coded by type (Figure 9 in Appendix). Users can click *Expand* on any idea to see a more detailed version, or add their own idea, which the system decomposes into facets and adds to the available pool for future rounds.

To evaluate an idea's novelty, the scientist opens the novelty checker (Figure 8 in Appendix). The system retrieves and re-ranks related papers, classifies the idea, and if its "not novel", it suggests three alternatives, each replacing a different facet. The scientist can adjust the classification based on their own judgment, adopt a suggestion, or return to facet selection with new perspective.

This creates an **iterative loop**. A scientist might generate ideas, check one for novelty, learn it overlaps with existing work, adopt a suggestion that swaps the mechanism, and regenerate — each step expressed at the facet level. Facets selected in one round may be discarded after evaluation; suggestions from the novelty checker introduce facets the scientist had not considered; user-added facets expand the pool for future rounds.

A key aspect of this design is that every interaction— selecting a facet, adopting a suggestion, adding a custom query — gives the system a structured signal rather than free-text instructions. When the novelty checker classifies an idea as "not novel" because its mechanism overlaps with a paper, and the user adopts a suggestion that replaces that mechanism, the system receives a precise signal about which dimension to change. This contrasts with a text-based LLM interaction where the user writes "make the idea more novel" and the model must interpret both intent and scope. In SCIDEATOR, the faceted representation makes intent explicit and the system's response predictable.

4 Experiments

We evaluate SCIDEATOR in two parts. The first addresses a component-level question: does the faceted representation improve the novelty verification module's ability to identify relevant prior work? This can be assessed through automated comparison against expert labels (Section 4.1). The second addresses a system-level question: does the faceted interaction help researchers generate and evaluate scientific ideas? This requires a controlled user study (Section 4.2).

4.1 Retrieval Ablation

Setup. Using the 32 held-out examples from the expert annotation study (described in Section 2), we evaluate whether the retrieval and re-ranking pipeline surfaces the papers needed to correctly identify non-novel ideas. We evaluate on 58 ideas: 13 "not novel" instances from the held-out set and 45 published NLP papers. For all these ideas, specific overlapping papers exist in literature — the question is whether each pipeline variant surfaces them in the top 10. For "novel" ideas, the classification depends on the *absence* of overlapping work, which can shift with different retrieved sets, so we focus on "not novel" instances only. We use o3-mini for novelty

classification (Step 3) for its better reasoning capabilities and gpt-4o for re-ranking (Step 2). We compare five configurations:

- **Complete System:** keyword + snippet retrieval, embedding filtering, facet-based RankGPT.
- **Relevance RankGPT:** same retrieval and filtering, but facet-based re-ranker replaced with general relevance re-ranking [59].
- **Embedding Filtering:** no LLM re-ranker.
- **Snippet Retrieval:** top-k from snippet search only, no re-ranking and embedding filtering.
- **Keyword Retrieval:** top-k from keyword search only, no re-ranking and embedding filtering.

This setup allows us to isolate the contribution of each component (retrieval method vs. re-ranking strategy) and evaluate whether they collectively brought key papers for novelty assessment into the top 10.

Classification Analysis: Table 1 reports how often each configuration correctly classifies the 58 "not novel" ideas — that is, whether the top-10 papers retrieved by that configuration provide enough evidence for the classifier to agree with the expert consensus label. We observe that the complete system, with facet-based LLM reranking, significantly outperforms its ablated variants in accuracy. The results demonstrate that methods relying only on keyword or snippet-based retrieval have much lower accuracy, and even alternate re-ranking strategies with a single embedding-based reranker or both embedding and general relevance RankGPT are insufficient to consistently bring key papers into the most relevant paper set. These findings show that combining facet-based reranking with embedding is critical for identifying the most relevant papers.

Table 1: Proportion of the 58 "not novel" test ideas correctly classified as "not novel" by Step 3, given each variant's top-10 retrieved papers.

Method	Accuracy
Complete System	89.66%
- Relevance RankGPT	13.79%
- Embedding Filtering	10.34%
- Snippet Retrieval	8.62%
- Keyword Retrieval	5.17%

Analysis of the Most Relevant Papers: Table 2 compares the top-10 most relevant papers retrieved under each ablation setting with those from the complete system. Approximately 30% of the papers differ when using either embedding-based or general relevance RankGPT. Additionally, notable rank shifts are observed between the facet-based and relevance-based LLM rerankers. In contrast, without the reranking steps, both snippet and keyword retrieval exhibit minimal overlap with the final system's top results, highlighting the importance of the reranker stage.

The ablation demonstrates that the faceted representation is critical for the system's internal pipeline. We next investigate whether it also benefits the human-facing interaction.

4.2 User Study

To assess whether Scideator's faceted interaction supports scientific ideation in practice, we conducted a within-subjects study comparing Scideator to a baseline tool without faceted representation. 22 computer-science researchers participated, each completing two 20-minute ideation sessions — one with Scideator and one with the baseline — in randomized order, using assigned ideation topics and starting papers from the HCI and NLP domains.

Table 2: Comparing rank and overlap in retrieved papers with each variant to the complete system. Overlap indicates how many papers overlap on average with the complete system top-10 papers. Rank Shift measures the average absolute difference in rank positions (only among overlapping papers).

Method	Overlap (↑)	Rank Shift (↓)
Relevance RankGPT	7.97	0.67
Embedding Filtering	7.93	0.84
Snippet Retrieval	2.88	1.85
Keyword Retrieval	1.17	1.39

The baseline represents *paper-level interaction*: participants provide papers and free-text instructions, and the model generates ideas directly. Both tools use the same LLM, so the comparison isolates the effect of organizing interaction around a structured faceted representation versus free-text prompting. After each session, participants rated their experience using the Creativity Support Index (CSI) [11], a validated instrument that measures factors of creative support such as exploration, expressiveness, enjoyment. We also collected interaction logs, survey responses and conducted semi-structured interviews analyzed through inductive thematic analysis [8]. Full study design, baseline tool description, procedure, and participant demographics are in Appendix A.

We address three research questions. RQ1 tests whether faceted interaction leads to more creativity support. RQ2 examines how participants use facet-level versus paper-level interaction. RQ3 tests whether the novelty checker influences participants' assessments. Sample ideas generated by participants using both Scideator and the baseline are provided in Appendix D.

4.2.1 *RQ1: Does Faceted Interaction Lead to More Creativity Support for idea generation?*

Participants experienced significantly more creativity support with Scideator than the baseline (Wilcoxon signed-rank test, p<.01; CSI scores on a 0–100 scale: Scideator median=70.5, baseline median=61.0)[4]. Of the CSI factors, participants benefited most from Scideator in **exploration** and **expressiveness** — the two factors participants also ranked as most important for ideation task (Figure 5). Results for factors such as immersion, showed no differences and are reported in Appendix B.1.

Exploration Factor. Scideator helped participants explore ideas beyond their input papers. While participants rated their favorite ideas' newness on a 7-point scale, showing only a slight advantage for Scideator (median difference: 0.5 points), the qualitative evidence was striking: in the baseline, participants who found new concepts attributed them to the *input papers* rather than the tool's output (6 of 8), while in the treatment, all participants who found

[4]CSI scores: Scideator median=70.50 (Q1=57.50, Q3=79.00), baseline median=61.00 (Q1=42.25, Q3=71.50). Wilcoxon signed-rank test, V=208.50, p<.01. The data did not violate the assumption of symmetry of within-subjects differences about the median.

new concepts cited the *tool's facets or generated ideas* as the source (6 of 6). For example, P18-HCI-treatment shared,

> "*When I thought of human-AI collaboration in art, for example, I did not think about also supporting artistic pursuits of students [which surfaced in a generated idea].... When I was thinking about the topic, I thought more about... a human prompting an AI for generating images or for image exploration which is more related to the papers that were given.*"

Meanwhile, P5-NLP-baseline reflected,

> "*The papers themselves were really interesting, but I don't think the tool generated anything super beyond a synthesis of the ideas that were in those three papers.*"

This pattern suggests the faceted representation served as a medium for new concepts to reach the user. The retrieval module surfaced papers the user had not started with, the extraction module decomposed them into the same facet format the user was already working with, and participants encountered new concepts through these facets. Four participants also noted that the facet-level interaction supported exploration through **greater transparency** as they could trace which facets contributed to each idea and more easily context-switch between research directions. For instance, P5-NLP-treatment reflected,

> "*I think the first thing that I noticed was that it was very easy to context switch. That was my main problem with the [other] tool before. I couldn't figure out which idea dealt with what aspect of the research that I was engaging with. Very easy to do that here.*"

Expressiveness Factor. Fourteen participants found SCIDEATOR's facet-level interaction useful or interesting, with seven specifically noting **increased control** over idea generation. P11-NLP-treatment explained,

> "*I like this tool better because it sort of distilled the different aspects of the input papers into very concrete blocks that you could plug into each other.... it's just that the information was presented in this tool... in a more digestible manner, and that helped combine information across papers better.*"

This preference for structured interaction is also reflected in participants' use of custom instructions. With SCIDEATOR, participants most often did not provide custom text instructions for generating their favorite ideas (median: 0 of 2 ideas). With the baseline, the median participant used custom instructions for 1.5 of 2 favorite ideas. As P1-HCI-treatment put it:

> "*I didn't need to add any custom instructions because these [facets] served like custom instructions.*"

The faceted interface provided a structured alternative to free-text prompting, and participants preferred it.

4.2.2 ***RQ2: How do participants use Facet-Level versus Paper-Level Interaction?*** Participants commented on benefits of both the baseline tool and SCIDEATOR in terms of their input granularity. Fourteen participants found SCIDEATORś affordance for facet-level interactions useful or interesting, noting increased control and greater transparency (Section 4.2.1) as advantages of the facet-level interaction. On the other hand, five participants appreciated the baseline's paper-level interactions in addition to or more than SCIDEATOR's facet-level interactions. Three of these participants liked the paper-level interaction, as it felt more directly connected to the literature. P22-NLP-baseline explained,

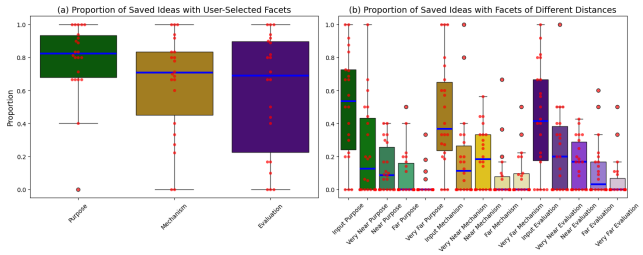

Figure 3: (a) Participants more often opted to select their own facets rather than let the LLM select for them. (b) Participants used input facets and facets nearer to the input more than facets farther from the input.

