

LitPivot: Developing Well-Situated Research Ideas Through Dynamic Contextualization and Critique within the Literature Landscape

Hita Kambhamettu*
Computer and Information
Science
University of Pennsylvania
Philadelphia, USA
hitakam@seas.upenn.edu

Bhavana Dalvi
Mishra†
Allen Institute for AI
Seattle, Washington, USA
bhavana.dalvi2023@gmail.com

Andrew Head
Computer and Information
Science
University of Pennsylvania
Philadelphia, USA
head@seas.upenn.edu

Jonathan Bragg
Allen Institute for Artificial
Intelligence
Seattle, Washington, USA
jbragg@allenai.org

Aakanksha Naik
Allen Institute for AI
Seattle, Washington, USA
aakankshan@allenai.org

Joseph Chee Chang
AI2
Allen Institute for AI
Seattle, Washington, USA
josephc@allenai.org

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

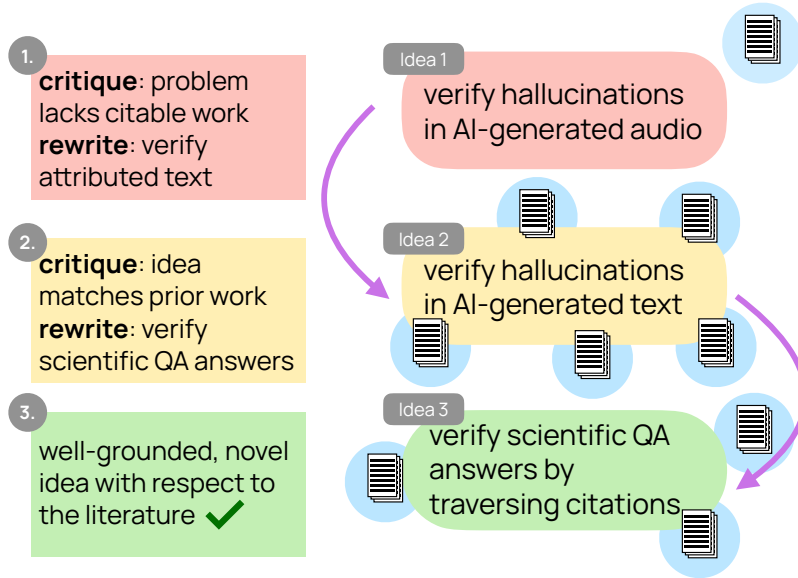


Figure 1: LitPivot supports the co-development of a research idea and its grounding in the literature through *literature-initiated pivots* (in purple). (1) The authors’ original idea (verifying hallucinations in AI-generated audio) has too little prior work to motivate it, so LitPivot suggests pivoting to verifying hallucinations in AI-generated text. (2) That idea is not novel, so LitPivot suggests traversing retrieved citations in scientific RAG systems. (3) The resulting idea builds on prior work while remaining differentiated from it.

*Work completed during an internship at Ai2.

†Now at Google; work conducted while at Ai2.



This work is licensed under a Creative Commons Attribution 4.0 International License.
UIST '26, Detroit, MI, USA

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2856-3/26/11

<https://doi.org/10.1145/3830398.3830577>

Abstract

Developing a novel research idea is hard. It must be distinct enough from prior work to claim a contribution while also building on it. This requires iteratively reviewing literature and refining an idea based on what a researcher reads; yet when an idea changes, the literature that matters often changes with it. Most tools offer limited support for this interplay: literature tools help researchers understand a fixed body of work, while ideation tools evaluate ideas against a static, pre-curated set of papers. We introduce *literature-initiated pivots*, a mechanism where engagement with literature prompts revision to a developing idea, and where that revision

changes which literature is relevant. We operationalize this in LIT-PIVOT, where researchers concurrently draft and vet an idea. LIT-PIVOT dynamically retrieves clusters of papers relevant to a selected part of the idea and proposes literature-informed critiques for how to revise it. A user study ($n=17$) shows researchers produced ideas rated by experts as better grounded, perceived their ideas as more novel and useful, and reported stronger understanding of the literature space. An open-ended study ($n=5$) reveals how researchers use LITPIVOT to iteratively evolve their own ideas.

CCS Concepts

• **Human-centered computing** → **Interactive systems and tools.**

Keywords

research ideation, literature synthesis, human-AI collaboration

ACM Reference Format:

Hita Kambhamettu, Bhavana Dalvi Mishra, Andrew Head, Jonathan Bragg, Aakanksha Naik, Joseph Chee Chang, and Pao Siangliulue. 2026. LitPivot: Developing Well-Situated Research Ideas Through Dynamic Contextualization and Critique within the Literature Landscape. In *The 39th Annual ACM Symposium on User Interface Software and Technology (UIST '26)*, November 02–05, 2026, Detroit, MI, USA. ACM, New York, NY, USA, 19 pages. <https://doi.org/10.1145/3830398.3830577>

1 Introduction

This paper was originally going to be about a human-AI system to verify hallucinations in AI-generated audio files. How did we get here? After drafting an initial idea, we did a literature search and found that only two research groups were interested in this space. We were one of them. To ensure we addressed a problem important to the community, we pivoted to verifying hallucinations in AI-generated attributed text. A second literature search revealed the opposite problem: the space was saturated. While this validated the community's interest, it meant we would need to articulate a novel contribution. Our focus narrowed to helping users make sense of scientific RAG QA answers by traversing their retrieved academic citations. This raised a question: who needs to do that, and why? One answer was researchers developing an idea, determining whether existing literature supports it. In other words, us. Supporting that need, a researcher checking a developing idea against the literature, is ultimately what this paper contributes.

Changes to our idea were informed by literature. This is consistent with prior work that states that literature search and idea development are intertwined [12, 15]: through literature review, researchers iteratively change their ideas as they compare against related work and reconsider what is novel, important, and feasible [21, 30, 34]. Existing literature can expose a flaw in an idea's motivation and inspire a stronger framing. The changes to the idea in turn shape which literature becomes most relevant. We call this phenomenon when engagement with literature prompts a revision to a developing idea, and where that revision changes which literature is relevant, a *literature-initiated pivot*.

Prior theories of design and scientific reasoning also conceptualize phenomena similar to *literature-initiated pivots*. Concept-Knowledge Theory describes design as movement between an evolving concept space and existing knowledge [23]; related perspectives in scientific reasoning and creativity similarly emphasize reciprocal movement between generating possibilities and evaluating them against available knowledge [16, 29]. These frameworks emphasize that the relevant knowledge landscape depends on how the hypothesis has evolved, and vice versa. These perspectives motivate a design opportunity for research ideation tools. Systems should support ideation as a tight, iterative loop: where actively developing an idea surfaces the literature that grounds, challenges, and reshapes it, which in turn meaningfully changes the idea.

Existing tools support understanding the literature space or developing the idea largely in isolation. On one hand, systems for augmented reading [3, 19, 24], scholarly synthesis [26, 27], and literature review [37] help researchers read papers, identify themes across them, and build overviews of a research space. However, they provide less support for carrying insights from specific papers back into an evolving research idea [27, 37]. On the other hand, ideation tools help researchers generate research questions [35], assess idea novelty [40], and iterate on an idea [39]. While some systems provide literature-grounded feedback [39, 40], they use a fixed, user-curated set of papers assembled before ideation begins.

In this work, we explore *literature-initiated pivots* in the context of early-stage research ideation, where ideas evolve quickly and developing them requires reciprocal engagement with the literature [12]. We introduce LITPIVOT, a system designed to support the co-development of a research idea and its grounding in a literature space. A researcher can check a facet of their idea (problem, contribution, or evaluation) against the literature that is retrieved and organized for that aspect. Using bipartite graph-based and LLM-simulated reasoning, LITPIVOT maps the idea aspect onto the aspects of the retrieved literature, showing how the idea follows or differs from prior work and how the idea can be re-articulated to be more grounded or distinct. The researcher can then review this output, which includes references back to the literature, and integrate this updated understanding to refine their idea. LITPIVOT is intended for researchers who already hold a nascent idea and want to develop it against the literature. While ideas can originate from many sources (e.g. conversations, observations, practical needs), once an idea has been articulated, engaging with the literature is an important way researchers refine, differentiate, and strengthen it. This is the stage of ideation that LITPIVOT supports.

Three studies inform and validate LITPIVOT. First, a formative study ($n = 5$) and research artifact analysis ($n = 4$) suggest challenges researchers face when making literature-initiated pivots. Second, in a user study ($n = 17$) comparing LITPIVOT to a baseline system that supports ideation in a popular current practice, AI chat-with-papers QA, we find that using LITPIVOT leads researchers to consult significantly more papers when developing their idea, report a stronger understanding of the literature space, and produce ideas that were rated by experts as better grounded in the literature. Finally, in a qualitative, open-ended study, we observe literature-initiated pivots in practice as participants used LITPIVOT to develop their ongoing ideas in relation to the literature.

We conclude by discussing how LITPIVOT may have improved researchers' engagement with the literature and the grounding of their ideas, and how it shaped the ideas themselves. We discuss how ideation in other knowledge-intensive domains may also benefit from interactive integration with a corpus of knowledge. Our work suggests that the corpus could be a more active participant in shaping new insights in these domains. In summary, this paper:

- Defines the *literature-initiated pivot* as the process where engagement with literature prompts the revision of a developing idea, which in turn shifts the relevant literature, and identifies it as a key mechanism to support during ideation.
- Contributes a system, LITPIVOT, that operationalizes support for literature-initiated pivots in early-stage research ideation through dynamic literature reorganization and graph-based reasoning over a literature space.
- Provides evidence from a user study ($n = 17$) that supporting literature-initiated pivots leads researchers to report deeper understanding of the literature and develop better-grounded ideas. A qualitative study ($n = 5$) illustrates how researchers benefit from making these pivots.

2 Related Work

2.1 Ideation as Dual-Space Search

Theories of creative cognition and scientific reasoning characterize discovery as an iterative movement between two distinct spaces: the *already known* and the *not-yet-known*. For instance, Concept-Knowledge (C-K) theory models design as reciprocal transitions between validated knowledge and unvalidated concepts [23]. Similarly, Klahr and Dunbar [29] describe scientific discovery as a dual search across interacting hypothesis and experiment spaces, while the Geneplore model emphasizes cycles of generative and exploratory processes for vetting partially developed creative ideas [17].

In the context of research ideation, this dual-space dynamic takes the shape of a continuous dialogue: a developing idea (the not-yet-known) is shaped by existing literature (the already known), and as the idea evolves, it shifts which literature becomes relevant. LITPIVOT operationalizes this recursive process. By facilitating *literature-initiated pivots*, the conceptual movements between these two spaces, LITPIVOT helps researchers iterate on nascent ideas. It supports this mechanism by segmenting ideas, dynamically clustering relevant literature to each segment, reasoning over the literature space, and incorporating insights back into the idea.

2.2 Tools for Literature Understanding

An important part of research ideation is developing a deep understanding of the landscape of related literature. Yet the volume of scientific literature is so large [2] that developing this understanding is often prohibitively difficult. To help researchers read literature at scale, prior HCI systems support activities such as paper comparison [8, 10, 22] and literature discovery [26–28, 32, 37, 45]. Many of these systems help users schematize a literature space, for example through semantic alignment [22], similarities between papers [28, 32], or methodological similarities [8].

These works demonstrate the utility of organizing a large literature space. The reported effects include lower cognitive load [22], a better understanding of paper relevance [28, 32], and the creation of

artifacts that accurately reflect the papers read [37]. However, they stop short of carrying those insights into the active development of a research idea. This gap is reflected in the future directions these works identify: Palani et al. [37] note that real-world information seeking often involves multiple queries and ill-defined goals, conditions that resemble research ideation, and suggests extending support for such tasks; and Kang et al. [27] note future opportunities to study literature search in the context of ideation outcomes. LITPIVOT does this by dynamically clustering and ranking literature based on the specific part of the idea a researcher is revising. For example, if a researcher is refining their proposed contribution, LITPIVOT anchors its reasoning on that specific facet, surfacing and comparing the contributions of adjacent papers so the researcher can articulate their idea's novelty.

2.3 Tools for Literature-Supported Ideation

A growing body of HCI work explores scaffolding research ideation with LLMs, contributing frameworks for human-AI co-creation of research questions [35] and accounts of AI-supported ideation and creativity support [33, 41, 46]. These systems focus on idea generation without necessarily grounding it in the knowledge on which it builds. In research ideation, where an idea's merit is defined by its relationship to prior literature, this grounding is critical.

This is not to say that the concept of "changing an idea" has been unexplored. A complementary set of tools actively supports idea refinement and assessment. Closest to our work is Scideator [40], which lets users input a set of research papers, decomposes them into facets (purposes, mechanisms, evaluations), allows users to mix-and-match these facets to form new ideas, and then evaluates their novelty. Similarly, IdeaSynth [39] maps idea facets to nodes on a canvas, enabling researchers to explore combinations through AI-based question-answering over user-provided literature. There are a number of ways that LITPIVOT differs from these tools. Most importantly, LITPIVOT focuses on the reciprocal relationship between understanding a literature space and developing an idea: the core mechanism of the literature-initiated pivot. This dynamic is largely elided in IdeaSynth and Scideator, where the main contributions lie in generating many variations of an idea. While the literature might be consulted in this generation, dynamically exploring and reacting to it is not the focus of these interactions. This difference is evident in the documented limitations of both systems. Scideator [40] retrieved a small number (10) of papers per idea and reported that users stated that the system classified ideas as novel too often and missed important related works. IdeaSynth [39] reports that literature curation in the system is largely manual. Participants felt its recommendations were too limited, making other work hard to discover. The outcomes evaluated with these systems reflect these differing goals. IdeaSynth states that users explored more alternative ideas using the system but did not observe any significant difference in users' reported confidence in their understanding of the literature, and Scideator reports increased creativity support (and did not measure difference in literature understanding). We evaluate LITPIVOT on the outcome its mechanism targets: participants reported significantly higher understanding of the literature, and experts rated their ideas as significantly more

grounded in relevant literature. We consider groundedness as one signal for evaluating early-stage research ideas.

3 Understanding literature-initiated pivots

We conducted an analysis of research meeting notes and a formative study to gain insight into the following research questions:

- FQ1: When and how do literature-initiated pivots occur during the process of developing a research idea?
- FQ2: What challenges do researchers encounter when making literature-initiated pivots?

3.1 Understanding when and how literature-initiated pivots occur (FQ1)

We conducted a document analysis of meeting notes (obtained with permission) from four completed computer science research projects (two HCI and two NLP) at the authors' institution. The first author did an iterative open coding on these notes [4, chapter 5], noting when literature-initiated pivots occurred¹.

Meeting notes referenced literature in an average of 7.5 meetings per project ($SD = 2.65$). Over the four projects, we identified 21 literature-initiated pivots. We observed two main types of pivots: pivots where the research idea differentiates itself from existing literature ($n = 9$) to articulate novelty, and pivots where the idea aligned with literature ($n = 12$) to adopt best practices or strengthen validity. We go into detail about literature-initiated pivots that occurred early on in the project, as these were the situations that LITPIVOT was designed to support.

3.1.1 Differentiating from existing literature. In these pivots, researchers refined their idea to establish its novelty relative to literature that closely mirrored their proposed contribution. For example, one project initially referenced literature about identifying relations between individual claims, such as detecting a contradiction between “The dog is blue” and “The dog is red.” Later, an author shared a paper suggesting that this problem was largely solved, but pointing to an open challenge in resolving information across hierarchies, such as relating high-level summaries to granular sub-claims. The project then pivoted from identifying claim relations to developing claim hierarchies. This illustrates how the idea was revised to differentiate it from existing literature.

3.1.2 Aligning with existing literature. Authors also referenced existing literature that was similar to parts of their idea. For example, the authors of one project were developing a system that relied on AI for specific text and image extraction tasks. The authors referenced a canonical paper's evaluation as being one to adopt. In this case, the authors sought out literature that was similar in one dimension, evaluation methods, to the project.

¹We identified a literature-initiated pivot as any instance where the meeting notes documented a researcher citing a specific paper as the rationale for changing a facet of their idea, or where the notes indicated that a particular paper or literature space should be reviewed to inform a change to the idea. The de-identified coding summary will be made available in supporting materials.

3.2 Understanding challenges in making literature-initiated pivots (FQ2)

We recruited 5 researchers (3 HCI, 2 NLP; 3F, 2M; median age = 27) via our organization's Slack channels and professional connections for a formative study. These researchers had a median of five years of research experience and were actively engaged in developing a research project. Each participant brought a research idea in progress, but not yet fully developed, to the study.

During the session, participants developed a short research proposal across three phases. In Phase 1 (10 minutes), participants used only a document editor to articulate their initial ideas. In Phase 2 (15 minutes), participants refined their ideas using standard search tools. This allowed us to observe how they naturally integrate literature into their ideation without augmentation. In Phase 3 (15 minutes), participants used a design probe² that provided literature-grounded feedback on the idea. This phase evaluated the benefits and challenges of using literature-informed support to revise ideas.

All sessions were conducted over Google Meet by the first author. The first author analyzed the transcripts using open coding followed by thematic analysis [4, Chapter 5] to identify common challenges and opportunities. The analysis was refined through group discussions with the other authors.

Two participants pivoted their idea during Phase 2. All five participants pivoted their idea during Phase 3. This suggests that literature-initiated pivots happen even within a relatively short duration (approximately 30 minutes across Phases 2 and 3). We detail the results of our formative study below. We identified two primary challenges: balancing the tension between expanding on the idea and critically evaluating it against related literature, and pivoting when literature challenges the novelty of the idea.

3.2.1 CH1: Tension between developing the idea and consulting related literature. When searching and reading literature in Phase 2, participants had two goals: curating the foundational work their idea builds upon ($n = 3$), and showing how their idea differs from it ($n = 5$). Both proved challenging. Consider the case of P4. After drafting an idea, P4 wanted to check if existing work had already proposed a similar contribution. Finding three such papers, they read each to note its motivation and contribution, then copied over the papers' future work and limitations sections. Reviewing these notes prompted P4 to revise their idea's methods and evaluation. By the end of the task, they reflected, “*before, I felt very strongly about this project...but this has made me question the novelty of it.*” This suggests a need for support when researchers are concurrently developing and evaluating their ideas against prior work.

3.2.2 CH2: Pivoting when literature challenges the idea's novelty. A central part of evaluating an idea is articulating its novelty, especially when existing literature raises doubts about it. Participants described the kind of support they wanted in these moments. Consider the case of P3. After reading two papers that substantially overlapped with their core contribution, P3 found themselves in what they called a “research rut.” To get out of it, they wanted

²Built off of [48], the design probe was an LLM-based system that analyzed the quality of the research idea. It generated an initial evaluation, recommending changes based on referenced literature, and producing arguments for and against the idea's core claims (see Appendix F).

guidance on “*how to recommend concretely what I can do to fix this problem in a novel way.*” Importantly, P3 did not want a completely new idea; instead, they said they “*want [support] to prompt me to think about certain aspects of the research problem, rather than thinking for me.*” This suggests that support should go beyond surfacing relevant literature, and instead provide specific, literature-backed ways to articulate an idea’s novelty over prior work.

3.2.3 Additional challenges. Our formative work surfaced other difficulties researchers face when engaging with literature during ideation. One was discovering literature that participants did not know how to search for: P1 described the anxiety of unknown related work, “*the fear is there are demons out there lurking around in the related work... like they did exactly this before.*” Another was the gradual internalization of literature, which unfolded over longer than a single session: P5 described ideation as an interleaved process of “*read literature and write, read literature and write,*” until “*there’s enough literature sustaining the motivation of the [work].*”

We designed LITPIVOT around CH1 and CH2 because they were the most-observed challenges specific to literature-initiated pivots. LITPIVOT’s cited-paper expansion (Section 4.2.1) also partially supports discovery, a complementary challenge to internalization. However, LITPIVOT was not designed to support literature discovery and internalization beyond the scope of CH1 and CH2.

3.3 Design goals

We identify two key opportunities:

- **DG1: Support dynamic curation to align and differentiate an idea.** Our document analysis (Section 3.1) showed that researchers seek literature both to align with and differentiate from specific parts of their idea. LITPIVOT supports this by iteratively decomposing the evolving idea into facets and organizing the literature around them (Section 4.2.2), helping researchers see the exact subset relevant to the part of the current idea they want to develop.
- **DG2: Use literature-based suggestions to inspire literature-initiated pivots.** Participants reported a tension between developing their ideas and consulting literature to refine them (Section 3.2.1), especially when prior work threatened their idea’s novelty (Section 3.2.2). Participants wanted more than conflicts surfaced; they wanted concrete recommendations for how to pivot. Therefore, LITPIVOT goes beyond flagging overlaps, applying graph-based checks and lightweight formalizations over the literature space (Section 4.3) to suggest specific, actionable rewrites to the idea.

4 System

LITPIVOT can be used to help researchers develop an initial idea in dialogue with the literature. LITPIVOT features a text editor where researchers can further develop their ideas (Figure 2) and break them down into specific facets. Following prior work on faceted research ideation [39, 40], we structure these as problem, contribution, and evaluation statements. These facets serve as clickable conduits for checking specific parts of the idea against the literature. When a researcher selects a facet, LITPIVOT dynamically clusters the literature space around that specific component, visually highlighting the most relevant clusters. After the researcher

selects a subset of these papers, LITPIVOT evaluates the facet against the literature. For example, to evaluate a problem statement, LITPIVOT extracts problem statements from the selected papers to check if the proposed idea addresses a problem that is of interest to the research community. Conversely, to evaluate a contribution, LITPIVOT identifies how it differentiates itself from existing work (detailed via our bipartite graph approach in Section 4.3.2). Finally, alongside a report on the facet’s literature fit, LITPIVOT provides concrete, actionable suggestions for revising that facet.

4.1 Narrative scenario

We describe the experience of interacting with LITPIVOT in a narrative walkthrough. Cy, an NLP researcher studying LLM evaluation on complex reasoning tasks, wants to develop this idea: using LLMs to translate informal plans into precise specifications that verification tools can test for correctness. She knows the NLP literature well but is less familiar with the relevant formal methods work. She turns to LITPIVOT and types her initial idea into the text editor.

Exploring faceted literature space. After writing out her idea, Cy presses a modal key, triggering LITPIVOT to retrieve all related papers and list them in the right-hand panel. Cy wants to know whether her contribution is novel. She clicks the “Enable facet highlighting” button, which divides her idea into facets.

She clicks on the contribution segment of her idea. The side panel reorganizes the literature into labeled clusters of prior contribution types, such as LLM-based formal verification frameworks or claim verification systems. These clusters help Cy understand the contribution types in the existing literature. The clusters most relevant to her proposed contribution are starred (see Figure 2).

Assessing and updating an idea facet. Cy selects papers in clusters to assess her proposed contribution against and generates an assessment. LITPIVOT notes the proposed contribution’s overlap with prior work: VERIPLAN [31] “already couples an LLM, a rule translator, and a model checker.” This tells Cy that her contribution as written is likely not novel. The assessment also highlights an open limitation articulated in another paper LEANREASONER [25]: “the brittleness in commonsense formalization.” This intrigues Cy. To understand this gap more before pivoting, she reads an evidence snippet from LEANREASONER’s discussion section. She learns that while prior work verifies rigid task plans, it fails when constraints rely on implicit, real-world knowledge.

She reviews LITPIVOT’s suggested rewrite, which proposes translating free-form creative narratives into formal constraints. Cy reflects on how narratives rely on the exact commonsense reasoning that LEANREASONER struggles with and reads through the LEANREASONER paper to get a better understanding of its methods. Cy reframes her research contribution. Instead of her original idea of translating rigid plans into specifications, she pivots to using an LLM to translate free-form narratives into temporal logic, using a model checker to verify the translation. Because her contribution has shifted in both domain and type, LITPIVOT dynamically updates the literature space. New clusters of papers become relevant, including work on evaluating LLMs in creative writing tasks and on combining LLMs with temporal logic. This updated view saves Cy from the friction of restarting her manual search from scratch.