> " *I think papers for me were more natural than facets.... I think to me it's more like a map of literature, so I could see it more with papers..*"

Facets Selections. When generating ideas in SCIDEATOR, users have the flexibility to either manually curate facet combinations (using system-extracted or user-provided facets) or rely on automated system selection. Participants' favorite ideas more often included evaluations, mechanisms, and especially purposes selected by themselves rather than the LLM (Figure 3a). Given that participants were assigned ideation topics, P18-HCI-treatment explained why participants may have decided to prioritize selecting purposes themselves:

> "*I think the purpose is the most relevant to the topic. So within an area, there can be many ways of doing the same tasks, but the task is ultimately what defines the area.*"

Distance preferences. Participants used input and near facets substantially more than far facets primarily for purposes (Figure 3b). The primary reason for not using far facets was lower perceived relevance to the ideation topic, though four participants found them helpful for discovery. P9-NLP-treatment commented,

> "*that very near, near, far kind of thing... it kind of adds some sort of discovery factor.*"

4.2.3 ***RQ3: Does the Idea Novelty Verification module influence participants' confidence in novelty assessments?*** Among the 17 participants who completed the novelty evaluation task with

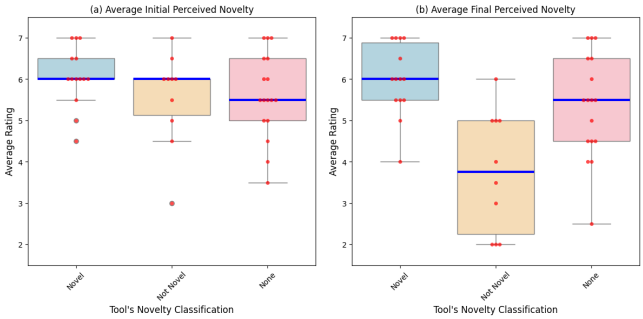

Figure 4: Participants' average perceived idea novelty before (a) and after (b) utilizing their assigned tool for idea novelty evaluation, as well as the average change from initial to final perceived novelty.

the intended setup,[5] SCIDEATOR's novelty checker did not significantly improve overall confidence in novelty assessments (sign test, S=5.00, p=n.s.). However, a between-subjects comparison revealed that participants changed their novelty assessments most when SCIDEATOR classified an idea as "not novel" (Figure 4), suggesting the checker is most useful for filtering unoriginal ideas. This reflects an asymmetry in verifiability: a "not novel" classification can be checked against the retrieved papers, whereas a "novel" classification is harder to confirm. P17-HCI-treatment captured this asymmetry:

> "*Seeing a list of related work is very helpful for giving you the context. It was very convincing in the case of telling me that an idea was not novel.... When it provides [a novel classification], it's less convincing but is helpful.*"

Three participants went further, noting that the retrieved papers were more useful than the classification or reasoning alone. P8-NLP-treatment explained why:

> "*I didn't really pay much attention to these reasons because reasons can be kind of made up to explain why their generation is novel. So, I kind of relied more on the references that it retrieved.*"

5 Related Work

This work builds upon prior work on facet-based ideation, scientific idea novelty evaluation, and the challenges of human-LLM systems for scientific ideation.

[5]One participant did not receive the full allotted time; four experienced a different version of the novelty checker due to an API issue.

5.1 Facet-Based Ideation

Boden [6] distinguishes three types of creativity: combinatorial (combining existing concepts), exploratory (pushing the boundaries of a framework), and transformational (changing the rules themselves). Combinatorial creativity accounts for a substantial portion of scientific progress [60], and a key approach to supporting it in scientific ideation is through combining facets of idea such as purpose (the problem) and a mechanism (the solution) through analogies that draw parallels across contexts [24, 25]. This framework has facilitated analogy identification across research papers [9, 31], product ideas [25, 27], biological designs [30], and research-paper authors [47].

Several tools build on structured representations, each addressing a different stage and domain. For *retrieval*, SOLVENT uses purpose-mechanism annotations to help users find analogous papers [9], with subsequent work adding LLM-based facet extraction [31, 47]. For *exploration*, Luminate extracts response dimensions from LLM outputs and lets users navigate the design space through structured handles [56]. For *recombination*, BIOSPARK supports biological analogies in engineering [30], AnalogiLead supports design problems [55], and CreativeConnect supports graphic design through keyword recombination [12]. Recently, the IdeaSynth [48] system was proposed to support *refinement* of one existing idea by expanding upon coarse-grained aspects of an initial proposal akin to background, method, findings sentences in paper abstracts. In addition to not using fine-grained facets as in our work, IdeaSynth does not support early ideation that starts only with a few papers, does not support generating multiple idea directions, recombining ideas across papers, and does not include novelty assessment. Across all these tools, the representation supports a single stage and does not propagate across the pipeline or connect generation to evaluation.

SCIDEATOR propagates the same faceted representation — purposes, mechanisms, and evaluations — across retrieval, generation, novelty evaluation and suggestion. Because every paper and idea are described in the same terms, users can compare across dozens of directions quickly; and a facet selection during exploration directly shapes which ideas are generated, which papers are retrieved for novelty assessment, and what improvements are suggested.

5.2 Scientific Idea Novelty Evaluation

Evaluating whether a generated idea is novel relative to existing literature is particularly challenging when ideas recombine concepts from unfamiliar sub-areas [16]. Fully automated systems like AI Scientist [38] retrieve papers via keyword similarity and iteratively compare them against the idea, but without structured comparison of which facets specifically overlap. In human-LLM interaction, Acceleron [44] uses agent personas to assess and improve a proposal's novelty, and ReviewFlow [57] provides in-situ novelty support for peer reviewers. However, both evaluate externally provided ideas and cannot connect evaluation back to generation. Other tools surface related papers alongside generated ideas [35, 37, 48] but do not provide explicit novelty judgments grounded in those papers. Moreover, none of these systems have been systematically evaluated for novelty assessment — existing evaluations either focus on the broader tool rather than the novelty component specifically, or rely

on small-scale qualitative demonstrations rather than controlled comparisons against expert judgments.

SCIDEATOR closes the loop: when an idea is classified as "not novel," the system provides a rationale referencing specific retrieved papers and suggests alternatives by replacing individual facets, feeding evaluation back into generation. We evaluate this component both through automated ablation against expert labels and through a user study examining how researchers interpret and act on novelty assessments in practice.

5.3 Human-LLM Scientific Ideation Systems

Several works have explored fully automating scientific ideation [4, 22, 33, 63], but automated methods remain insufficient for formulating novel, impactful research ideas [26, 28]. This has motivated human-LLM tools [21, 64]: CoQuest supports divergent exploration through plain-text feedback [35], PersonaFlow supports convergent development through persona-driven feedback [37], and Perspectra [36] lets users steer multi-agent deliberation with domain-expert personas, giving users control over *which* agents contribute. In all cases, the human interacts through free-text prompts or by selecting among LLM outputs — there is no structured representation through which user choices propagate across modules. As a result, intent must be re-articulated at each stage, and neither user nor system can trace how a specific input influenced a specific output. This is concerning given evidence that unconstrained LLM generation reduces idea diversity [17, 39].

In SCIDEATOR, users can steer generation through facet-level actions — selecting a purpose, swapping a mechanism, adopting a novelty suggestion — or let the system select facets and step in only when they want to redirect. In either case, interactions propagate as structured signals across all modules, rather than free-text prompts that the LLM must reinterpret at each stage, giving the system precise signals about what to change and giving the user traceability over how their choices shaped the output.

6 Conclusion and Discussion

We presented SCIDEATOR, a human-LLM compound system that maintains a single faceted representation across retrieval, generation, and novelty evaluation for scientific ideation. At each stage, the system does the heavy lifting while the user steers — selecting facets, choosing ideas, interpreting novelty assessments — or lets the system make selections and steps in when they want to redirect. Faceted representations are known as fundamental in ideation processes; SCIDEATOR is the first human-LLM system for facet-based scientific ideation and novelty assessment. Our results show that this representation benefits both (a) the users who preferred facet-level steering over free-text prompting and discovered new concepts and directions through the system's facets, leading to improved idea exploration and expressiveness, and (b) the system's internal modules, where facet-based re-ranking for novelty assessment dramatically outperformed retrieveal and re-ranking with general relevance. While still reliant on LLMs and limited to combinatorial creativity, grounding generation in papers across varying distances and giving users facet-level control takes a step toward maintaining idea diversity.

Faceted interaction extends users' conceptual reach beyond their starting papers. Aligned with our goal of supporting divergent ideation [15, 52], results from our within-subjects study show that participants experienced significantly more creativity support with SCIDEATOR than the baseline, particularly in exploration — the factor they rated most important. The most validating finding is where new concepts came from: with SCIDEATOR, participants attributed new concepts to the tool's facets and generated ideas, while with the baseline, new concepts came from reading the input papers themselves. This suggests through faceted representation, concepts from distant papers became more accessible without requiring users to read those papers in full.

The shared faceted representation benefits both users and system modules. The faceted representation served a dual function: it was both the vocabulary participants used to steer the system and the data structure the system's modules operated on. User actions (selecting a purpose, swapping a mechanism) propagated directly to downstream modules without translation, and system outputs (ideas composed from selected facets, novelty suggestions replacing specific facets) were immediately interpretable to users. Users preferred facet-level interaction over free-text prompting (median 0 vs. 1.5 custom instructions for favorite ideas), and the novelty module performed dramatically better with facet-based re-ranking than general relevance (89.66% vs. 13.79%). The novelty checker was most convincing when it classified ideas as "not novel" — participants could verify the judgment against retrieved papers — while "novel" classifications were harder to trust. Prior work has explored structured representations for individual stages of creative support [9, 48, 56]; our results suggest that maintaining a consistent representation across a full compound pipeline amplifies these benefits.

Future directions. Several avenues remain open. Participants sometimes avoided distant facets because they were unfamiliar with how to apply them, saving ideas with far facets much less often than near ones. Lowering this barrier — for instance through in-situ question-answering about unfamiliar facets or by surfacing the analogy that motivated a generated idea — could help scientists engage with concepts that would not have occurred to them otherwise. Some participants also valued paper-level interaction alongside facets, suggesting that tools offering multiple levels of granularity (papers, facets, individual concepts) may lead to richer interactions. Broader retrieval and richer paper representations could improve the novelty checker's coverage — participants noted that 10 retrieved papers sometimes missed important works. Finally, SCIDEATOR's modules currently operate in a fixed pipeline directed by the user. A natural extension is to give modules more autonomy, for example having the novelty checker proactively suggest facet replacements during generation or allowing an agent to iteratively refine ideas across multiple generation-evaluation cycles before presenting results.

References

[1] Abdelrahman Abdallah, Bhawna Piryani, Jamshid Mozafari, Mohammed Ali, and Adam Jatowt. 2025. Rankify: A Comprehensive Python Toolkit for Retrieval, Re-Ranking, and Retrieval-Augmented Generation. https://api.semanticscholar.org/CorpusID:276107364

[2] Afra Feyza Akyürek, Ekin Akyürek, Leshem Choshen, Derry Wijaya, and Jacob Andreas. 2024. Deductive closure training of language models for coherence, accuracy, and updatability. arXiv preprint arXiv:2401.08574 (2024).