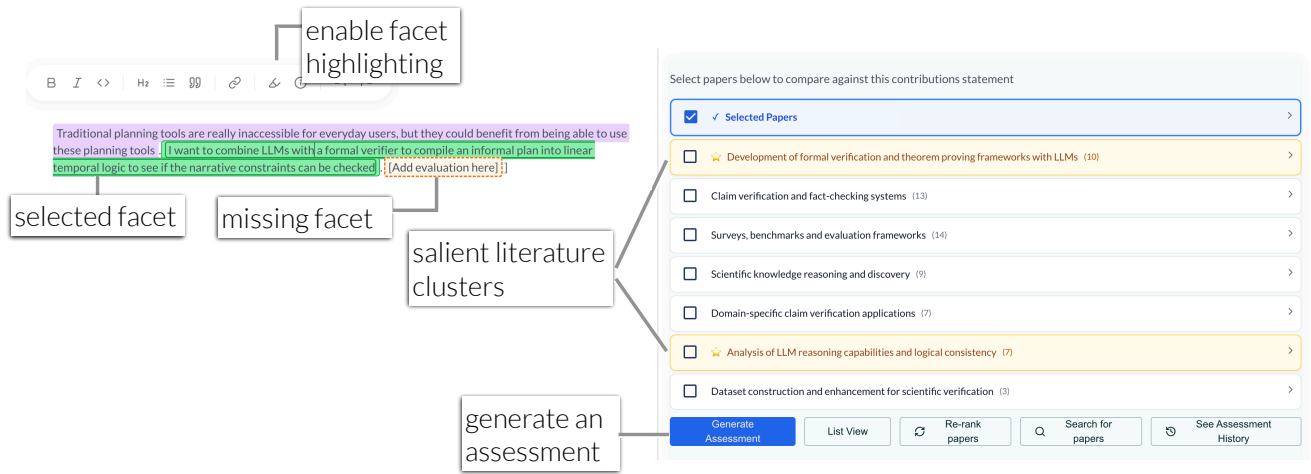


Figure 2: The LitPivot interface. The left pane is a document editor with a toolbar toggle to highlight facets. The right pane is a paper panel that surfaces literature clusters. The interface colors text by idea facet: problem (purple), contribution (green), and evaluation (orange; currently missing). Clusters most relevant to the selected facet (here, the contribution statement) are starred. After selecting papers, the researcher can generate an assessment to compare the idea against selected work.

Instead, she can stay focused on developing her idea while the literature dynamically adapts to her idea.

Maintaining idea congruency. Cy must now re-articulate what problem motivates a system like this. LITPIVOT brings this to her attention by flagging the now-outdated problem facet, signaling that the idea’s internal congruency has broken.

Cy clicks the problem segment of her idea. LITPIVOT organizes the literature into clusters based on the problems addressed in prior work. She selects several papers that relate to the problem her new contribution might address, such as inconsistencies in free-form narrative generation, to check her idea against. LITPIVOT surfaces a well-documented, open challenge in that literature: maintaining plot lines in long narratives often exceeds an LLM’s context window. Seeing the direct connection between this documented gap and her new direction, Cy sees that her contribution is well-suited to solve it. She uses insights from LITPIVOT to rewrite her problem statement, focusing on how formal verification can preserve temporal consistency in long-form storytelling, a task that heavily relies on the commonsense reasoning she identified earlier. Cy now feels ready to further develop this refined idea. Ultimately, LITPIVOT supported Cy in making several literature-initiated pivots: reshaping an already-existing contribution into one that addresses an open problem in the research community.

4.2 Constructing the literature space

The literature space was organized to: 1) curate a large set of relevant literature, and 2) inspire inquiry into the details most useful to the researcher. The pipeline is in Figure 3.

4.2.1 Retrieving relevant literature. LITPIVOT passes the researcher’s idea as a seed query to Asta Paper Finder³, an open-sourced LLM-powered scientific search engine. Paper Finder retrieves an initial

set of relevant papers by Semantic Scholar ID. The IDs are passed into the Semantic Scholar Snippet Text API⁴ to retrieve the full text for each paper. Each paper’s limitations, future work, contributions, motivations, evaluations, and methods sections are extracted. If an extracted section cites a work not already in the literature set, LITPIVOT adds the referenced paper to the corpus. The final literature set thus includes both the most directly relevant papers and the foundational work that informed them. An LLM categorizes each paper based on the researcher’s idea⁵ as perfectly relevant, somewhat relevant, complementary, or not relevant. In the UI, papers appear in descending order of relevance. Researchers can request additional literature at any point, either by adding papers via the Semantic Scholar Academic Graph API⁶ or by prompting the system to fetch new results based on a revised idea.

4.2.2 Organizing the literature space. To surface the literature most useful to the researcher, LITPIVOT clusters papers conditioned on specific facets of the research idea. LITPIVOT groups extracted problem, contribution, and evaluation sections from related work. Papers sharing similar facets are then clustered and labeled by type (Appendix B.0.10). To surface the most relevant clusters during editing, LITPIVOT passes the current facet text and all cluster names to an LLM, which identifies the clusters most useful (Appendix B.0.9). Literature-initiated pivots may shift which papers are most relevant, and researchers can trigger an update to refresh the literature view. When the user clicks the re-rank button in the global view, LITPIVOT recalculates paper relevance (detailed in Appendices B.0.10 and B.0.9). We designed this as a user-initiated interaction to manage system latency and give researchers control over when to reorganize the entire literature space around their revised idea.

⁴<https://api.semanticscholar.org/api-docs/graph#tag/Snippet-Text>

⁵Prompt adapted from Asta Paper Finder: <https://paperfinder.allen.ai/>

⁶<https://api.semanticscholar.org/api-docs/graph>

³Paper Finder, <https://github.com/allenai/asta-paper-finder>

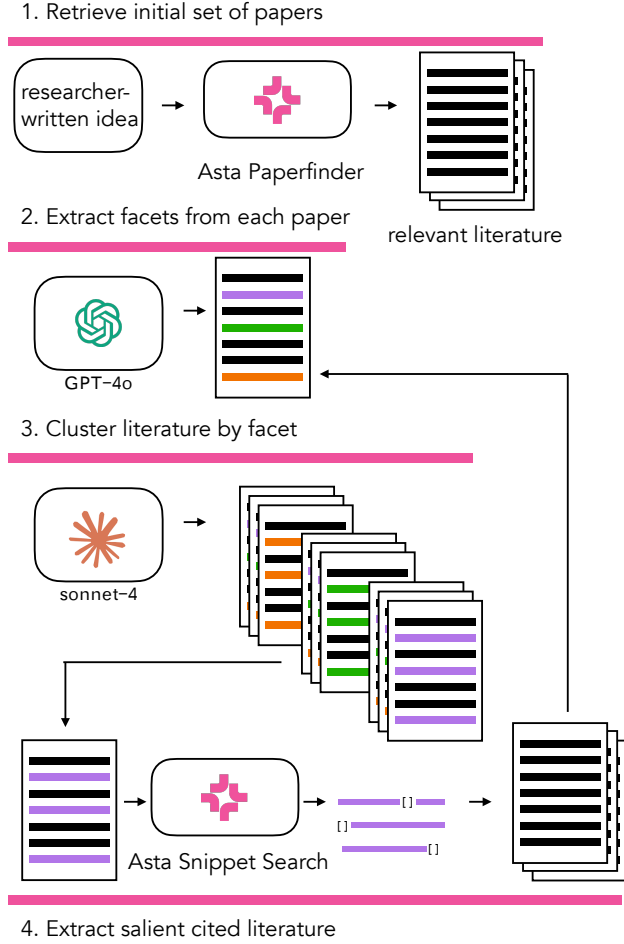


Figure 3: LitPivot pipeline for literature retrieval and faceted paper clustering. (1) The researcher’s idea is used to retrieve an initial set of relevant papers via Asta PaperFinder. (2) GPT-4o extracts idea-relevant facets from each paper. (3) Sonnet-4 clusters the literature by facet, surfacing groups of papers most relevant to each part of the idea. (4) The system extracts papers that are meaningfully cited within the extracted facets and includes this in the literature set.

4.3 Supporting the literature-initiated pivot

We employ graph-based checks and retrieval-augmented generation to check an idea facet against selected literature.

4.3.1 Align facet with literature. Our formative study revealed that participants seek literature to either motivate a problem or adopt an evaluation methodology. Inspired by [13], we assert that a robust problem statement must: (1) be grounded in citable prior work, (2) establish the problem’s existence through evidence of failures, gaps, or pain points, and (3) argue for its significance through impact, risks, or opportunity. A robust evaluation statement must: (1) assess whether the contribution aligns with the problem, (2) test whether it answers the research questions, and (3) verify its plausibility against how similar contributions are evaluated in prior work.

These criteria are translated into prompts (see Appendices B.0.6 and B.0.3). Since LitPivot prioritizes literature-driven framing, the problem assessment checks whether the draft can cite at least one matching statement from the selected papers. The evaluation assessment conditions on the proposed problem, contribution, and few-shot examples of how selected papers evaluate similar work. As prior work may sometimes use suboptimal methods, LitPivot surfaces literature patterns for the researcher to review and selectively adopt. Researchers can accept or reject suggestions and request additional context to steer the search. Following these assessments, an LLM generates literature-informed rewrites: the problem rewrite must specify (1) the problem, (2) evidence it is real, and (3) why it matters; the evaluation rewrite must specify (1) whether the evaluation shows the contribution addresses the problem, and (2) whether the plan is feasible and sensitive enough to detect intended effects.

4.3.2 Differentiate facet from literature. A proposed contribution must articulate its relationship to and distinction from existing work, a task our formative study identified as particularly challenging (Section 3). We formalize this by constructing a bipartite graph: one vertex set represents contributions extracted from selected papers, the other represents their limitations or future work statements, and edges indicate that a contribution addresses a specific limitation. An LLM prompt constructs this graph by taking all extracted contributions and limitations/future-work statements as input and returning a bipartite graph in JSON. A second prompt links the proposed contribution to known limitations and future work: links solely to already-addressed items suggest an incremental contribution, replication, or extension, while links to unaddressed items suggest conceptual novelty (see Appendices B.0.4 and B.0.5). LitPivot then rewrites the contribution against three criteria: (1) does it directly address the stated problem, (2) is success plausible given the methods and constraints, and (3) how is it positioned relative to the linked limitations and future work?

4.3.3 Maintaining consistency between facets. Because the problem, contribution, and evaluation facets that make up a researcher’s idea are interdependent, if a researcher revises one facet, they might need to go back to other parts of the idea and revise them to make sure the idea is internally congruent. When a facet is changed, LitPivot passes the revised idea, the literature analysis, and the newly edited facet to an LLM, instructing it to identify which other facets should be updated to maintain consistency. If a mismatch is detected, for example, if a newly revised contribution no longer addresses the original problem or requires different evaluation metrics, LitPivot visually highlights the affected facets to prompt the researcher for further revision (Appendix B.0.7). Additionally, after segmenting the initial idea, LitPivot flags any core facets that have not yet been articulated, prompting the researcher to draft the missing facet (Appendix B.0.8). A researcher can also request a full assessment of the entire idea at any time (Appendix B.0.11).

4.4 Implementation

LitPivot is implemented as a web application. The frontend uses the React framework in JavaScript, while the backend is implemented in Python using Flask. Once retrieved, paper data is stored as JSON files. Different language models are used across the pipeline

based on the task. Broadly, OpenAI’s GPT-4o (temperature=0.7) is used for low-level tasks: segmenting ideas into facets, extracting facets from paper full-text, identifying relevant clusters, detecting missing facets, and flagging affected segments after an edit. OpenAI’s o3 (temperature=0.7) is used for tasks involving reasoning: graph instantiation and generating evaluations and suggestions for each facet. Anthropic’s sonnet-4 (temperature=0.7) is used for tasks involving long contexts, such as clustering papers by facet.⁷

4.5 Technical Evaluation

This section assesses the LITPRIVOT pipeline: does a pivot-supporting pipeline produce better literature-grounded research idea feedback than that of prior systems? We compare LITPRIVOT’s pipeline against another literature-based ideation system, IdeaSynth [39]. Rather than benchmarking against the growing landscape of AI writing tools [36], into which we conjecture LITPRIVOT’s pipeline could eventually be integrated, we focus on assessing it against a research-grade baseline. Two domain experts, recruited through professional networks, rated outputs from both systems across 5 seed ideas drawn from their areas of expertise, each decomposed into problem, contribution, and evaluation facets (15 segments total). The LLM and literature corpus were held constant across conditions, with each idea’s source paper withheld. Blinded to condition, experts rated on 1–7 Likert scales: (i) how well the evaluation identified strengths and weaknesses relative to the literature, and (ii) whether incorporating the suggestions would yield a stronger paper.