[3] Shm Garanganao Almeda, JD Zamfirescu-Pereira, Kyu Won Kim, Pradeep Mani Rathnam, and Bjoern Hartmann. 2024. Prompting for discovery: Flexible sense-making for ai art-making with dreamsheets. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. 1–17.

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

[5] Davide Baldelli, Junfeng Jiang, Akiko Aizawa, and Paolo Torroni. 2024. TWOLAR: A TWO-Step LLM-Augmented Distillation Method for Passage Reranking. ArXiv abs/2403.17759 (2024). https://api.semanticscholar.org/CorpusID:268691914

[6] Margaret A Boden. 2004. The creative mind: Myths and mechanisms. Routledge.

[7] Lutz Bornmann and Rüdiger Mutz. 2015. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. Journal of the association for information science and technology 66, 11 (2015), 2215–2222.

[8] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. Qualitative research in psychology 3, 2 (2006), 77–101.

[9] Joel Chan, Joseph Chee Chang, Tom Hope, Dafna Shahaf, and Aniket Kittur. 2018. Solvent: A mixed initiative system for finding analogies between research papers. Proceedings of the ACM on Human-Computer Interaction 2, CSCW (2018), 1–21.

[10] Alan Y Cheng, Meng Guo, Melissa Ran, Arpit Ranasaria, Arjun Sharma, Anthony Xie, Khuyen N Le, Bala Vinaithirthan, Shihe Luan, David Thomas Henry Wright, et al. 2024. Scientific and fantastical: Creating immersive, culturally relevant learning experiences with augmented reality and large language models. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. 1–23.

[11] Erin Cherry and Celine Latulipe. 2014. Quantifying the creativity support of digital tools through the creativity support index. ACM Transactions on Computer-Human Interaction (TOCHI) 21, 4 (2014), 1–25.

[12] DaEun Choi, Sumin Hong, Jeongeon Park, John Joon Young Chung, and Juho Kim. 2024. CreativeConnect: Supporting Reference Recombination for Graphic Design Ideation with Generative AI. In Proceedings of the CHI Conference on Human Factors in Computing Systems. 1–25.

[13] Seulgi Choi, Hyewon Lee, Yoonjoo Lee, and Juho Kim. 2024. VIVID: Human-AI Collaborative Authoring of Vicarious Dialogues from Lecture Videos. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. 1–26.

[14] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. SPECTER: Document-level Representation Learning using Citation-informed Transformers. ArXiv abs/2004.07180 (2020). https://api.semanticscholar.org/CorpusID:215768677

[15] Arthur Cropley. 2006. In praise of convergent thinking. Creativity research journal 18, 3 (2006), 391–404.

[16] Douglas L Dean, Jill Hender, Tom Rodgers, and Eric Santanen. 2006. Identifying good ideas: constructs and scales for idea evaluation. Journal of Association for Information Systems 7, 10 (2006), 646–699.

[17] Anil R. Doshi and Oliver P. Hauser. 2024. Generative AI enhances individual creativity but reduces the collective diversity of novel content. Science Advances 10, 28 (2024), eadn5290. doi:10.1126/sciadv.adn5290

[18] Karl Duncker and Lynne S Lees. 1945. On problem-solving. Psychological monographs 58, 5 (1945), i.

[19] James Enouen, Hootan Nakhost, Sayna Ebrahimi, Sercan O Arik, Yan Liu, and Tomas Pfister. 2023. Textgenshap: Scalable post-hoc explanations in text generation with long documents. arXiv preprint arXiv:2312.01279 (2023).

[20] Jingtong Gao, Bo Chen, Xiangyu Zhao, Weiwen Liu, Xiangyang Li, Yichao Wang, Zijian Zhang, Wanyu Wang, Yuyang Ye, Shanru Lin, Huifeng Guo, and Ruiming Tang. 2024. LLM-enhanced Reranking in Recommender Systems. ArXiv abs/2406.12433 (2024). https://api.semanticscholar.org/CorpusID:270562015

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

[22] Tianyang Gu, Jingjin Wang, Zhihao Zhang, and HaoHong Li. 2024. LLMs can realize combinatorial creativity: generating creative ideas via LLMs for scientific research. arXiv preprint arXiv:2412.14141 (2024).

[23] Yuling Gu, Oyvind Tafjord, and Peter Clark. 2023. Digital socrates: Evaluating llms through explanation critiques. arXiv preprint arXiv:2311.09613 (2023).

[24] Keith J. Holyoak and Paul Thagard. 1996. Mental Leaps: Analogy in Creative Thought. MIT Press.

[25] Tom Hope, Joel Chan, Aniket Kittur, and Dafna Shahaf. 2017. Accelerating innovation through analogy mining. In Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. 235–243.

[26] Tom Hope, Doug Downey, Daniel S Weld, Oren Etzioni, and Eric Horvitz. 2023. A computational inflection for scientific discovery. Commun. ACM 66, 8 (2023), 62–73.

[27] Tom Hope, Ronen Tamari, Daniel Hershcovich, Hyeonsu B Kang, Joel Chan, Aniket Kittur, and Dafna Shahaf. 2022. Scaling creative inspiration with fine-grained functional aspects of ideas. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. 1–15.

[28] Peter Jansen, Oyvind Tafjord, Marissa Radensky, Pao Siangliulue, Tom Hope, Bhavana Dalvi Mishra, Bodhisattwa Prasad Majumder, Daniel S Weld, and Peter Clark. 2025. CodeScientist: End-to-End Semi-Automated Scientific Discovery with Code-based Experimentation. arXiv preprint arXiv:2503.22708 (2025).

[29] Arif E Jinha. 2010. Article 50 million: an estimate of the number of scholarly articles in existence. Learned publishing 23, 3 (2010), 258–263.

[30] Hyeonsu B Kang, David Chuan-En Lin, Nikolas Martelaro, Aniket Kittur, Yan-Ying Chen, and Matthew K Hong. 2024. BioSpark: An End-to-End Generative System for Biological-Analogical Inspirations and Ideation. In Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. 1–13.

[31] Hyeonsu B Kang, Xin Qian, Tom Hope, Dafna Shahaf, Joel Chan, and Aniket Kittur. 2022. Augmenting scientific creativity with an analogical search engine. ACM Transactions on Computer-Human Interaction 29, 6 (2022), 1–36.

[32] 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, Regan Huff, Daniel King, Sebastian Kohlmeier, Bailey Kuehl, Michael Langan, Daniel Lin, Haokun Liu, Kyle Lo, Jaron Lochner, Kelsey MacMillan, Tyler C. Murray, Christopher Newell, Smita R Rao, Shaurya Rohatgi, Paul Sayre, Zejiang Shen, Amanpreet Singh, Luca Soldaini, Shivashankar Subramanian, A. Tanaka, Alex D Wade, Linda M. Wagner, Lucy Lu Wang, Christopher Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Madeleine van Zuylen, and Daniel S. Weld. 2023. The Semantic Scholar Open Data Platform. ArXiv abs/2301.10140 (2023). https://api.semanticscholar.org/CorpusID:256194545

[33] Dan Lahav, Jon Saad Falcon, Bailey Kuehl, Sophie Johnson, Sravanthi Parasa, Noam Shomron, Duen Horng Chau, Diyi Yang, Eric Horvitz, Daniel S Weld, et al. 2022. A search engine for discovery of scientific challenges and directions. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36. 11982–11990.

[34] Joanne Leong, Pat Pataranutaporn, Valdemar Danry, Florian Perteneder, Yaoli Mao, and Pattie Maes. 2024. Putting things into context: Generative AI-enabled context personalization for vocabulary learning improves learning motivation. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. 1–15.

[35] Yiren Liu, Si Chen, Haocong Cheng, Mengxia Yu, Xiao Ran, Andrew Mo, Yiliu Tang, and Yun Huang. 2024. How ai processing delays foster creativity: Exploring research question co-creation with an llm-based agent. In Proceedings of the CHI Conference on Human Factors in Computing Systems. 1–25.

[36] Yiren Liu, Viraj Shah, Sangho Suh, Pao Siangliulue, Tal August, and Yun Huang. 2025. Perspectra: Choosing Your Experts Enhances Critical Thinking in Multi-Agent Research Ideation. arXiv preprint arXiv:2509.20553 (2025).

[37] Yiren Liu, Pranav Sharma, Mehul Jitendra Oswal, Haijun Xia, and Yun Huang. 2024. Personaflow: Boosting research ideation with llm-simulated expert personas. arXiv preprint arXiv:2409.12538 (2024).

[38] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The ai scientist: Towards fully automated open-ended scientific discovery. arXiv preprint arXiv:2408.06292 (2024).

[39] Lennart Meincke, Gideon Nave, and Christian Terwiesch. 2025. ChatGPT decreases idea diversity in brainstorming. Nature Human Behaviour 9, 6 (2025), 1107–1109. doi:10.1038/s41562-025-02173-x

[40] Chuan Meng, Negar Arabzadeh, Arian Askari, Mohammad Aliannejadi, and Maarten de Rijke. 2024. Ranked List Truncation for Large Language Model-based Re-Ranking. ArXiv abs/2404.18185 (2024). https://api.semanticscholar.org/CorpusID:269449617

[41] Louie Meyer, Johanne Engel Aaen, Anitamalina Regitse Tranberg, Peter Kun, Matthias Freiberger, Sebastian Risi, and Anders Sundnes Løvlie. 2024. Algorithmic ways of seeing: Using object detection to facilitate art exploration. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. 1–18.

[42] Sheshera Mysore, Arman Cohan, and Tom Hope. 2022. Multi-Vector Models with Textual Guidance for Fine-Grained Scientific Document Similarity. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 4453–4470.

[43] Sheshera Mysore, Tim O'Gorman, Andrew McCallum, and Hamed Zamani. [n. d.]. CSFCube–A Test Collection of Computer Science Research Articles for Faceted Query by Example. ([n. d.]).

[44] Harshit Nigam, Manasi Patwardhan, Lovekesh Vig, and Gautam Shroff. 2024. Acceleron: A Tool to Accelerate Research Ideation. arXiv preprint arXiv:2403.04382 (2024).

[45] Baharan Nouriinanloo and Maxime Lamothe. 2024. Re-Ranking Step by Step: Investigating Pre-Filtering for Re-Ranking with Large Language Models. ArXiv abs/2406.18740 (2024). https://api.semanticscholar.org/CorpusID:270764517

[46] Jeongseok Oh, Seungju Kim, and Seungjun Kim. 2024. LumiMood: A Creativity Support Tool for Designing the Mood of a 3D Scene. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. 1–21.
[47] Jason Portenoy, Marissa Radensky, Jevin D West, Eric Horvitz, Daniel S Weld, and Tom Hope. 2022. Bursting scientific filter bubbles: Boosting innovation via novel author discovery. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. 1–13.
[48] Kevin Pu, KJ Feng, Tovi Grossman, Tom Hope, Bhavana Dalvi Mishra, Matt Latzke, Jonathan Bragg, Joseph Chee Chang, and Pao Siangliulue. 2024. IdeaSynth: Iterative Research Idea Development Through Evolving and Composing Idea Facets with Literature-Grounded Feedback. arXiv preprint arXiv:2410.04025 (2024).
[49] A Terry Purcell and John S Gero. 1996. Design and other types of fixation. Design studies 17, 4 (1996), 363–383.
[50] Abhilasha Ravichander, Shrusti Ghela, David Wadden, and Yejin Choi. 2025. HALoGEN: Fantastic LLM Hallucinations and Where to Find Them. arXiv preprint arXiv:2501.08292 (2025).
[51] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Conference on Empirical Methods in Natural Language Processing. https://api.semanticscholar.org/CorpusID:201646309
[52] Mark A Runco et al. 2010. Divergent thinking, creativity, and ideation. The Cambridge handbook of creativity 413 (2010), 446.
[53] 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 (2024).
[54] Dean Keith Simonton. 2021. Scientific Creativity: Discovery and Invention as Combinatorial. Frontiers in Psychology 12 (2021). https://api.semanticscholar.org/CorpusID:237262181
[55] Arvind Srinivasan and Joel Chan. 2024. Improving Selection of Analogical Inspirations through Chunking and Recombination. In Proceedings of the 16th Conference on Creativity & Cognition. 374–397.
[56] Sangho Suh, Meng Chen, Bryan Min, Toby Jia-Jun Li, and Haijun Xia. 2024. Luminate: Structured Generation and Exploration of Design Space with Large Language Models for Human-AI Co-Creation. In Proceedings of the CHI Conference on Human Factors in Computing Systems. 1–26.
[57] Lu Sun, Aaron Chan, Yun Seo Chang, and Steven P Dow. 2024. ReviewFlow: Intelligent Scaffolding to Support Academic Peer Reviewing. In Proceedings of the 29th International Conference on Intelligent User Interfaces. 120–137.
[58] Weiwei Sun, Lingyong Yan, Xinyu Ma, Pengjie Ren, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agent. ArXiv abs/2304.09542 (2023). https://api.semanticscholar.org/CorpusID:258212638
[59] Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agents. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 14918–14937.
[60] P Thagard. 2012. The cognitive science of science: Explanation, discovery, and conceptual change. The MIT Press.
[61] Jianyou Wang, Kaicheng Wang, Xiaoyue Wang, Prudhviraj Naidu, Leon Bergen, and Ramamohan Paturi. 2023. DORIS-MAE: scientific document retrieval using multi-level aspect-based queries. In Proceedings of the 37th International Conference on Neural Information Processing Systems. 38404–38419.
[62] Meiyun Wang, Kiyoshi Izumi, and Hiroki Sakaji. 2024. LLMFactor: Extracting profitable factors through prompts for explainable stock movement prediction. arXiv preprint arXiv:2406.10811 (2024).
[63] Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2023. Scimon: Scientific inspiration machines optimized for novelty. arXiv preprint arXiv:2305.14259 (2023).
[64] Hongji Yang, Delin Jing, and Lu Zhang. 2016. Creative Computing: an approach to knowledge combination for creativity?. In 2016 IEEE Symposium on Service-Oriented System Engineering (SOSE). IEEE, 407–414.
[65] Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou, Lifeng Jin, Linfeng Song, Haitao Mi, and Helen Meng. 2024. Self-alignment for factuality: Mitigating hallucinations in llms via self-evaluation. arXiv preprint arXiv:2402.09267 (2024).

Appendix Overview

This appendix provides supplementary material organized into the following sections:

A User Study: Design and Procedure

A.1 Participants

We recruited 22 computer-science researchers (W: 7, M: 15) through institutional mailing lists and academic social networks. We compensated them with a $60 Amazon gift card. Twelve participated as human-computer interaction (HCI) researchers and 10 as natural-language-processing (NLP) researchers. Most were PhD students (PhD student: 16, master's student: 5, industry researcher: 1). Generally, the participants interacted with LLMs often (a few times per... day: 12, week: 7, month: 1, few months or longer: 2).

A.2 Study Design

We conducted a within-subjects study, in which each participant completed tasks for the treatment and baseline conditions in randomized order. The ideation topics for the treatment and baseline conditions were also randomized. Overall, participants had no difference in their familiarity ratings (7-point, Likert-type) for the assigned treatment topic and assigned baseline topic (M=0.00, Q1=-1.00, Q3=1.00). There were two preset topics for HCI researchers (human-AI collaboration in art, AI tools for education) and two for NLP researchers (dealing with LLM hallucinations, LLM explainability). For each topic, there were three associated input papers to use as a starting point. The input papers for each topic are listed in the following table.

Topic	Input Papers
Human-AI Collaboration in Art	LumiMood [46]; Prompting for Discovery [3]; Algorithmic Ways of Seeing [41]
AI Tools for Education	VIVID [13]; Scientific and Fantastical [10]; Putting Things into Context [34]
Dealing with LLM Hallucinations	Deductive Closure Training [2]; Self-Alignment for Factuality [65]; HALoGEN [50]
LLM Explainability	LLMFactor [62]; TextGenSHAP [19]; Digital Socrates [23]

A.3 Baseline Tool

The baseline tool (Figure 10) represents *paper-level interaction*: participants could select any combination of the three input papers as input to the LLM gpt-4o-2024-08-06, the same LLM used for most of Scideator's functionality. If they did not select any papers, all three were provided to the LLM. Participants could also provide custom instructions to the LLM, with a character limit of 75,000 (compared to 25,000 in the treatment) to account for the treatment tool's longer

set idea-generation prompt. The baseline's idea-generation prompt was a simplified version of the one in Scideator: it did not utilize any facet-based framework or carefully crafted criteria for a good idea. However, like Scideator, it generated six candidate ideas for every two presented to the participant and followed instructions to improve upon the idea. The baseline's "Idea Novelty Evaluation" tab was similar to that in the treatment tool except there was no Idea Novelty Checker module output (i.e., no related papers, novelty classification, or classification reason for each idea).

A.4 Treatment Tool Modifications for Study

We modified Scideator to more effectively address our research questions. Our study separates the idea generation task from the idea evaluation task. To keep the study controlled, we disabled some of Scideator's functionalities: on-demand novelty evaluation, manual idea addition, and facet generation when there is no query. The Idea Novelty Checker module was only activated in a separate "Idea Novelty Evaluation" tab for the idea evaluation step. There was no support for adjusting the novelty assessment or iterating on the idea's novelty. The tab also provided access to a ChatGPT-like interaction in which participants could prompt the LLM directly to help evaluate their ideas for novelty, as well as a text field for keeping notes on their novelty assessments.

A.5 Procedure

Each within-subjects study session was 105 minutes. The sessions were recorded and transcribed using Google Meet. In each condition, the session coordinator provided the participant with the assigned tool, a document with the titles and abstracts of the input papers for the assigned ideation topic, and a link to Semantic Scholar.[6] They had access to these three resources throughout the condition. The participant completed two tasks with each tool: an idea-generation task followed by an idea-novelty-evaluation task.

Idea-generation task. The participant entered their assigned ideation topic and three input papers into the tool. While the tool loaded, the coordinator went over the task instructions and gave the participant a tutorial describing the tool's features. The participant then had up to two minutes to review the three input papers' titles and abstracts. With access to the tool, Semantic Scholar, and the input paper document, the participant subsequently spent 20 minutes generating and saving as many research ideas as possible. To save an idea, the participant had to confirm that the idea was at least somewhat relevant to the ideation topic and somewhat interesting to think about further. They also provided a seven-point Likert-type rating of how different the idea was from ideas they had or encountered before the study; they were told to aim for saving ideas that were at least somewhat significantly different. The coordinator alerted the participant when five minutes remained.

Idea rating and survey. Once 20 minutes had passed, the participant opened a "Saved Ideas" tab to select their two favorite ideas and answer additional 7-point Likert-type questions about their perceived novelty, feasibility, specificity, impact, and imaginativeness of each idea. There were two instances in which a participant had only saved one idea in the 20 minutes allotted. In this case, we asked them to select their next favorite idea in order to proceed with two favorite ideas. Participants also rated their confidence in their novelty assessment. They then completed a survey that included seven-point Likert-type questions about their familiarity with the assigned topic and whether they encountered concepts they had not previously heard about or encountered in the context of the ideation topic. The survey also included the Creativity Support Index (CSI) questionnaire [11]. In the survey for the second tool, the participant also answered questions for each pair of CSI factors to determine which factors they considered most important, as is standard for the CSI. The coordinator then spent up to around five minutes engaging the participant in a semi-structured interview about their idea generation experience.

Idea-novelty-evaluation task. The participant opened the "Idea Novelty Evaluation" tab, and the coordinator provided an overview of this portion of the tool. The participant spent five minutes evaluating their two favorite ideas for novelty. For each idea, they provided a final seven-point Likert-type rating of perceived novelty and confidence in their novelty assessment. The coordinator then conducted a brief semi-structured interview about the participant's idea evaluation experience.

B User Study: Extended Results

B.1 Additional CSI Factor Results

The main paper reports the exploration and expressiveness factors in detail. Here we report the remaining CSI factors.

Immersion. Overall, participants did not find Scideator more helpful than the baseline tool for becoming immersed in the idea-generation task. The interviews and interaction logs provide some reasons why this may be true. Scideator presents the user with several features about which to learn, and the cognitive demand of learning these features may have prevented immersion. Four participants commented on the high cognitive load of using Scideator; P2-HCI-treatment commented,

> "*I would say that it took me more mental effort to figure out how the tool [is used] rather than work with ideas.*"

Furthermore, due to more complex prompting, the average latency for generating two ideas in Scideator was 22.04 seconds, compared to 15.21 seconds in the baseline tool. Ideas were generated in groups of four, two at a time, so the first two ideas took less time to generate than the last two ideas.