LITPRIVOT outperformed the baseline on both measures. We use the Wilcoxon signed-rank test as our primary inference and report paired *t*-tests as a convergent robustness check, including mean differences, 95% CIs, and paired effect sizes (d_z). For suggestions, both raters independently favored LITPRIVOT over baseline with large effects (Rater 1: $M=3.40 \rightarrow 5.27$, $M_{diff} = +1.87$, $t(14)=4.00$, $p=.001$, $W=8$, $p=.005$, $d_z=1.03$; Rater 2: $M=3.73 \rightarrow 5.73$, $M_{diff} = +2.00$, $t(14)=3.94$, $p=.002$, $W=8$, $p=.005$, $d_z=1.02$). Aggregating both raters ($n=30$ paired judgments) yielded a mean improvement of +1.93 points ($t(29)=5.71$, $p<10^{-4}$; $W=30$, $p=7.1 \times 10^{-5}$; $d_z=1.04$). For evaluations, one rater showed a significant preference for LITPRIVOT ($M=3.53 \rightarrow 5.47$, $M_{diff} = +1.93$, $t(14)=2.91$, $p=.012$, $W=14.5$, $p=.016$, $d_z=0.75$), while the other showed a non-significant trend ($M=3.60 \rightarrow 4.40$, $M_{diff} = +0.80$, $t(14)=1.52$, $p=.15$, $W=20.5$, $p=.14$, $d_z=0.39$); the combined analysis was significant with a moderate effect ($M_{diff} = +1.37$, $t(29)=3.18$, $p=.004$, $W=66$, $p=.005$, $d_z=0.58$). This suggests that our pipeline’s design for supporting literature-initiated pivots has the intended positive effects of 1) correctly relating the idea to a selected literature space, and 2) providing actionable ways the idea can be better situated within that space.⁸

5 Study 1: User Study

Participants. We recruited 17 researchers in computer and information sciences whose research aligned with the study tasks using

⁷We assigned models by latency rather than defaulting to the most capable model for every step: LITPRIVOT is an interactive system, and reasoning models are considerably slower, so we reserved them for the steps that most benefit from deeper reasoning. Users can swap in larger models as desired.

⁸Inter-rater reliability was low across conditions (Krippendorff’s α ranging from -0.33 to 0.11), consistent with prior work on the subjectivity of single-item scholarly judgments [5, 38, 42]. Given this, we emphasize within-rater paired contrasts over absolute scores, and interpret the evaluations as such.

professional networks and social media. Because LITPRIVOT aims to help researchers concurrently evolve an idea and their understanding of a literature space, we primarily targeted early-career researchers. While these individuals possess foundational research skills, they presumably benefit more from literature scaffolding than senior researchers who already possess rich, internalized domain knowledge [11, 18]. Accordingly, our participants consisted mostly of PhD students (82%), along with one postdoc and two industry researchers. Participants had varying levels of research experience: 59% reported 2–5 years and 41% reported 6–10 years. Their research areas spanned NLP (9), HCI (3), robotics/RL (2), biomedical/health AI (2), and ML/AI (1). Participants also reported varying use of AI tools for ideation: 18% always, 35% frequently, 35% occasionally, and 12% rarely.

Study setting. The study was conducted over Google Meet. Sessions lasted 90 minutes, and participants received \$75 USD.⁹

Interfaces. In the baseline condition, participants used a QA interface that let them select papers and ask questions (see Figure 6). To isolate the effect of literature-initiated pivots, this baseline did not segment the idea into facets or support targeted pivots through automated assessments, unlike LITPRIVOT. Instead, it reflected a common alternative: a standard text editor paired with RAG-based LLM assistance. Its backend matched the literature-grounded QA feature in IdeaSynth [39] and used the same paper corpus as LITPRIVOT. The QA interface answered questions using the full text of the selected papers and the current idea draft as context.¹⁰ Both conditions used the same set of ~100 papers, retrieved by Asta PaperFinder [1], an agentic search system with sub-agents for query analysis, planning, and iterative retrieval expansion. Before each task, participants received a tutorial for the assigned condition. The experimenter demonstrated the interface’s features and asked participants to practice using them by briefly editing a sample idea.

Test Tasks. Each participant developed two research ideas, one with LITPRIVOT and one with the baseline. Based on participants’ reported familiarity with subareas of HCI and NLP, they were assigned two seed ideas¹¹. We used standardized seed ideas rather than using participants’ own ideas to control for differences in project maturity. Participants had 30 minutes per task. We collected:

- (1) *Behavioral interaction logs.* We recorded logs of text edits, papers selected or retrieved, assessments generated (in LITPRIVOT), and questions asked (in the baseline).
- (2) *Self-reported data.* Participants rated their agreement on 7-point Likert scales (1=strongly disagree, 7=strongly agree) across several dimensions. To measure how perceptions of the ideas changed, we collected pre- and post-task ratings for: (1) perceived novelty, (2) perceived feasibility, and (3)

⁹The study was approved by an internal review board. Our study protocol and materials are included in the supplemental material

¹⁰We counterbalanced condition order to control for learning and ordering effects. Although LITPRIVOT normally supports adding new papers, we did not allow participants to add their own during the study to ensure both conditions used the same paper set.

¹¹An example task topic, “biomedical claims verification,” used the following seed idea: “Verifying biomedical claims is difficult, especially when you have to look through a paper’s citations, and those citations’ citations, to get to the actual study that conducted the experiment that yields primary evidence in support of a claim. We will create a dataset of biomedical claims and the citation ‘hops’ it takes from paper to paper to entirely verify the claim.”

perceived utility of the idea. Additionally, participants provided post-task ratings across three additional metrics: (1) understanding of the literature space, (2) helpfulness of the system’s outputs, and (3) trust in those outputs. We corrected families of related outcomes using the Holm–Bonferroni method. We report estimated marginal means with 95% CIs and effect sizes; for non-parametric data, we use Wilcoxon signed-rank tests with Cliff’s δ .

- (3) *Qualitative data.* At the end of the study, participants reflected on their experiences during both tasks. The lead author analyzed participants’ responses through a thematic analysis process [4, chapter 5]. Analyses were refined through discussion with another author to derive key themes.
- (4) *Expert evaluation of ideas.* The last two authors, who are domain experts in the areas the ideas were from, rated the final ideas, blind to condition. Assessing research ideas requires long-term execution to judge feasibility or impact [6, 39, 49], so we focused on one criterion that encapsulates the primary effect of LITPIVOT: “*This idea is well-grounded and well-argued using relevant literature.*” Ratings used a 7-point scale. We measured agreement with Krippendorff’s alpha and iteratively refined the guidelines to reach consensus. Agreement improved from 0.389 to 0.804 after three rounds of discussion and guideline refinement (about 140 minutes total; Appendix A); remaining conflicts were resolved.

We scope our measures of idea novelty, feasibility, and utility to self-report data. Judging the significance of a research idea before it has been executed can be noisy [38]; self-report data can be assessed at the idea stage and follows practices in prior ideation work [39, 44]. For an early-stage idea, whether the researcher perceives it as novel, feasible, and useful itself might be a practical and useful signal. Measuring downstream idea significance and impact is an important direction for future work (Section 7.2).

We hypothesized that:

- H1: LITPIVOT helps researchers understand the literature space better than the baseline.
- H2: LITPIVOT helps researchers generate better-grounded ideas than the baseline.

5.1 LITPIVOT helps researchers better understand the literature space (H1).

Participants selected significantly more unique papers with LITPIVOT to generate assessments (median=7, $M = 7.1$, $SD = 3.9$) than they did for Q&A in the baseline (median=3, $M = 4.4$, $SD = 4.3$; $W = 15.0$, $p = 0.010$). While an increased document selection count might not indicate quality of interaction, our proxies showed some evidence that participants better understood the literature space that they engaged with. Participants reported a significantly better understanding of the literature space with LITPIVOT ($M = 5.71$) than with the baseline ($M = 3.88$; $t(16) = 5.85$, $p < .001$, $d_z = 1.42$).

Literature search in the baseline. In the baseline, participants ($n = 13$) typically chose the first few papers in the interface ($M = 2.92$ papers, $SD = 1.19$) to ground the chat interaction. A smaller number of participants ($n = 4$) read through the list of papers first and selected a larger set of papers ($M = 9.25$ papers, $SD = 7.23$). On

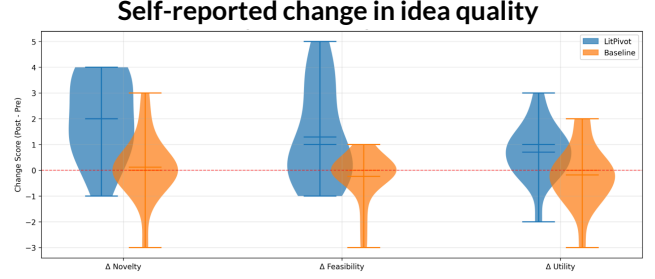


Figure 4: Distribution of change in Likert scores (post task - pre task) for perceived idea novelty, feasibility, and utility across conditions. Horizontal bars indicate median and quartile values; the red dashed line marks no change.

average, participants asked 3.9 questions in the baseline. Common question types were requests to evaluate the idea (13; “Is this idea novel?”), summarize the literature (10; “What do these papers do?”), and clarify details (7; “What benchmarks do the paper report?”).

Selective literature pruning. Using LITPIVOT, we observed 11 participants follow an iterative literature selection process: beginning with all papers in starred clusters, then narrowing the set by deselecting papers, and in some cases adding papers from non-starred clusters ($n = 6$), for subsequent assessments of the same idea facet. Participants revised their idea based on gaps identified in the literature. For example, P1 was editing a contribution about generating “synthetic medical notes that are consistent and factual,” initially selecting 12 papers across three starred clusters. After LITPIVOT identified “temporally coherent multi-visit corpora” as an open problem, P1 reviewed abstracts, deselected six papers focused on single-visit notes, and revised the contribution. After another assessment, P1 further refined the contribution to a technique for “check/reward signals for contradictions across multiple hospital visit notes.” By restructuring their literature space and acting on specific gaps surfaced by LITPIVOT, P1 pivoted their contribution.

Expanding literature assessment. Six participants increased the number of papers used across assessments as they refined their ideas. For example, P9 generated four assessments using 2, 3, 4, and finally 6 papers. Their final idea reflected this expanding literature set: it used statistics from prior work to motivate the problem, combined technical contributions from two papers into a novel contribution, and drew on another paper to ground the evaluation.

5.2 LITPIVOT helps researchers generate better-grounded ideas (H2).

Improvements in self-reported idea qualities. Analyzed as per-participant change scores, participants reported that their ideas improved in the following dimensions (see Figure 4):

- Participants reported a significantly larger increase in their idea’s novelty with LITPIVOT ($\Delta = 2.29$ vs. $\Delta = 0.47$; $t(16) = 3.23$, $p_{\text{corr}} = .016$, $d_z = 0.78$), with a mean difference of 1.82 (95% CI [0.63, 3.01]).
- Participants reported a significantly larger increase in their idea’s utility with LITPIVOT compared to the baseline ($\Delta =$

0.53 vs. $\Delta = -0.76$; $t(16) = 2.81$, $p_{\text{corr}} = .037$, $d_z = 0.68$), with a mean difference of 1.29 (95% CI [0.32, 2.26]).

- Participants reported a larger increase in perceived feasibility, though the difference was not significant ($\Delta = 0.88$ vs. $\Delta = -0.35$; $t(16) = 2.11$, $p_{\text{corr}} = .152$, $d_z = 0.51$), with a mean difference of 1.23 (95% CI [-0.01, 2.47])¹².