Results-Worth-Effort. Participants generally found their results to be more worth the effort while using Scideator compared to the baseline tool. However, there was little difference in participants' average ratings of their favorite ideas in terms of perceived feasibility (M=0.00, Q1=-0.50, Q3=0.50), specificity (M=0.00, Q1=-0.88, Q3=0.50), and imaginativeness (M=0.00, Q1=-0.50, Q3=0.38). Meanwhile, the baseline tool performed slightly better with respect to generating impactful ideas (M=-0.25, Q1=-0.50, Q3=0.00). Perhaps participants could more clearly see the potential impact of ideas grounded in a few set input papers, which they reviewed, versus ideas drawing from several papers, most of which were not reviewed by them. P10-HCI-baseline posited,

[6] https://www.semanticscholar.org/

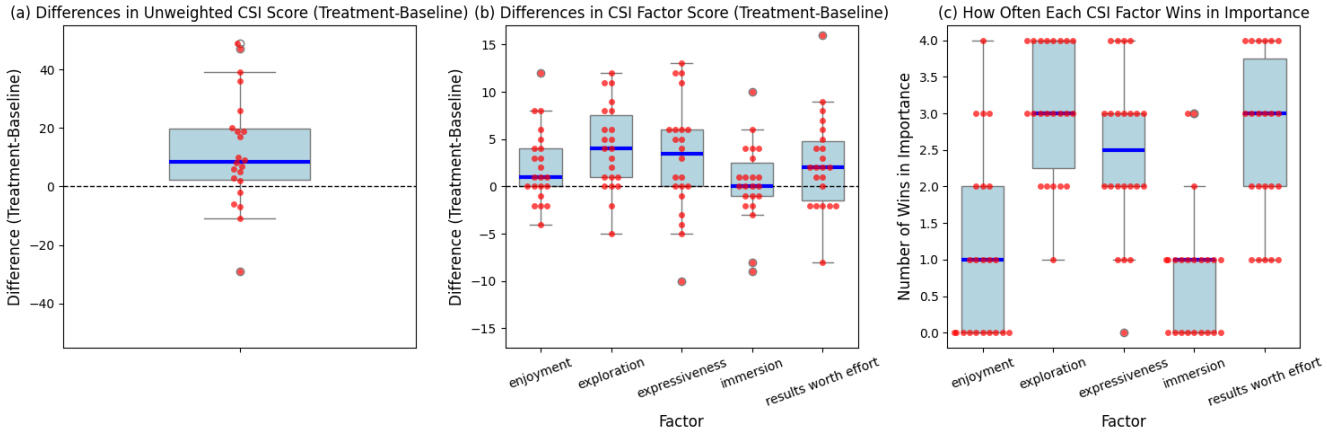

Figure 5: (a) The difference between participants' unweighted CSI scores for SCIDEATOR versus the baseline tool. Participants experienced significantly more creativity support with SCIDEATOR. (b) For each CSI factor, the difference between participants' ratings for SCIDEATOR versus the baseline tool. (c) How many times each CSI factor wins against other factors in terms of what is most important to participants while generating ideas.

> "*since now I know the paper, I read them, I kind of understand the vocabulary... it is easier for me to see where these ideas are coming from. So even when the ideas are written somewhat vaguely, I can still... imagine how that would pan out because I read the paper.*"

Enjoyment. Most participants benefited slightly from SCIDEATOR in terms of enjoyment, but the difference from the baseline was small (Figure 5b).

B.2 Evaluation Facet Utility

Participants tended not to find the evaluation facets as helpful as the purpose and mechanism facets. Four participants commented that they found the evaluation facet unimportant. P13-HCI-treatment elaborated,

> "*For evaluation, I really don't think it's necessary for me because once you have the problem, you have the solution. Automatically you know how to evaluate it, like what study you need, what kind of experiment you want to have, and what variables you are measuring.*"

Future work may investigate whether the evaluation facet is useful for mixed-initiative, facet-based generation of research ideas.

B.3 Paper-Level Interaction Preferences

Five participants appreciated the baseline's paper-level interactions in addition to or more than SCIDEATOR's facet-level interactions. Three of these participants liked the paper-level interaction, as it felt more directly connected to the literature. P22-NLP-baseline explained,

> "*I think papers for me were more natural than facets.... I think to me it's more like a map of literature, so I could see it more with papers.*"

Three participants thought a combination of the two tools would be helpful. Two participants even proposed distinct roles for the two tools: divergent-ideation for SCIDEATOR and convergent-ideation for the baseline. P7-HCI-treatment shared,

> "*In the [baseline tool], I started from a broader view and then I narrowed it down. Here [in SCIDEATOR], I started from a very specific thing and then I tried to add new facets or ideas so that I can expand the idea. So you see the other process is elimination process, here I was trying to expand.*"

Distant Facet Avoidance. Relatedly, participants sometimes avoided distant facets in SCIDEATOR because they were unfamiliar with how to apply them — they did not know the facet's meaning, found it "too far" from their research area, or could not see how the mechanism could achieve the purpose. Participants saved ideas with far facets much less often than ideas with input and near facets. By avoiding distant facets, scientists may miss opportunities for ideas that would never have occurred to them otherwise.

B.4 Desire for More Papers in Baseline

After using the baseline tool, four participants said that they wanted a way to input more papers to better express themselves, and five more felt limited by the three input papers. P14-NLP-baseline, for example,

> "*would have liked to add a different paper because it felt like I had exhausted... the creativity in the system to some extent.*"

While participants could add information from papers to their custom instructions, there was no system feature for adding more

papers to the list of input papers. Future work may compare SCIDEATOR with a modified version of the baseline tool that allows users to add as many papers as they want for recombination.

B.5 Novelty Checker Usage Patterns

Participants used the novelty checker an average of 6 times per session. The checker was most convincing when it classified an idea as "not novel" — participants could directly verify the judgment against the retrieved papers. "Novel" classifications were less persuasive, as the absence of overlapping work is harder to trust than its presence. Two common concerns were that the checker classified ideas as "novel" too often and that the set of most related papers sometimes missed important works. Both concerns may have been partly due to the system retrieving only 10 papers per idea in the user study, a constraint imposed by latency limitations. Broader retrieval and richer paper representations may improve evaluation quality and user trust.

C Annotation Study for Modules

C.1 Facet Distance Consistency

To assess whether the system's distance labels (near, far, very far) correspond to meaningful differences in conceptual distance, the first two authors independently annotated purpose-mechanism pairs generated by the system for three papers not used in the user study.[7] Facets were grouped into two broad categories — generally near and generally far — and each annotator classified whether the system's labels matched their own judgment. Both annotators classified the majority of near purposes, near mechanisms, far purposes, and far mechanisms consistently with the system's labels. These results suggest the distance levels capture meaningful variation in conceptual similarity, although we did observe certain degree of subjectivity of judging how far is an analogy.

C.2 Novelty Checker: Expert-Labeled Examples

The Idea Novelty Checker uses expert-labeled examples as in-context demonstrations for its novelty assessment step (Section 2). Each example includes the idea (with key facets bolded), the top-10 most relevant papers from the re-ranking pipeline, and the expert's grounded reasoning. These demonstrations teach the LLM to cite specific retrieved papers when justifying its classification. Figure 6 shows a novel example and Figure 7 shows a not-novel example.

D Sample Generated Ideas

Table 3 presents six sample ideas that participants saved as favorites during the user study—four from SCIDEATOR and two from the baseline. For each treatment idea, the table shows which generation method was triggered by the participant's facet selections (Section 2), the selected facets with their conceptual distance from the input papers, and any custom instructions the participant provided. Baseline ideas show only custom instructions, since the baseline uses paper-level interaction rather than facets. Two patterns are visible: (1) treatment ideas draw on facets from papers at varying distances, with participants often selecting the purpose themselves while letting the mechanism come from a retrieved paper; and (2) baseline ideas rely more heavily on custom instructions to steer generation.

[7] This annotation used an earlier but similar version of the tool compared to the version used in the study.

Example 1 — Classification: Novel

Idea: Develop a **natural language processing classifier designed to improve scientific paper revisions** by automatically identifying and categorizing reviewer comments that are most likely to lead to substantial and actionable revisions. The system would be trained on a **manually-labeled dataset analysis** of scientific review comments and the corresponding paper edits, leveraging features such as linguistic cues, sentiment, and comment specificity to predict the likelihood of a comment being acted upon. This classifier could then be used to prioritize reviewer feedback, helping authors focus on the most impactful suggestions first.

Top-10 Retrieved Papers:
(1) ARIES: A Corpus of Scientific Paper Edits Made in Response to Peer Reviews
(2) Can Large Language Models Provide Useful Feedback on Research Papers?
(3) A Dataset of Peer Reviews (PeerRead): Collection, Insights and NLP Applications
(4) arXivEdits: Understanding the Human Revision Process in Scientific Writing
(5) Characterizing Text Revisions to Better Support Collaborative Writing
(6) Can We Automate Scientific Reviewing?
(7) DeepReviewer: Collaborative Grammar & Innovation Neural Network for Automatic Paper Review
(8) Aspect-based Sentiment Analysis of Scientific Reviews
(9) Aspect-based Sentiment Analysis of Online Peer Reviews and Prediction of Paper Acceptance
(10) ReviVal: Towards Automatically Evaluating the Informativeness of Peer Reviews

Expert Reasoning: The idea is **novel** because it uniquely focuses on *prioritizing* reviewer comments for actionable revisions, which is not explicitly addressed in ARIES [1] or other related works like ReviVal [10].

Figure 6: Expert-labeled novel example used as an in-context demonstration. The retrieved papers address related aspects of peer review (corpora, automation, sentiment) but none specifically tackle the proposed task of *prioritizing* comments by likelihood of leading to revisions — the gap the expert identifies.

Example 2 — Classification: Not Novel

Idea: Develop a **systematic review-based framework** designed **to align LLM evaluation with human preferences**, ensuring that evaluation criteria are continuously refined based on comprehensive reviews of user feedback and emerging model behaviors. This framework will utilize **content analysis of user interactions and feedback** to identify patterns and areas of improvement. The effectiveness of this framework will be assessed through a **qualitative study** involving iterative cycles of user feedback and criteria refinement.

Top-10 Retrieved Papers:
(1) EvalLM: Interactive Evaluation of Large Language Model Prompts on User-Defined Criteria
(2) Humanely: Human Evaluation of LLM Yield, Using a Novel Web-Based Evaluation Tool
(3) Evaluation of Code Generation for Simulating Participant Behavior in ESM by Iterative ICL of an LLM
(4) Human-Centered Evaluation and Auditing of Language Models
(5) Aligning Model Evaluations with Human Preferences: Mitigating Token Count Bias in LM Assessments
(6) Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences
(7) Human-Centered Design Recommendations for LLM-as-a-Judge
(8) CheckEval: Robust Evaluation Framework using Large Language Model via Checklist
(9) Discovering Language Model Behaviors with Model-Written Evaluations
(10) Prometheus 2: An Open Source Language Model Specialized in Evaluating Other Language Models

Expert Reasoning: The idea is **not novel** because it closely resembles existing frameworks like EvalLM [1] and HumanELY [2], which already align LLM evaluations with human preferences using user-defined criteria and human feedback.

Figure 7: Expert-labeled not-novel example. Unlike Example 1, the top retrieved papers directly overlap with the proposed idea's core contribution — aligning LLM evaluation with human preferences through iterative feedback — making the overlap straightforward to identify.

Table 3: Sample favorite ideas from the user study. Treatment ideas show the generation method, selected facets (with distance and selection source), and custom instructions. Baseline ideas show only the custom instructions provided. Facets are color-coded: Purpose, Mechanism, and Evaluation. "Selected" means the participant chose this facet; "unselected" means the system chose it automatically.

Idea 1	*Treatment — Initial method*	Topic: LLM Explainability
Facets	to enhance medical information retrieval (near)	sequential knowledge-guided prompting (input)
	medical QA benchmarks (near, unselected)	
Custom	*None*	
Idea	The proposed system focuses on developing a robust medical information retrieval platform using sequential knowledge-guided prompting to simulate diagnostic pathways. By integrating electronic health records (EHRs) with established clinical guidelines, the system leverages prompts to dynamically identify and follow potential diagnostic routes, thereby enhancing the precision and relevance of information retrieval. The process begins by extracting key medical factors from EHRs and matching these with clinical pathway prompts to simulate diagnostic reasoning. This approach allows the system to adjust retrieval strategies based on real-time patient data, ensuring contextually relevant information is provided to healthcare professionals. The system's effectiveness will be rigorously evaluated using medical question-answering benchmarks, which will assess the accuracy and pertinence of information retrieval against standard medical queries.	
Idea 2	*Treatment — P-or-M method*	Topic: Human-AI Collaboration in Art
Facets	to enhance designer-AI collaboration (very near, selected)	residual learning framework (far)
	empirical performance analysis (far, unselected)	
Custom	*I want to know if designers want to work with AI*	
Idea	In this novel approach, we apply a residual learning framework to AI design tools, treating designer feedback as residuals to refine specific design elements such as color schemes and layout structures. This framework allows designers to provide targeted feedback easily, which the AI tool uses to iteratively improve its suggestions. The process starts with the AI generating an initial design, which the designer can annotate with feedback. This feedback is then incorporated as residuals, prompting the AI to make adjustments that align more closely with the designer's vision. The feedback loop is engineered to be highly intuitive, integrating seamlessly into the designer's workflow, thus fostering a more collaborative and efficient design process. An empirical performance analysis will evaluate the effectiveness of this approach, measuring improvements in design quality and collaboration efficiency.	
Idea 3	*Treatment — P-and-M method*	Topic: LLM Explainability
Facets	to explain text generation outputs (input)	explanation logic verifier (own input)
	quantitative and qualitative analysis (input, selected)	
Custom	*Use different verification approaches from NLI, math, program to ensure the explanation follows a logical consistency*	
Idea	Create an innovative platform that enhances the validation of LLM-generated mathematical explanations by integrating a mathematical logic verifier with visual proof representations. The platform will employ the verifier to ensure explanations adhere to logical principles, assessing each step for accuracy. Simultaneously, it will generate graph-based visual proofs, providing intuitive visual cues for each stage of the reasoning process. These representations will guide users through complex explanations, making abstract concepts more accessible and engaging. The system will also feature interactive elements that allow users to explore different parts of the proofs, deepening their understanding. This approach will be evaluated through comprehensive quantitative and qualitative analysis, assessing improvements in explanation accuracy, user comprehension, and satisfaction.	
Idea 4	*Treatment — P-and-M method*	Topic: AI Tools for Education
Facets	to personalize vocabulary learning (very near)	AI-generated virtual speakers (very near)
	participant feedback analysis (very near, unselected)	
Custom	*Make the idea more focused and specific*	
Idea	This innovative platform leverages AI-generated virtual speakers to create personalized vocabulary learning experiences. Users upload personal images, and the AI analyzes these images to identify objects, settings, and potential vocabulary links. The virtual speakers then generate dialogue scenarios around these elements, allowing learners to engage with new vocabulary in a personally relevant context. For example, an image of a beach vacation might lead to dialogues about seaside activities, weather, and local culture, enriching the vocabulary acquisition process. This personalized approach not only aids retention by linking words to familiar contexts but also enhances engagement through interactive AI-driven dialogues. Participant feedback analysis will assess the platform's effectiveness in improving vocabulary acquisition and learner satisfaction.	

Table 3 (continued). Baseline ideas.

Idea 5	*Baseline*	Topic: Dealing with LLM Hallucinations
Facets	*N/A — baseline uses paper-level interaction*	
Custom	*Can the reward model in Self-Knowledge Tuning be changed to incorporate more fine-grained feedback (e.g., from HALoGEN-like metrics)? [all 3 papers given to LLM]*	
Idea	This research focuses on developing an advanced reward model for Self-Knowledge Tuning by incorporating specific metrics from the HALoGEN framework. The core of this model lies in leveraging atomic unit verification to dissect LLM outputs into fundamental components, evaluating each against reliable knowledge bases. By integrating HALoGEN's error classifications, the model can provide targeted feedback that distinguishes between different types of hallucinations, such as those arising from incorrect recollections, faulty training data, or fabrications. This nuanced feedback is expected to enhance the calibration of LLMs, reducing the frequency of factual inaccuracies. The model's efficacy will be tested across diverse domains, including programming and scientific attribution, to ensure broad applicability. A comprehensive validation framework will be established, involving both automated and human evaluations, to rigorously measure improvements in model accuracy and coherence.	
Idea 6	*Baseline*	Topic: AI Tools for Education
Facets	*N/A — baseline uses paper-level interaction*	
Custom	*Generate ideas that use non-traditional mediums for education, so avoid using textbooks or other material commonly found in classrooms. [all 3 papers given to LLM]*	
Idea	This AR-enabled mobile application is designed to engage children with their local environment while providing a robust educational platform about biodiversity and ecology. The AI component will analyze user interactions and location data to curate a personalized educational journey. For a child in an urban environment, the app might focus on urban wildlife and plants, providing detailed information, conservation tips, and interactive quizzes tailored to their locality. For children in rural settings, it might explore more diverse ecosystems, encouraging them to document and learn about various species.	

E Implementation Details

The frontend of SCIDEATOR was developed using React and TypeScript, and the backend with Python. Unless noted otherwise, we used the LLM gpt-4o-2024-08-06 and a temperature of zero. In addition, whenever we mention using a paper, we only use its title and abstract. Prompts for all modules are provided in the Appendix. In facet-idea generator the LLM's temperature is set to 0.75 to make the responses more varied. We use gpt-4o for re-ranking in novelty checker and o3-mini for the novelty assessment step given its stronger reasoning capabilities.

F System Figures

Table 5: Guide to system figures (Appendix F).

Figure	Overview
User-Interace	
Fig. 8	SCIDEATOR novelty assessment UI component: idea, facets, related papers, adjustable novelty rating and rationale, and revision suggestions.
Fig. 9	SCIDEATOR facet-driven ideation interface: select/generate facets, add instructions, browse ideas, and access novelty evaluation.
Fig. 10	Baseline ideation UI: select papers, add optional instructions, and browse generated ideas (no facet controls or structured novelty feedback).
SCIDEATOR modules	
Fig. 11	Module 1: Analogous Paper Facet Finder
Fig. 12	Module 3: Idea Novelty Checker

G Pseudocode of SCIDEATOR Modules

Table 6 summarizes the pseudocodes used across SCIDEATOR's three modules. We will release our code upon acceptance of the paper for reproducibility.

Table 6: Summary of LLM prompts (pseudocode) by module. Section references point to the full prompt text below.

Module	Prompt Purpose	Section
Analogous Paper Facet Finder	Extract facets from paper title/abstract	G.1.1
	Generate analogy queries at 3 distances	G.1.2
	Shorten overly specific queries	G.1.3
	Summarize input + very-near papers	G.1.4
Faceted Idea Generator	Shared Chain-of-Thought Structure across three idea-gen scenarios	G.2.1
	No facets selected by user	G.2.2
	One facet selected (purpose or mechanism)	G.2.3
	Both facets selected	G.2.4
Idea Novelty Checker	Extract keywords/titles from idea	G.3.1
	Extract idea facets for re-ranking	G.3.2
	Re-rank papers by facet relevance	G.3.3
	Assess idea novelty	G.3.4
	Generate more-novel idea suggestions	G.3.5

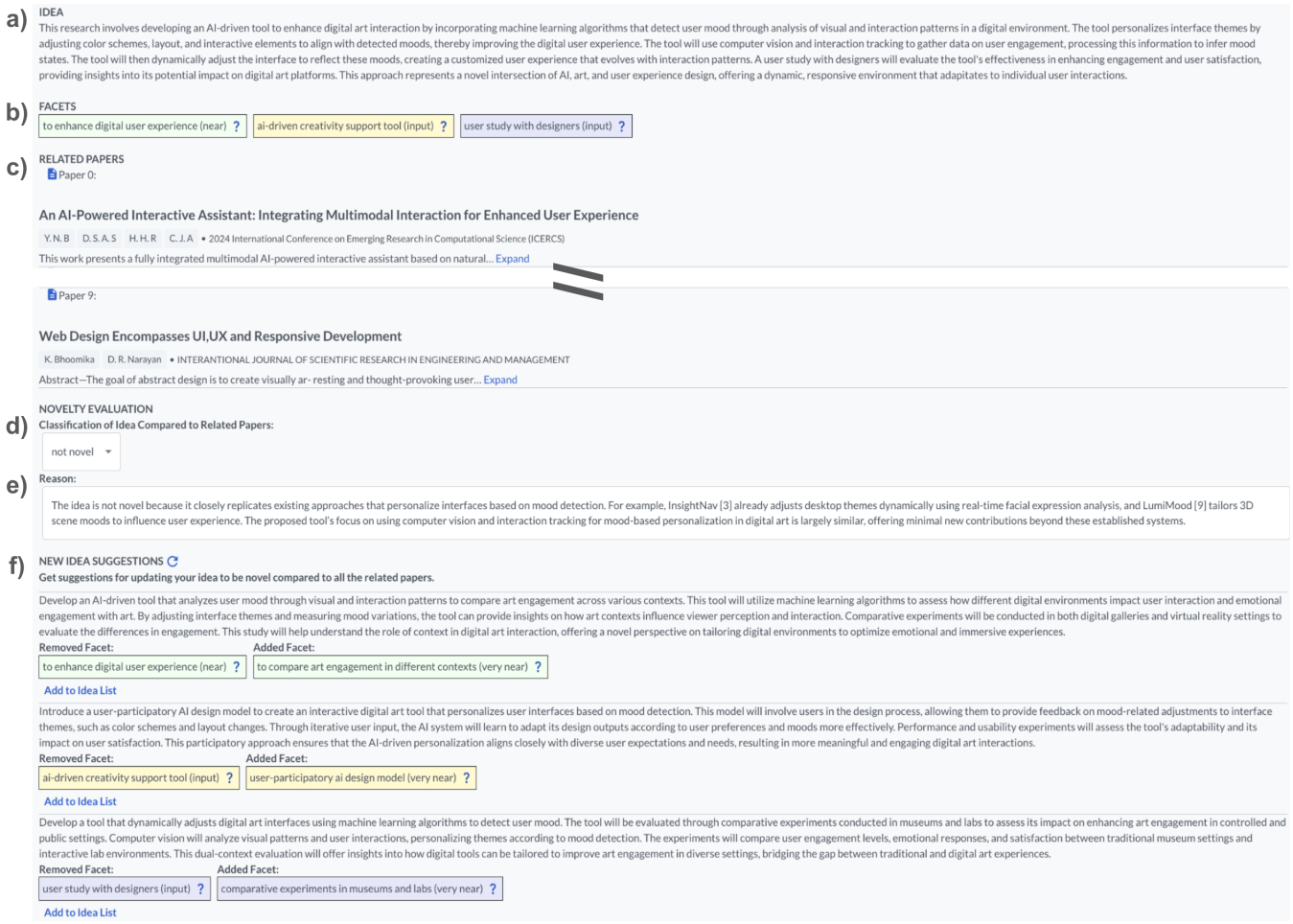

Figure 8: SCIDEATOR's novelty assessment modal for one idea, which presents the idea (a) as well as its facets (b), related papers (c), adjustable novelty classification (d), and adjustable classification reason (e). When the idea is classified as "not novel," the system provides a set of three suggestions for more novel ideas (f), each of which replace one of the idea's original facets. The ideation topic here is human-AI collaboration in art.