Effect of LITPIVOT on groundedness of final ideas. To complement self-reported metrics¹³, we conducted a blinded expert evaluation of idea grounding. Ideas produced with LITPIVOT were rated as significantly more well-argued and grounded than those from the baseline. The median rating rose from 2 to 5 out of 7, and the mean doubled (Baseline $M=2.59$, $SD=1.33$; LITPIVOT $M=5.18$, $SD=1.85$; both $n=17$). A Wilcoxon signed-rank test confirmed higher ratings for LITPIVOT ($W=4.5$, $p<.001$), and a paired t -test showed the same pattern ($t(16)=5.33$, $p<.001$; mean difference = 2.59, 95% CI [1.56, 3.62]; $d_z=1.29$). See Table 1 for examples.

Using LITPIVOT to make literature-initiated pivots. We observed participants using LITPIVOT to develop their idea facets in two ways. Most participants ($n=13$) moved from one out-of-sync facet to another, not necessarily sequentially. For example, P4 first modified a problem statement, the first sentence of the idea. LITPIVOT flagged that the proposed evaluation, the last sentence of the idea, no longer matched the new problem statement. P4 edited their contribution last after it was flagged as out-of-sync. Alternatively, some participants edited their ideas sequentially, revising one facet multiple times before editing the next ($n=4$). For example, P5 generated three assessments for each facet before moving to the next facet segment. These two revision behaviors suggest that LITPIVOT supports different literature-grounded idea development workflows.

Participants rated LITPIVOT's suggestions as significantly higher than the outputs from the baseline in terms of helpfulness ($M = 6.24$ vs. $M = 4.12$; $t(16) = 3.96$, $p = .001$, $d_z = 0.96$) and trust ($M = 6.00$ vs. $M = 3.88$; $t(16) = 5.47$, $p < .001$, $d_z = 1.33$). There is some evidence that participants actively iterated on the suggestions rather than accepting LITPIVOT's outputs as is: across all sessions, 50% of the assessments were incorporated as written, 45.6% were partially incorporated with edits, and 4.4% were not incorporated.

6 Study 2: Open-ended qualitative study

To observe how researchers use LITPIVOT to develop their own ideas, we conducted a qualitative study. Five researchers who had participated in the user study in Section 5 were recruited. Before the session, participants shared a current idea and any relevant literature they had, which we pre-populated into the system. During a 30-minute session, they used LITPIVOT to develop the idea.

LITPIVOT was used across the research cycle. Researchers used LITPIVOT for vetting the validity of new ideas against relevant literature ($n = 2$) and framing mature ideas ($n = 3$). Most participants ($n = 4$) stated they would reuse the system for research writing. As P12 stated, “this is super useful...in the early stages of when I don't

know about a research idea and my idea is not fully shaped...and when I do have a fully formed research idea and I want to...make sure I'm not copying someone's existing research idea.”

LITPIVOT helps researchers relate their ideas to literature. All participants used facet-based clusters to search for literature. P3 stated, “it's way easier to navigate to a potential[ly relevant paper] when everything's grouped by topics.” Notably, participants selected different literature when iterating on different facets, highlighting the utility of dynamic literature curation. P2, for instance, initially believed their idea contributed a result specific to fine-tuning. However, after identifying a contribution cluster discussing reinforcement learning (RL) approaches, they reframed their contribution to generalize across both fine-tuning and RL. P2 reflected: “[LITPIVOT] could be useful for finding similarities between papers...especially in this case where this concept from one area...has parallels in another.”

Participants ($n = 4$) used assessments to evaluate their ideas' potential contributions, leading to pivots in how they framed their contribution. For example, P3 began with an early-stage idea focused on question generation in a specific high-stakes domain. After generating three assessments, each with a different set of literature, they noted that this initial contribution was too incremental. P3 then pivoted to defining and evaluating “situational appropriateness” of question-generation across multiple domains. By the end of the session, they articulated this new contribution, a taxonomy of QA constraints across domains. During the session, they saved seven papers for future reference, incorporated five LITPIVOT-suggested rewrites to their idea, and copied two evidence excerpts included in LITPIVOT's output to the idea editor.

7 Discussion

7.0.1 How LITPIVOT's features supported idea development. Our findings suggest that two design choices of LITPIVOT contributed to its observed effects. The first is dynamic literature curation. Participants selected significantly more unique papers with LITPIVOT (median 7 vs. 3), and the user study revealed two curation behaviors this enabled: selective pruning ($n=11$), where participants narrowed starred clusters in response to gaps the system surfaced (e.g., P1's pivot to multi-visit medical notes), and expansion ($n=6$), where participants grew their literature set across successive assessments (e.g., P9's four assessments over 2–6 papers).

The second is that LITPIVOT's critiques are specific enough to act on while leaving the decision with the researcher. LITPIVOT articulates how a facet relates to specific selected papers and proposes a revision, and participants rated these outputs as significantly more helpful ($M = 6.24$ vs. 4.12) and more trustworthy ($M=6.00$ vs. 3.88) than the baseline's. Researchers also engaged with these assessments: 50% of assessments were incorporated as written, 45.6% were edited before incorporation, and 4.4% were rejected. This pattern of active mediation, rather than passive acceptance, echoes insights from our formative study (P3: wanted support that “prompt[s] me to think...rather than thinking for me”).

7.0.2 Principles for broader applications. Existing tools for knowledge-intensive work like clinical reasoning or legal argumentation assume the corpus is static: it can be queried, but it does not change.

¹²We hypothesize that feasibility was unchanged because participants were assigned ideas in this study and, though they were assigned ideas based on their research areas, they might not have been able to effectively assess feasibility.

¹³Self-reported metrics are common in HCI ideation research [20, 39, 47], following established Creativity Support Tool evaluation practices [7, 9].

This assumption shapes more than retrieval interfaces. Search systems rank results by relevance to a query, not by their relationship to the argument a user is developing. Our results suggest a different model, in which the corpus reorganizes as the idea changes and the idea is revised in response to what the corpus surfaces. This model carries two design implications. First, retrieval could be idea-sensitive rather than query-sensitive, modeling the current state of the user’s developing argument. Second, feedback could be relational: tools could articulate how a retrieved document bears on the specific claim the user is making. The model could also extend beyond ideation. Expository writing, as defined by Shen et al. [43], produces artifacts that both summarize existing knowledge and generate new insight: authors must understand source texts while revising the artifact they produce, which requires pivots that recalibrate the evidence space and re-contextualize the artifact. Tools could support evidence-initiated pivots broadly, and their effects in these high-stakes settings are a promising future direction.

Adapting LITPIVOT’s mechanisms to other fields might require some adaptation. The facet vocabulary (problem, contribution, evaluation) can be used to assess and develop empirical HCI and NLP research. It might need restructuring for fields with different argument structures. The bipartite graph novelty check assumes a literature that explicitly states limitations and future work, a convention that might vary across disciplines and paper types.

7.0.3 Ethical considerations. LITPIVOT could encourage the development of a research idea that diverges from one the researcher would have naturally come to. For example, the facet structure asserts that an idea consists of a problem, contribution, and evaluation; the bipartite graph asserts that novelty comes from addressing an unaddressed limitation; and LLM-based extraction and clustering might encode further assumptions about the most influential literature relating to the idea. On one hand, these mechanisms set a lower bound on novelty: when a researcher’s articulated contribution is too close to prior work, LITPIVOT pushes the researcher to differentiate it, and it does not reject ideas for being unconventional. However, the same grounding that strengthens an idea could, if followed uncritically, pull an idea closer to existing literature than the researcher intends, or crowd out early, unformed directions that do not yet map cleanly onto articulated limitations. Paradigm-shifting ideas, in particular, may not correspond to any limitation stated in prior work. Our study offers a partial safeguard signal; participants edited or rejected half of LITPIVOT’s suggestions rather than accepting them without any edits. This practice of editing LITPIVOT-generated suggestions depends on researcher diligence. Designers of similar systems should make these embedded values legible to users, and future designs could offer modes with varying degrees of literature-boundedness suited to different phases of ideation (Section 7.2).

7.1 Limitations

Our evaluation focused on HCI and NLP researchers performing solo, time-boxed tasks largely at a specific stage of the ideation process. While our formative study (Section 3) indicates that ideas can evolve within short time frames, the generalizability of our findings could be limited. LITPIVOT’s utility is contingent on the

availability of a robust and accessible literature corpus; its effect may be less apparent when this is not the case.

LITPIVOT assesses two dimensions of an idea: how it is grounded in prior work and how it is differentiated from it. Other dimensions (e.g. the significance of a problem, an idea’s interestingness, the feasibility of executing it) are left to the researcher, and our evaluation does not measure them. Our formalization of idea differentiation is also one operationalization among many: the bipartite graph treats novelty as coverage of limitations that prior work has left unaddressed, which does not capture novelty that arises from reframing a problem or from work whose limitations are never articulated. LITPIVOT encodes specific values about what makes an idea good, such as grounding and structural coherence, that may not hold across disciplines. Because LITPIVOT does not execute code or run experiments, claims about idea feasibility cannot be verified, and the longitudinal effects of the system on long-term research development remain to be studied.

7.2 Future Work

Future work could explore the effects of different literature curation strategies on idea development. LITPIVOT lets researchers control which literature they evaluate against, in contrast to systems like Scideator that use automated methods [40]. Automated literature curation can surface long-tail papers a human might miss, but manual workflows might result in deeper understanding of the literature. Our literature clustering aims to bridge this gap, pointing users to different parts of the literature (framing, method, evaluation, datasets). Future work could further support human-AI collaboration through mixed-initiative search, clustering, and vetting [14]. Additionally, as richer agentic research workflows mature [14], comparing LITPIVOT against them, and layering pivot support onto them, could be a useful next step. Longitudinal studies could also track downstream outcomes such as significance and impact, and future designs could let researchers adjust the degree of literature-boundedness across phases of ideation.

Finally, dynamically re-situating evolving text against its sources could extend to other research tasks. For example, a system could provide dynamic feedback that surfaces related research threads to what the user is currently writing. Similar ideas could potentially be applied to update deep research agent reports. This includes tracing citations of existing references to check for new relevant literature published since the report was generated, situating the report within the new documents to generate updated suggestions.

8 Conclusion

We present LITPIVOT, an ideation system that helps researchers iteratively develop a research idea and explore relevant literature. LITPIVOT supports this process through *literature-initiated pivots*: moments when engaging with relevant literature prompts a researcher to revise the idea’s framing, and where that revision changes which literature is relevant. This coupling is driven by reorganizing the literature based on changes to the idea and surfacing whether an idea follows from, departs from, or is already addressed by the literature. Our studies show that ideas produced with LITPIVOT were rated by experts as better grounded in the literature, were perceived by

researchers as more novel and useful, and that researchers developed a stronger understanding of the literature space. We argue this points to a broader design principle: a knowledge corpus could be an active participant in shaping new ideas.

Acknowledgments

We thank our formative and user study participants for their time. We also thank Dan Weld and Doug Downey for discussions on project framing and the system prototypes. Hita thanks the bright and kind Ai2 intern cohort, especially Yapei Chang, Nishant Balepur, Alexis Ross, Federica Bologna, Xinran Zhao, Peiling Jiang, Anej Svete, Akhila Yerukola, Amanda Bertsch, and Michael Nounkhovitch, for making the summer in Seattle fun and memorable.

References

- [1] Ai2. Retrieved August 22, 2025. *Asta PaperFinder*. <https://allenai.org/blog/paper-finder>
- [2] Philip G Altbach and Hans De Wit. 2018. Too much academic research is being published. *International Higher Education* 96, 2 (2018).
- [3] Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A. Hearst, Andrew Head, and Kyle Lo. 2022. Paper Plain: Making Medical Research Papers Approachable to Healthcare Consumers with Natural Language Processing. *ACM Transactions on Computer-Human Interaction* 30 (2022), 1–38. <https://api.semanticscholar.org/CorpusID:247187606>
- [4] Ann Blandford, Dominic Furniss, and Stephann Makri. 2016. *Qualitative HCI Research: Going Behind the Scenes*. Morgan & Claypool Publishers.
- [5] Lutz Bornmann, Rüdiger Mutz, and Hans-Dieter Daniel. 2010. A reliability-generalization study of journal peer reviews: A multilevel meta-analysis of inter-rater reliability and its determinants. *PLoS one* 5, 12 (2010), e14331.
- [6] Kevin J Boudreau, Eva C Guinan, Karim R Lakhani, and Christoph Riedl. 2016. Looking across and looking beyond the knowledge frontier: Intellectual distance, novelty, and resource allocation in science. *Management Science* 62, 10 (2016), 2765–2783.
- [7] Erin A Carroll and Celine Latulipe. 2012. Triangulating the personal creative experience: self-report, external judgments, and physiology. In *Proceedings of Graphics Interface 2012*. 53–60.
- [8] 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.
- [9] 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.
- [10] Hai Dang, Chelse Swoopes, Daniel Buschek, and Elena L Glassman. 2025. CorpusStudio: Surfacing Emergent Patterns In A Corpus Of Prior Work While Writing. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–19.
- [11] Susan B Davidson and Juliana Freire. 2008. Provenance and scientific workflows: challenges and opportunities. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. 1345–1350.
- [12] Oscar Diaz and Jeremias P Contell. 2022. Developing research questions in conversation with the literature: operationalization & tool support. *Empirical Software Engineering* 27, 7 (2022), 174.
- [13] Timothy J Ellis and Yair Levy. 2008. Framework of problem-based research: A guide for novice researchers on the development of a research-worthy problem. *Informing Science* 11 (2008), 17.
- [14] KJ Feng, Kevin Pu, Matt Latzke, Tal August, Pao Siangliulue, Jonathan Bragg, Daniel S Weld, Amy X Zhang, and Joseph Chee Chang. 2026. Cocoa: Co-planning and co-execution with ai agents. *Proceedings of the ACM on Human-Computer Interaction* (2026).
- [15] Deborah Finfgeld-Connett and E Diane Johnson. 2013. Literature search strategies for conducting knowledge-building and theory-generating qualitative systematic reviews. *Journal of advanced nursing* 69, 1 (2013), 194–204.
- [16] Ronald A Finke. 1996. Imagery, creativity, and emergent structure. *Consciousness and cognition* 5, 3 (1996), 381–393.
- [17] Ronald A Finke, Thomas B Ward, and Steven M Smith. 1996. *Creative cognition: Theory, research, and applications*. MIT press.
- [18] Sarah Rose Fitzgerald. 2017. *Information seeking of scholars in the field of higher education*. Michigan State University.
- [19] Raymond Fok, Hita Kambhamettu, Luca Soldaini, Jonathan Bragg, Kyle Lo, Marti Hearst, Andrew Head, and Daniel S Weld. 2023. Scim: Intelligent skimming support for scientific papers. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 476–490.
- [20] Katy Ilonka Gero, Vivian Liu, and Lydia Chilton. 2022. Sparks: Inspiration for science writing using language models. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference*. 1002–1019.
- [21] Darcy Haag Granello. 2001. Promoting cognitive complexity in graduate written work: Using Bloom’s taxonomy as a pedagogical tool to improve literature reviews. *Counselor Education and Supervision* 40, 4 (2001), 292–307.
- [22] Ziwei Gu, Joyce Zhou, Ning-Er Lei, Jonathan K Kummerfeld, Mahmood Jasim, Narges Mahyar, and Elena L Glassman. 2025. AbstractExplorer: Leveraging Structure-Mapping Theory to Enhance Comparative Close Reading at Scale. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*. 1–25.
- [23] Armand Hatchuel, Benoit Weil, et al. 2003. A new approach of innovative Design: an introduction to CK theory.. In *DS 31: Proceedings of ICED 03, the 14th International Conference on Engineering Design, Stockholm*. 109–110.
- [24] Andrew Head, Kyle Lo, Dongyeop Kang, Raymond Fok, Sam Skjonsberg, Daniel S Weld, and Marti A Hearst. 2021. Augmenting scientific papers with just-in-time, position-sensitive definitions of terms and symbols. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–18.
- [25] Dongwei Jiang, Marcio Fonseca, and Shay B Cohen. 2024. Leanreasoner: Boosting complex logical reasoning with lean. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 7497–7510.
- [26] Hyeonsu Kang, Joseph Chee Chang, Yongsung Kim, and Aniket Kittur. 2022. Threddy: An interactive system for personalized thread-based exploration and organization of scientific literature. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–15.
- [27] 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.
- [28] Hyeonsu B Kang, Tongshuang Wu, Joseph Chee Chang, and Aniket Kittur. 2023. Synergi: A mixed-initiative system for scholarly synthesis and sensemaking. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–19.
- [29] David Klahr and Kevin Dunbar. 1988. Dual space search during scientific reasoning. *Cognitive science* 12, 1 (1988), 1–48.
- [30] Jeffrey W Knopf. 2006. Doing a literature review. *PS: Political Science & Politics* 39, 1 (2006), 127–132.
- [31] Christine P Lee, David Porfirio, Xinyu Jessica Wang, Kevin Chenkai Zhao, and Bilge Mutlu. 2025. Veriplan: Integrating formal verification and llms into end-user planning. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–19.
- [32] Yoonjoo Lee, Hyeonsu B Kang, Matt Latzke, Juho Kim, Jonathan Bragg, Joseph Chee Chang, and Pao Siangliulue. 2024. Paperweaver: Enriching topical paper alerts by contextualizing recommended papers with user-collected papers. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–19.
- [33] Q Vera Liao, Hariharan Subramonyam, Jennifer Wang, and Jennifer Wortman Vaughan. 2023. Designing understanding: Information needs for model transparency to support design ideation for AI-powered user experience. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–21.
- [34] Weng Marc Lim, Satish Kumar, and Faizan Ali. 2022. Advancing knowledge through literature reviews: ‘what’, ‘why’, and ‘how to contribute’. *The Service Industries Journal* 42, 7–8 (2022), 481–513.
- [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 2024 CHI Conference on Human Factors in Computing Systems*. 1–25.
- [36] OpenAI. 2026. Introducing Prism. <https://openai.com/index/introducing-prism/>. Accessed: 2026-02-04.
- [37] Srishti Palani, Aakanksha Naik, Doug Downey, Amy X Zhang, Jonathan Bragg, and Joseph Chee Chang. 2023. Relatedly: Scaffolding literature reviews with existing related work sections. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.
- [38] Elizabeth L Pier, Markus Brauer, Anna Filut, Anna Kaatz, Joshua Raclaw, Mitchell J Nathan, Cecilia E Ford, and Molly Carnes. 2018. Low agreement among reviewers evaluating the same NIH grant applications. *Proceedings of the National Academy of Sciences* 115, 12 (2018), 2952–2957.
- [39] Kevin Pu, KJ Kevin Feng, Tovi Grossman, Tom Hope, Bhavana Dalvi Mishra, Matt Latzke, Jonathan Bragg, Joseph Chee Chang, and Pao Siangliulue. 2025. Ideasynt: Iterative research idea development through evolving and composing idea facets with literature-grounded feedback. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–31.
- [40] Marissa Radensky, Simra Shahid, Raymond Fok, Pao Siangliulue, Tom Hope, and Daniel S Weld. 2024. Scideator: Human-llm scientific idea generation grounded in research-paper facet recombination. *arXiv preprint arXiv:2409.14634* (2024).
- [41] Orit Shaer, Angelora Cooper, Osnat Mokryn, Andrew L Kun, and Hagit Ben Shoshan. 2024. AI-Augmented Brainwriting: Investigating the use of LLMs

- in group ideation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–17.
- [42] Nihar B Shah, Behzad Tabibian, Krikamol Muandet, Isabelle Guyon, and Ulrike Von Luxburg. 2018. Design and analysis of the NIPS 2016 review process. *Journal of Machine Learning Research* 19, 49 (2018), 1–34.
- [43] Zejiang Shen, Tal August, Pao Siangliulue, Kyle Lo, Jonathan Bragg, Jeff Hammerbacher, Doug Downey, Joseph Chee Chang, and David Sontag. 2023. Beyond summarization: Designing AI support for real-world expository writing tasks. In *Proceedings of the 2nd Workshop on Intelligent and Interactive Writing Assistants (In2Writing)*.
- [44] Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2025. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. In *International Conference on Learning Representations*, Vol. 2025. 94003–94092.
- [45] Amanpreet Singh, Joseph Chee Chang, Dany Haddad, Aakanksha Naik, Jena D Hwang, Rodney Kinney, Daniel S Weld, Doug Downey, and Sergey Feldman. 2025. Ai2 scholar qa: Organized literature synthesis with attribution. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*. 513–523.
- [46] 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 2024 CHI Conference on Human Factors in Computing Systems*. 1–26.
- [47] Sangho Suh, Bryan Min, Srishti Palani, and Haijun Xia. 2023. Sensecape: Enabling multilevel exploration and sensemaking with large language models. In *Proceedings of the 36th annual ACM symposium on user interface software and technology*. 1–18.
- [48] Rosni Vasu, Chandrayee Basu, Bhavana Dalvi Mishra, Cristina Sarasua, Peter Clark, and Abraham Bernstein. 2025. HypER: Literature-grounded Hypothesis Generation and Distillation with Provenance. *The Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2025).
- [49] Dashun Wang, Chaoming Song, and Albert-László Barabási. 2013. Quantifying long-term scientific impact. *Science* 342, 6154 (2013), 127–132.

A Guidelines for rating idea artifacts

- Focus on whether the idea description leveraged prior work to argue for its novelty and explain its relation to relevant literature. For example, arguments and methods are stronger if they are backed by relevant prior work to show importance or promise.
- The idea description is well reasoned (i.e., no jump of logic); Descriptions of problem, contribution, and evaluation should be relevant and coherent.
- Do not consider personal research preferences or the “interestingness” of the idea.
- Assume participants described prior work referenced in their idea descriptions accurately.
- Execution feasibility is not part of the evaluation, as it may be highly dependent on the amount of resources the participants have access to.
- Rate the ideas holistically, but -1 if the idea description is missing evaluation. (All ideas have background, problem, and contribution statements, but a few were missing evaluation.)
- Two example ideas that were agreed upon received a rating of 7 and 1, respectively.

B Prompts

B.0.1 Prompt to segment idea. A researcher wants help in refining their research idea into a research proposal. To do this, you are tasked with figuring out what parts of their current research idea relates to the following important facets of a research proposal: problem, contribution, and evaluation.

The research problem is any text that describes the core problem a researcher is trying to solve. It answers the following questions: 1. What is the problem this researcher is trying to solve? 2. Why bother solving this problem?

The research contribution is any text that describes the proposed solution to this problem. It answers the following questions: 1. How do current solutions fail at solving the problem? 2. What would a solution to this problem look like? This is the core contribution of the researcher, the core questions the researcher wants to explore in the project. The core contribution might be a proposed system or it might be a study or benchmark that the researchers want to run, or a dataset they are introducing. It is the proposed way that a researcher will answer their research questions. It is what they are developing in this project.

The research evaluation is any text that describes how a researcher will evaluate their contribution. It answers the following questions: 1. How would I know I've solved this problem? 2. How do I plan on solving this problem?

IMPORTANT: The researcher might not have each of these facets, so only classify a section of the text as a particular category if it is the case, and leave the other categories blank.

IMPORTANT: One segment of text should only have one category.

IMPORTANT: Each segment of text should be a complete phrase or sentence.

IMPORTANT: There should be no overlap between faceted segments.

IMPORTANT: Each segment should be matched and returned verbatim as the input, matching case.

IMPORTANT: If any text states Add evaluation here it should be classified as evaluation.

IMPORTANT: If any text states Add problem here it should be classified as problem.

IMPORTANT: If any text states Add contribution here it should be classified as contribution.

Here is the researcher's idea: IDEA

B.0.2 Prompt to extract the contribution facet from a paper. You are helping a researcher understand the core contributions in the provided paper. Extract and return exact quoted excerpts from this paper that represent the core contributions and novelty of this paper. The return text will be a list of exact quotes from the paper. Limit to 3 excerpts. Each excerpt should be 1–2 sentences long.

This is the paper: TEXT

Note: We used a similar template for extracting other facets (problems, limitations, evaluations, methods, results, future work).