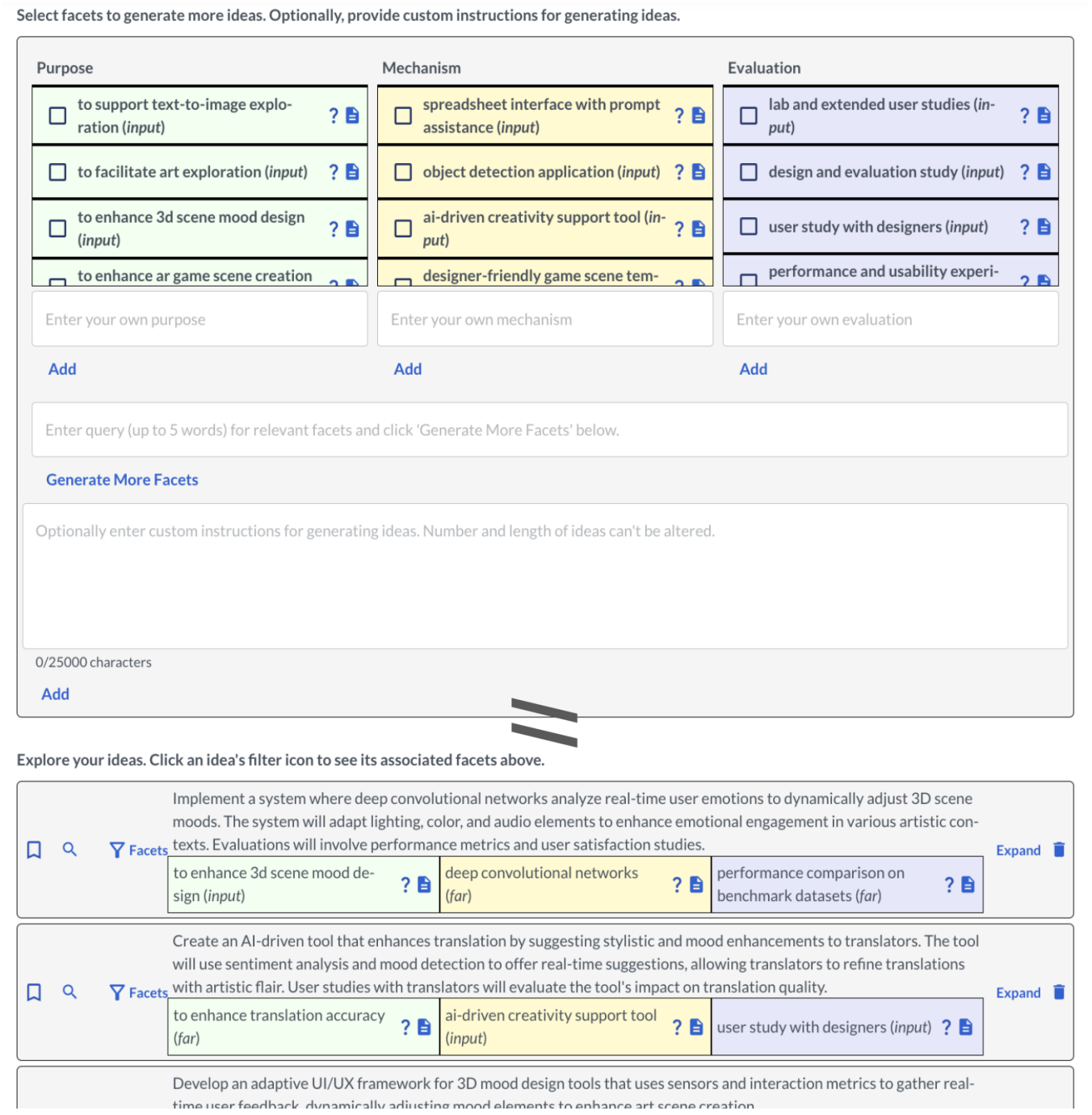

Figure 9: Scideator's cold start. Above, the user selects or adds facets to generate ideas. They can also generate more facets to consider, and add custom instructions for the idea generation. Below, the user peruses their ideas and evaluates an idea for novelty by clicking the search icon to its left. The ideation topic here is human-AI collaboration in art.

Figure 10: The cold start of the baseline UI for the user study's idea-generation task. The ideation topic here is human-AI collaboration in art.

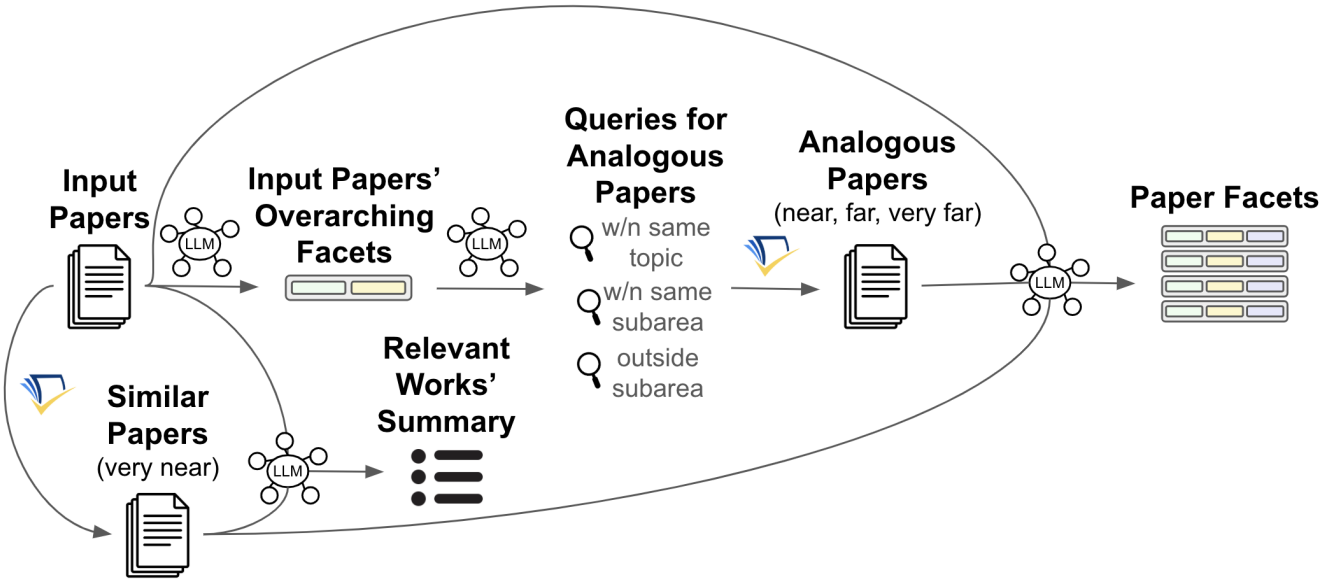

Figure 11: The Analogous Facet Generation and Paper Retrieval module.For a set of input papers, SCIDEATOR uses Semantic Scholar's API to retrieve similar papers (very near). It uses the input and very-near papers to create a summary of relevant works. Next, the tool extracts key facets from the input papers and determines the input papers' overarching purpose and mechanism, which it uses to come up with three queries for papers with an analogous purpose and mechanism. The queries are for analogous papers with varying distances from the input paper: same topic (near), same subarea (far), and different subarea (very far). Those queries are fed to the Semantic Scholar API to retrieve analogous papers. Finally, the facets of all the analogous papers are extracted by the LLM.

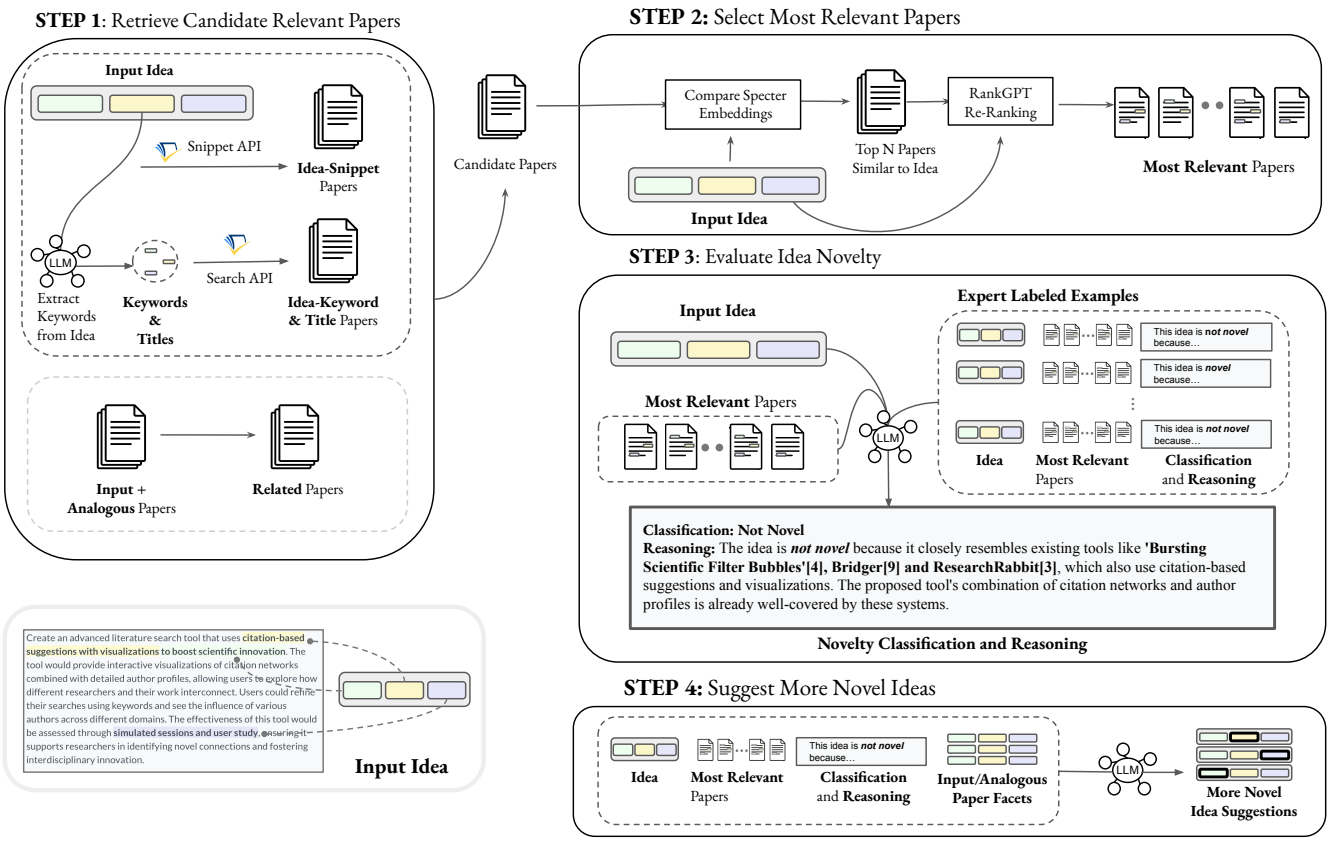

Figure 12: The Idea Novelty Checker module follows a retrieve-then-re-rank approach for novelty evaluation. In Step 1, it gathers a comprehensive set of papers relevant to an idea. This includes papers originally used to generate the idea, related papers, and additional papers retrieved through keyword and title searches extracted directly from the idea, as well as snippet searches using the entire idea as input. In Step 2, a two-stage re-ranking process is applied, where an embedding-based ranking strategy filters the large collection to top-N papers, followed by a facet-based LLM re-ranker to identify the top-k most relevant papers. In Step 3, these top-k papers are used to assess the idea's novelty, guided by in-context examples that evaluate novelty with grounded reasoning. In Step 4, if an idea is classified as "not novel" by the tool or user, the LLM generates three idea suggestions, each replacing a different facet in the original idea in order to make the idea more novel compared to the relevant papers.