B.0.3 Prompt to evaluate a problem facet. You will be executing two tasks: first, you will evaluate the researcher's problem statement of their idea and second, you will provide literature-grounded, concrete suggestions for how they can improve their problem statement.

This is their entire idea: IDEA

This is the part they need to refine: IDEA_SEGMENT

IMPORTANT: This is additional context that the researcher wants you to know: ADDITIONAL_CONTEXT. If there is no context, ignore this.

To help the researcher refine this part of their idea, here is how current, successful papers describe the problem statements that they are aiming to solve:

CONTEXT

For your evaluation, you want to make sure that the researcher is answering the following broad questions in their problem statement: 1. What is the problem the researcher is trying to solve? and 2. Why bother solving this problem?

More specifically, with the current formulation of the researcher's problem, can they cite any of the papers above to support their problem statement? Using the excerpts from the papers provided as examples, evaluate whether the problem statement specific enough and if it is an actual problem, and if there is enough support to know that this is the right problem formulation for the researcher to work on. Your evaluation should be no more than 4 sentences, and should include excerpts from and references to as many related papers as possible provide the best assessment of the paper. The references of papers should include the first few words of the paper title and should be of the following form: "<Paper corpusId="268248445">PaperWeaver</Paper>".

After the evaluation, you will provide 1-3 concrete suggestions to the researcher on how they can improve their problem formulation, if applicable. The suggested rewrite should be something that the researcher can directly replace their current idea segment with. More formally, rewrite the researcher's idea so that the questions below are clearly answered. Then, include excerpts from relevant literature that could be used to improve the problem formulation. These excerpts should be direct quotes of related problem statements or supporting evidence. If at any point, any of the following questions are currently inadequately answered by the researcher, then you should use the literature provided above and your knowledge of the researcher's idea to ensure that your suggested rewrite answers all of these questions. Here are the questions:

1. What is the problem? This should be of the form "We currently can't do Y" or "We want to do Y"
2. What is the specific problem we are trying to solve?
3. How do we know this is an actual problem?
4. Why should we bother solving this problem?

IMPORTANT: Here is an example of the output format:

"evaluation" : "The current idea omits any concrete metrics or procedures, so it cannot yet demonstrate that the produced citation-hop dataset is correct or useful for verifying biomedical claims. Prior work such as "<Paper corpusId="268248445">Piecing It All Together</Paper>" and "<Paper corpusId="123456789">Retrieval-Augmented Scientific Claim Verification</Paper>" shows that both expert-led quality audits and baseline system benchmarks are needed to answer "Have we solved the problem?". Without similar expert agreement scores and model-based baselines, success cannot be measured.",

"suggestions" :

"We will sample automatically extracted citation chains and confirm that every cited paper is extant and accurately quoted. We will report the hallucination rate (false or malformed citations) and compare it to baseline large-language-model generation without our pipeline." :

"278962365" : "<Quote corpusId=278962365>In a sample of 10 claims containing 211 citations, every work cited by Valsci was confirmed to be extant...</Quote>",

...

,

...

IMPORTANT: Your output should strictly be of the following form:

"evaluation" : "an evaluation of the idea segment",

"suggestions" :

"suggested rewrite 1" :

"corpusId1" : "<Quote corpusId=278962365>quote 1</Quote>",

...

,

...

IMPORTANT: Your rewrites should only be rewrites of this problem statement, NOT the contribution or evaluation. Focus only on making this problem statement better.

IMPORTANT: The response should be entirely self-contained, meaning any quotes from prior literature should be directly quoted. Additionally the reader should understand the answer without knowing the question.

IMPORTANT: All references of papers should include the first few words of the paper title and should be of the following form: "<Paper corpusId="268248445">PaperWeaver</Paper>".

IMPORTANT: Your assessment should be in two sections. The first should be an evaluation, which should be 3 sentences long. And the second should be dictionary of idea suggestions, which should be a direct rewrite of the selected segment along with a quote from related literature and a short explanation to support the suggestion.

IMPORTANT: The response should be as useful as possible to the researcher, so define any jargon by inferring the researcher's experience based on the idea they wrote.

IMPORTANT: The suggestions should be based one what you think the researcher would be most interested in given their original idea.

IMPORTANT: The structure of the output should be self-contained and not enumerated.

B.0.4 Prompt to construct a bipartite graph. You are helping a researcher assess their research idea. To do this, you will be helping the researcher understand what the contributions are of current literature and what future work and limitations have been identified by current literature. This will ultimately be used to help make the researcher's idea more concrete and novel.

Ultimately, the researcher wants to refine their research idea so they are addressing an exciting, existing limitation that has not been addressed yet or work on a project that has a lot of contributions which is one that the research community is really excited about addressing.

Given a list of limitations and future work described in current, related research papers and a list of contributions of these papers, you will note whether a contribution directly addresses a limitation or future work.

Decide the match between contribution and future work or limitation with deep care and wisdom, so that they are the most meaningful to the researcher who wants to understand what future work or limitations identified by papers related to their idea has been addressed by current contributions of these papers.

Here is a list of papers with their limitations and future works: LIMITATIONS

Here is a list of papers with their current contributions: CONTRIBUTIONS

Here is the researcher's idea: IDEA

IMPORTANT: Make sure ALL papers appear in your output. Do not miss any papers.

The output will be a JSON where the key is the name of the limitation or future work and the value is a dictionary where the key is "addressed_by" and the value is a list of paper titles and contributions that address the future work mapped to short, self-contained explanations. If no papers address

the future work or limitation, the "addressed_by" key should be mapped to an empty list.

Here is an example of the output structure:

```
"(paper title X) Limitation X":
  "addressed_by": [
    "(paper1 title) paper1 contribution" : "short explanation",
    "(paper2 title) paper2 contribution" : "short explanation"
  ],
"(paper title Y) Future work Y":
  "addressed_by": [
    "(paper3 title) paper3 contribution": "short explanation",
    "(paper4 title) paper4 contribution": "short explanation"
  ],
...
```

B.0.5 Prompt to evaluate a solution facet. You will be executing two tasks: you will evaluate the researcher's proposed solution in their research idea and you will provide literature-grounded, concrete suggestions for how they can improve their proposed solution/research contribution.

This is their entire idea: IDEA

This is the part they need to refine: IDEA_SEGMENT

IMPORTANT: This is additional context that the researcher wants you to know: ADDITIONAL_CONTEXT. If there is no context, ignore this.

To help the researcher refine this part of their idea, you will rely on the following information. This dictionary maps current limitations and future work in related literature to contributions of current literature.

LIM_FW_CONT_MATCHING

This is a list of all of the current contributions of the selected papers: CONTRIBUTIONS

For your evaluation, you want to make sure that the researcher is answering the following broad questions in their proposed solution: 1. How do current solutions fail at solving the problem the researcher has identified? 2. Does the researcher's solution better address the problem?

To formally evaluate the novelty of the proposed solution, you will do the following: The proposed solution is novel if it addresses future work or limitations in the set of literature that has not yet been addressed. Given this list of current limitations and future work described in related research papers and how current contributions of some papers address these limitations or future work, you are tasked with deciding if the researcher's proposed solution addresses a limitation or future work in this dictionary. If the proposed solution addressess a limitation or future work that has not been addressed by any other work, then it is novel. If the researcher's contribution matches an existing contribution, then it is not novel. Your evaluation should be no more than 4 sentences, and should include excerpts from and references to as many related papers as possible provide the best assessment of the paper. The references of papers should include the first few words of the paper title and should be of the following form: "<Paper corpusId="268248445">PaperWeaver</Paper>".

After the assessment, you will provide 1-3 concrete suggestions for how the current formulation of the researcher's solution/contribution can be more novel, if applicable. The suggested rewrite should be something that the researcher can directly replace their current idea segment with. Be specific, but don't make up any names or numbers such as intended results or system/benchmark/dataset names. More formally, rewrite the researcher's idea so that the

questions below are clearly answered. Then, include excerpts from relevant literature that could be used to improve the contribution formulation. In particular, quote the future work and limitations described in current work to affirm that the suggestion is novel, or quote any very related literature that supports the suggestion. If at any point, any of the following questions are currently inadequately answered by the researcher, then you should use the literature provided above and your knowledge of the researcher's idea to ensure that your suggested rewrite answers all of these questions. Here are the questions:

1. Does the solution accurately address the problem?
2. Will the solution succeed?
3. Is the solution novel?

IMPORTANT: Here is an example of the output format:

```
"evaluation" : "The current idea omits any concrete metrics
or procedures, so it cannot yet demonstrate that the produced
citation-hop dataset is correct or useful for verifying biomedical
claims. Prior work such as "<Paper corpusId="268248445">Piecing
It All Together</Paper>" and "<Paper corpusId="123456789">Retrieval-Augmented
Scientific Claim Verification</Paper>" shows that both expert-led
quality audits and baseline system benchmarks are needed to
answer "Have we solved the problem?". Without similar expert
agreement scores and model-based baselines, success cannot be
measured.",
"suggestions" :
```

```
"We will sample automatically extracted citation chains and
confirm that every cited paper is extant and accurately quoted.
We will report the hallucination rate (false or malformed
citations) and compare it to baseline large-language-model
generation without our pipeline." :
```

```
"278962365" : "<Quote corpusId=278962365>In a sample of 10
claims containing 211 citations, every work cited by Valsci
was confirmed to be extant...</Quote>",
```

```
...
,
...

```

IMPORTANT: Your output should strictly be of the following form:

```
"evaluation" : "an evaluation of the idea segment",
"suggestions" :
"suggested rewrite 1" :
"corpusId1" : "<Quote corpusId=278962365>quote 1</Quote>",
...
,
...

```

IMPORTANT: Your rewrites should only be rewrites of this contribution statement, NOT the problem or evaluation. Focus only on making this contribution statement better.

IMPORTANT: The response should be entirely self-contained, meaning any quotes from prior literature should be directly quoted. Additionally the reader should understand the answer without knowing the question.

IMPORTANT: All citations of papers should include the first few words of the paper title and should be of the following form: "<Paper corpusId="268248445">PaperWeaver</Paper>".

IMPORTANT: Your assessment should be in two sections. The first should be an evaluation, which should be 3 sentences long. And the second should be dictionary of idea suggestions, which should be a direct rewrite of the selected segment along with a quote from related literature and a short explanation to support the suggestion.

IMPORTANT: The response should be as useful as possible to the researcher, so define any jargon by inferring the researcher’s experience based on the idea they wrote.

IMPORTANT: The suggestions should be based on what you think the researcher would be most interested in given their original idea.

IMPORTANT: The structure of the output should be self-contained and not enumerated.

B.0.6 Prompt to evaluate an evaluation facet. You will be executing two tasks: you will evaluate the researcher’s proposed evaluation in their research idea and you will provide literature-grounded, concrete suggestions for how they can improve their proposed evaluation.

This is their entire idea: IDEA

This is the part they need to refine: IDEA_SEGMENT

IMPORTANT: This is additional context that the researcher wants you to know: ADDITIONAL_CONTEXT. If there is no context, ignore this.

For your evaluation, you want to make sure that the researcher is answering the following broad question in their proposed evaluation: How would they know they’ve solved the problem?

These are how evaluations of current, successful papers show how a solution solves a problem:

PROL_SOL_EVAL

To formally evaluate the proposed evaluation, you will do the following: given the problem and solution statement of the idea, the evaluation, and the method, does the evaluation and method correctly ensure that the solution is addressing the problem? Given a list of problem statements, solution statements, methods, and evaluations of related research papers, you will use these as few-shot examples to calibrate yourself to what a successful evaluation method would look like. Then you will decide whether the proposed evaluation will correctly evaluate whether the solution correctly addresses the problem. Your evaluation should be no more than 4 sentences, and should include excerpts from and references to as many related papers as possible provide the best assessment of the paper. The references of papers should include the first few words of the paper title and should be of the following form: “<Paper corpusId=“268248445”>PaperWeaver</Paper>“.

After the assessment, you will provide 1–3 concrete suggestions for how the current formulation of the researcher’s solution/contribution can be more novel, if applicable. The suggested rewrite should be something that the researcher can directly replace their current idea segment with. More formally, rewrite the researcher’s idea so that the questions below are clearly answered. Then, include excerpts from relevant literature that could be used to improve the problem formulation. These excerpts should be direct quotes of related evaluations or supporting evidence. If at any point, any of the following questions are currently inadequately answered by the researcher, then you should use the literature provided above and your knowledge of the researcher’s idea to ensure that your suggested rewrite answers all of these questions. Here are the questions:

1. Will the proposed evaluation ensure the solution accurately address the problem?

2. Will the evaluation succeed?

IMPORTANT: Your output should strictly be of the following form:

Evaluation: [3 sentences of evaluation, following citing paper formats]

Suggestion: full rewrite1 with concrete suggestion:
"evidence1": quoted evidence from a cited paper,
"paper1" : corpusId of paper1
"explanation1": explanation of why this is a good rewrite grounded in the evidence,

...

,

full rewrite2 with concrete suggestion:
"evidence2": quoted evidence from a cited paper,
"paper2" : corpusId of paper2
"explanation2": explanation of why this is a good rewrite grounded in the evidence,

...

...

IMPORTANT: Your rewrites should only be rewrites of this evaluation statement, NOT the contribution or problem. Focus only on making this evaluation statement better.

IMPORTANT: The response should be entirely self-contained, meaning any quotes from prior literature should be directly quoted. Additionally the reader should understand the answer without knowing the question.

IMPORTANT: Ground all your responses in the literature provided. All quotes from papers should be of the following form: “<Quote corpusId=“268248445”>PaperWeaver is good</Quote>“

IMPORTANT: All citations of papers should include the first few words of the paper title and should be of the following form: “<Paper corpusId=“268248445”>PaperWeaver</Paper>“

IMPORTANT: Your assessment should be in two sections. The first should be an evaluation, which should be 3 sentences long. And the second should be dictionary of idea suggestions, which should be a direct rewrite of the selected segment along with a quote from related literature and a short explanation to support the suggestion.

B.0.7 Prompt to identify affected facet segments. A researcher wants to edit part of their research idea. You are helping a researcher understand what other parts of their research idea they will need to edit.

Here are all of the segments of the research idea that could be edited: IDEA_SEGMENTS

Here is the idea that was just edited: IDEA_SEGMENT

Here is the assessment of that edit: NOVELTY_ASSESSMENT

You are tasked with identifying which segments should also be edited. Return only a list of exactly quoted segments that should be edited.

B.0.8 Prompt to identify missing segments. You are tasked with helping a researcher understand where they should add information about a particular facet of their research idea that is currently missing. You are given a researcher’s idea, and the following dictionary. The dictionary keys are three facets, problem, solution, and evaluation. For each of these three facets, the values are sentences in the original idea that describe this facet. If any of the values in the facet is empty, meaning there is no text yet describing this facet, then you will infer where text describing this facet should be placed. You will return the original seed idea text, with [FACET] in places where the particular missing facet should be included.

Here is the original idea: ORIGINAL_IDEA

Here is the dictionary of facets: FACETS

IMPORTANT: If there is already text that states [Add evaluation here], [Add problem here], [Add contribution here], or [Add solution here], assume that this facet is accounted for and not missing. So do not include that that facet is missing.

B.0.9 Prompt to identify relevant clusters. A researcher is refining a part of their research idea. To refine their research idea, they will be reading papers in the following clusters of papers. You will be tasked with identifying which paper clusters are most relevant to the part of the idea that the researcher is refining.

Here is the part of the idea that the researcher is editing: IDEA_SEGMENT

Here are the names of the paper clusters: CLUSTERS

IMPORTANT: Please return ONLY the exact cluster names, one per line, with no explanations or numbering. Select 1-3 clusters that most relate to the part of the idea.

Example output format:

Challenges in verifying health and medical claims
Citation integrity and provenance tracking

B.0.10 Prompt to cluster by the contribution facet. You will be helping a researcher with a research idea understand the clusters of current contributions in the research space that this idea is in so that they can understand what the main contribution patterns are that current projects make. Given a list of passages of current contributions from related research papers, you will come up with meaningful clusters to describe groups of contributions in this research area. You should make sure that every passage is included in a cluster. Choose the clusters with deep care and wisdom, so that they are the most meaningful to the researcher who wants to understand what the current contribution patterns are in papers related to their idea.

Here is a list of corpusId and contribution passages: CONTRIBUTIONS

Here is the seed idea: SEED

IMPORTANT: There are NUM_UNIQUE_CORPUS_IDS unique corpus IDs in the data below. Make sure ALL of them appear in your clustering output. Do not miss any corpus IDs.

The output will be a JSON where the key is the name of the cluster, typically a sentence that describes the contribution cluster, and the value is a list of just corpusIds that are in this cluster. Do not include the contribution passages in the output. Only include the corpusId.

Here is an example of the output structure:

```
"Novel algorithmic approaches and theoretical frameworks":
["corpusID1", "corpusID2"],
"Empirical evaluation and benchmark improvements": ["corpusID3"]
...
```

Note: We use a similar prompt structure to cluster by other facet types.

B.0.11 Prompt to generate complete assessments. You will be giving a full assessment of this researcher's project proposal. The proposal should answer these questions:

1. What is the problem? This should be of the form "We currently can't do Y" or "We want to do Y"
2. What is the specific problem we are trying to solve?
3. How do we know this is an actual problem?
4. Why should we bother solving this problem?
5. Does the solution accurately address the problem?
6. Will the solution succeed?
7. Is the solution novel?

8. Will the proposed evaluation ensure the solution accurately address the problem?

9. Will the evaluation succeed?

Here are previous assessments of parts of this idea.

PREVIOUS_ASSESSMENTS

Here is the proposal: IDEA

Your assessment should be no more than 5 sentences. Be thorough.

C Example of ideas from LITPIVOT and the baseline from the user study

Table 1: Side-by-side comparison of baseline ideas and their LitPivot counterparts. Diagonal entries are from the same participant.

	Baseline	LitPivot
Initial Idea 1: Current RAG QA systems are touted as more interpretable and explainable because they point to existing sources. But sometimes these sources aren't even relevant to the generation. We propose a tool that helps users engage more with source texts and better discern the quality of source texts.	Recent research has highlighted a critical gap in RAG QA systems where retrieved documents that appear relevant based on surface-level similarity, don't actually support the specific claims being made in the generation. While RAG systems theoretically provide transparency by surfacing source documents, the granularity of attribution is often too coarse (typically at the document or sentence level) to be truly useful for verification, as demonstrated in works on fine-grained attribution. The mismatch between retrieved sources and generated content not only undermines the system's interpretability claims but also potentially misleads users who trust the system's ability to ground its responses in reliable sources. We propose a tool that combines fine-grained attribution techniques with an interactive interface, where users can click on specific claims in the generated text to see the exact supporting quotes from source documents highlighted in context. The tool also employs a traffic light system to visually indicate the strength of attribution for each claim (green for direct quotes/strong support, yellow for partial/indirect support, red for unsupported claims), helping users quickly identify potential issues. We can evaluate our tool through a multi-faceted approach combining: (1) precision/recall metrics comparing system-identified attributions to human-annotated gold standards, (2) semantic similarity scores between generated claims and source text segments using established embedding models, (3) user studies measuring verification time and accuracy with and without the tool.	Current RAG QA systems are touted as more interpretable and explainable because they point to existing sources. However, these systems still tend to fail for certain tasks, e.g. they can hallucinate information, miss key references, and find wrong or unrelated sources. Existing literature focus on attribution (and how to use them for improving QA for LLMs, but not how to engage with these attributions, nor how to help users discern the quality of these attributions. Enabling users to do so empowers users to be more diligent in checking the answers and sources' quality and improves the trustworthiness and transparency of such RAG QA systems. We propose a tool that helps users engage more with source texts and better discern the quality of source texts. We introduce a voice- and pointer-driven RAG interface in which (i) users can issue spoken follow-ups such as show me the sentence that supports that claim or read the most reliable paragraph aloud, and (ii) every cited span is accompanied by an automatically generated reliability badge that aggregates contradiction, sufficiency and source-authority scores computed with CiteEval-AUTO and AutoAIS. The system also supports LAQuer-style localized queries so that users can drill down into sub-sentences, with the interface highlighting the exact supporting tokens. Our tool provides 1) a novel way for users to interact with these sources through multimodal interaction paradigms including source navigation (via attribution) and voice interaction, rather than treating RAG as a blackbox component in the QA pipeline, and 2) a novel way for users to easily verify and discern source quality. We focus on textual content as the source, leaving multimodal content (images, videos, PDFs, voices) as future work. A controlled user study will measure whether the badge + drill-down workflow improves verification speed and accuracy compared to the baseline clickable-citation UI used in TruthReader.
Initial Idea 2: Existing methods for verifying scientific claims across research papers typically lack systematic checks for logical consistency. We want to develop an interactive, two-way "semi-formal proof navigation" system that 1) translates natural-language claims from research papers into structured logical propositions using various formal logic representations for flexibility, 2) enables visualizing and interacting with the claims and their logical relationships. The system automatically highlights parts with problematic claims and suggests remediations, and enables the user to modify visualized logical claims where the results will be updated in the original content, and vice versa.	Existing methods for verifying scientific claims across research papers typically lack systematic checks for logical consistency. We want to develop an interactive, two-way "semi-formal proof navigation" system that 1) translates natural-language claims from research papers into structured logical propositions using various formal logic representations for flexibility, 2) enables visualizing and interacting with the claims and their logical relationships. The system automatically highlights parts with problematic claims and suggests remediations, and enables the user to modify visualized logical claims where the results will be updated in the original content, and vice versa.	Scientists currently lack automated tools that can detect when claims drawn from different papers form a logically inconsistent chain of reasoning. Existing scientific fact-checking resources focus on single-claim, single-paper links for example, SCIFACT targets abstract-level support or refutation, and CLAIM-BENCH evaluates claim-evidence reasoning within one paper leaving cross-paper logical consistency unchecked. Without such cross-document verification, meta-analyses and policy decisions risk propagating contradictions or unsupported citation leaps in an ever-growing literature, making a principled consistency checker both timely and necessary. We will build an interactive proof-navigation environment that (i) harvests claims and supporting sentences from multiple research papers, (ii) auto-formalises each statement into FOL while attaching citation metadata, (iii) dynamically chooses the most reliable symbolic solver per claim type, (iv) stitches the resulting formulas into a cross-paper proof graph, and (v) lets users step through, edit, and re-verify any node to expose contradictions or unsupported leaps.

D Affected Facets feature

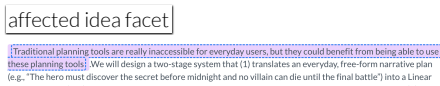


Figure 5: When a change to one idea facet might require editing another idea facet, LITPIVOT highlights the corresponding affected idea facet in the text editor. Here, the problem statement facet is outlined.

E Baseline

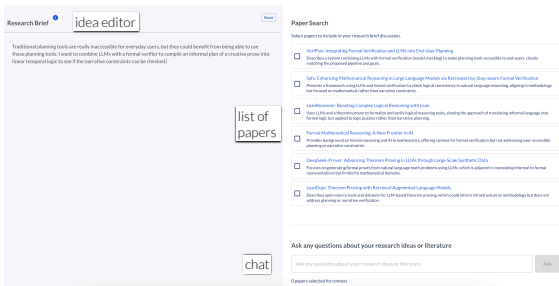


Figure 6: Baseline interface. The left pane contains an idea editor where researchers can write out their research idea. The right pane shows a list of retrieved papers. At the bottom, a chat box allows researchers to ask questions about their idea or the listed literature.

F Design Probe

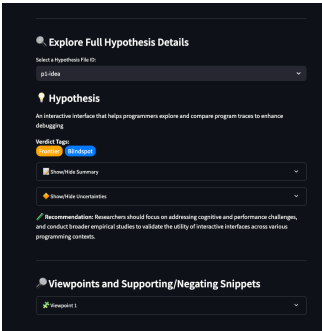


Figure 7: Design probe used in the formative study. This is an LLM-powered prototype. The research hypothesis is displayed at the top. Verdict tags are assigned below. There are recommendations, supporting and negating viewpoints that use related literature to ground feedback.