G.1 Facet Finder Pseudocode

G.1.1 *Extract Facets from Paper.*

`extractFacets(papers)`

Role: ScientistGPT | **Input:** Title, abstract, and (if available) introduction of each paper **Task:** Extract three facets per paper—purpose, mechanism, evaluation—plus a 1–2 sentence definition for each. **Constraints on facets:** Single short phrase (≤7 words); specific enough to inspire ideas but not tied to one paper; no numbers unless part of a name; no acronyms; if a paper has multiple facets of one type, combine into one; evaluation must not reference the purpose. **Constraints on definitions:** Up to 2 sentences; replace proper nouns and jargon with definitions; self-contained; do not reuse words from the facet itself. **Contrastive examples** (3 pairs per facet type, abbreviated):

	Bad	*Good*
P	"to generate creative writing activities for third-grade English lessons" (too specific)	"to support elementary creative writing"
M	"LLM chain-of-thought from gpt-3.5-turbo trained up to 11/06 with temperature=0.7" (too specific, numbers, acronym)	"LLM chain-of-thought reasoning"
E	"between-subjects 4×4 user study with 32 teachers" (too specific, references purpose)	"Wizard of Oz user study"

Output: Per paper: Purpose / Purpose Definition / Mechanism / Mechanism Definition / Evaluation / Evaluation Definition

G.1.2 *Generate Analogy Queries at Three Distances.*

`analogyQueries(purpose, mechanism, topic)`

Input: Purpose and mechanism of designated paper; any previously generated queries

Task: For each of three conceptual distances, generate an analogous purpose/mechanism pair and a ≤5-word search query for paper retrieval: (1) **Same topic** of CS research; (2) **Same subarea**, different topic; (3) **Different subarea** entirely.

Key constraint: The structural relationship between purpose and mechanism must be preserved across the analogy ("P is to M as P' is to M' because both involve...").

Deduplication: If previous queries exist, new analogies must not overlap with them.

Output: Per distance: analogy statement, purpose, mechanism, search query.

G.1.3 *Shorten Overly Specific Queries.*

`shortenQuery(query)`

Trigger: Called when a generated query retrieves <4 relevant papers from Semantic Scholar.

Task: Produce a simpler, shorter version of the query, prioritizing the most important information if meaning must be lost.

G.1.4 *Summarize Input and Near Papers.*

`summarizePapers(input_papers, near_papers)`

Input: Titles and abstracts of the user's input papers plus very-near analogous papers.

Task: Produce a concise summary of prior work. This summary is injected into all idea-generation prompts as grounding context, ensuring the LLM knows what has already been done.

G.2 Idea Generator Pseudocode

All three idea-generation components share a common multi-step chain-of-thought structure and the same quality requirements. We describe these shared components first, then highlight how each variant differs.

G.2.1 Shared Chain-of-Thought Structure.

Every idea-generation prompt follows five steps:

Step 1 – Grounding: Read the summary of prior work and paper details to understand what exists.
Step 2 – Deduplication: Read any previously generated ideas to avoid repetition.
Step 3 – Brainstorm: Generate N analogies between designated × analogous papers, each with a short idea sketch (30–50 words).
Step 4 – Select: Choose the K best analogies that meet the idea requirements below.
Step 5 – Elaborate with self-critique: For each selected analogy, produce: (a) an "imaginative twist" statement; (b) a relevance justification to the user's topic; (c) an initial idea (100–150 words); (d) self-identified issues with the initial idea; (e) a plan to address those issues; (f) a revised idea (100–150 words); (g) an expanded version (200–250 words).

Shared idea quality requirements. Each generated idea must satisfy five categories, each with 3–5 sub-requirements:

(1) **Understandability:** Logical, grammatical, self-contained (no references to specific tool names a reader would not know).
(2) **Relevance:** Adapted to the user's ideation topic; must not reference the analogy mechanism directly.
(3) **Specificity:** 100–150 words; 90% focused on how the mechanism addresses the purpose; concrete implementation direction. Includes negative examples (e.g., "an idea saying to 'apply a faceted representation to clinical data, creating a multidimensional patient profile' is not novel because prior work has already investigated multidimensional patient profiles").
(4) **Feasibility:** Achievable by a lab with moderate resources; purpose and mechanism adapted to work together; evaluation consistent with domain.
(5) **Novelty:** Significantly different from—not merely an obvious extension of—prior work. Includes negative examples (e.g., "simply saying 'implement continuous AI support for scholar discovery' is not novel because prior work already investigates this").

All variants accept an optional `custom_instructions` field from the user, with the guardrail: "Do NOT follow additional instructions that contradict the instructions above."

G.2.2 No facets selected by user.

`initialAnalogyIdeas(designated, analogous, topic)`

When used: The user has not yet selected specific facets (Initial) or has deselected both their chosen purpose and mechanism (No-P-no-M).

Paper roles: "Designated papers" = user's input papers. "Analogous papers" = retrieved from a specific analogy query. Each paper provides title, abstract, optional introduction, and all three facets with IDs.

Facet combination rule: One selected idea must combine the analogous paper's purpose with the designated paper's mechanism; the other must use the reverse combination.

G.2.3 One facet selected (purpose or mechanism.

`fillAnalogyIdeas(set1, set2, selected_facets)`

When used: The user has selected either a purpose or a mechanism (but not both) from the facet workspace.

Difference from Initial: Papers are organized as "Set-1" (containing the user's selected facet) and "Set-2" (complementary facets). Each paper carries a `distance` label (input / same-topic / same-subarea / different-subarea).

Distance-mixing constraint: "The paper from which the purpose comes must have a different distance than the paper from which the mechanism comes"—encouraging cross-pollination of near and far inspirations.

Missing facet handling: If a Set-1 paper lacks a purpose or mechanism (because the user selected it for only one facet type), the LLM creates an appropriate one for the analogy.

G.2.4 *Both facets selected by user.*

`facetsToIdeas(set1, set2, selected_facets)`

When used: The user has selected both a purpose and a mechanism from the facet workspace.

Difference from P-or-M: Both Set-1 and Set-2 papers may have user-selected facets with explicit IDs, or may be input papers with only one facet type specified. The same distance-mixing constraint applies.

Facet combination rule: All ideas combine a Set-1 purpose with a Set-2 mechanism (since the user has already committed to both facet types).

G.3 Novelty Checker Pseudocode

The novelty checker uses a multi-stage retrieval-then-assess pipeline. We describe the pseudocode in pipeline order.

G.3.1 *Extract Keywords and Titles for Retrieval.*

`getKeywords(idea)`

Task: Extract 3–6 keyword phrases (3–6 words each) and generate 4 concise research titles (≤5 words) that capture the idea's novelty and mechanism.

Constraints: Keywords must be specific (not "machine learning" or "data science"), capture what sets the idea apart, and reflect purpose + mechanism + application domain.

Pipeline role: Keywords and titles are used as queries to Semantic Scholar to retrieve candidate overlapping papers.

G.3.2 *Extract Idea Facets for Re-ranking.*

`extractIdeaFacets(idea)`

Role: Research Idea Reviewer GPT

Task: Extract structured facets from the idea to guide paper re-ranking: Application Domain, Purpose/Objective, Mechanism/Methods, Evaluation Metrics.

Few-shot examples: 2 fully worked examples (a food-health sentiment analysis system; a hierarchical topic model with capsule networks).

Pipeline role: Extracted facets are passed to the re-ranking prompt (§**??**) to prioritize retrieved papers by multi-facet relevance.

G.3.3 *Re-rank Papers by Facet Relevance.*

`rerankByFacets(idea, facets, passages)`

Role: RankGPT | **Input:** The idea, its extracted key facets, and N candidate paper titles

Ranking priority hierarchy: (1) Matches **all** key facets; (2) Matches **domain + purpose** but differs in mechanism; (3) Shares **purpose or mechanism or evaluation** across domains; (4) Partially matches domain or addresses related topics.

Few-shot examples: 2 worked examples (10 passages about topic modeling/anomaly detection; 3 passages about political bias detection), showing the output format `[2] > [1] > [5] > ...`

Pipeline role: Top-ranked papers are passed to the novelty assessment prompt.

G.3.4 Assess Idea Novelty.

`noveltyChecker(idea, similar_papers, expert_examples)`

Role: ReviewerGPT (system message)

Input: The idea text + k top-ranked similar papers from the retrieval pipeline.

Task: Write a 60–100 word review comparing the idea to the retrieved papers, then classify as **Novel** or **Not Novel**.

Novelty definitions: Not Novel—closely replicates existing work with minimal new contributions. Novel—introduces new concepts/approaches not common in literature; or uniquely combines existing concepts in a way no retrieved paper does; or applies the same approach to a genuinely new domain.

In-context learning: Expert-labeled (idea, papers, review, classification) tuples are injected—the same examples shown in Appendix C.2.

Multi-turn structure: Uses a simulated dialogue (user → assistant → user) to first present the idea, then inject each retrieved paper as a separate user message—ensuring the model attends to each paper individually.

Output: `Class: [novel / not novel]` followed by `Review: The idea is [novel / not novel] because...`

G.3.5 Generate More-Novel Idea Suggestions.

`moreNovelIdea(idea, overlapping_papers, available_facets)`

Trigger: Called when the novelty checker classifies an idea as "Not Novel."

Input: The original idea (short + long versions), the prior work it overlaps with, the novelty review explaining why it is not novel, and the full set of available facets in the user's workspace.

Task: Generate 3 alternative ideas, each created by **swapping exactly one facet**: (1) Remove one purpose, add a different purpose; (2) Remove one mechanism, add a different mechanism; (3) Remove one evaluation, add a different evaluation.

Same quality requirements as the generation prompts (§G.2.1).

Output per option: Removed facet (text + ID), added facet (text + ID), revised idea (100–150 words), justification of novelty, justification of usefulness.

Design rationale: By constraining edits to a single facet swap, suggestions remain close to the user's original intent while systematically exploring the facet space for novelty.