# Supporting Novice Researchers to Write Literature Review using Language Models

Kiroong Choe
Seoul National University
Seoul, Republic of Korea
krchoe@hcil.snu.ac.kr

Seokhyeon Park
Seoul National University
Seoul, Republic of Korea
shpark@hcil.snu.ac.kr

Seokweon Jung
Seoul National University
Seoul, Republic of Korea
swjung@hcil.snu.ac.kr

Hyeok Kim
Northwestern University
Evanston, IL, U.S.A.
hyeok@northwestern.edu

Ji Won Yang
Seoul National University
Seoul, Republic of Korea
yangji921@snu.ac.kr

Hwajung Hong
KAIST
Daejeon, Republic of Korea
hwajung@kaist.ac.kr

Jinwook Seo
Seoul National University
Seoul, Republic of Korea
jseo@snu.ac.kr

## ABSTRACT

A literature review requires more than summarization. While language model-based services and systems increasingly assist in analyzing accurate content in papers, their role in supporting novice researchers to develop independent perspectives on literature remains underexplored. We propose the design and evaluation of a system that supports the writing of argumentative narratives from literature. Based on the barriers faced by novice researchers before, during, and after writing, identified through semi-structured interviews, we propose a prototype of a language-model-assisted academic writing system that scaffolds the literature review writing process. A series of workshop studies revealed that novice researchers found the support valuable as they could initiate writing, co-create satisfying contents, and develop agency and confidence through a long-term dynamic partnership with the AI.

## CCS CONCEPTS

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

## KEYWORDS

literature review, novice researcher, human-AI collaboration

**ACM Reference Format:**
Kiroong Choe, Seokhyeon Park, Seokweon Jung, Hyeok Kim, Ji Won Yang, Hwajung Hong, and Jinwook Seo. 2024. Supporting Novice Researchers to Write Literature Review using Language Models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3613905.3650787

## 1 INTRODUCTION

Writing is an essential component of literature review. While often viewed as the final output, such as the related work section in a paper, writing is also a critical procedural aspect that works in conjunction with searching, reading, and organizing, actively contributing to the understanding of the literature [43]. Novice researchers—individuals entering the research field with no or little research experience—may have barriers to writing literature reviews due to a lack of confidence (i.e., *"I'm not a good writer"* [10]) and misconceptions about writing (e.g., *"I should fully understand literature before starting to write"* [29]). While getting feedback from an advisor or colleagues is a common remedy, novice researchers might not be able to access those feedback due to time constraints or the absence of mentors [31].

The recent advancements in large language models (LLMs), such as ChatGPT [48], have led to the development of various applications that provide intelligent assistance for academic tasks, including searching for literature, analyzing research papers, and enhancing academic writing [41]. Although these tools provide complementary guidance to novice researchers, they primarily focus on content delivery rather than helping researchers develop their own perspectives on the literature. Effective literature reviews require creative integration of previous research to pinpoint gaps and suggest future directions [18]. While numerous applications of LLMs have been proposed in creative domains [12, 17, 19, 32, 38], there is still a gap in understanding how to incorporate these models for argumentative (or expository) writing, where researchers generate knowledge based on evidence [45].

We contribute an empirical study in which novice researchers conducted a literature review of their own research topic with the support of LLMs. We first derived the challenges and needs of novice researchers for writing a literature review through a formative study with 11 junior graduate students. Then, we designed a prototype system, LitWeaver, that utilizes LLM-driven support to assist with the complete iteration cycle of writing a literature review, including paper reviewing, topic organization, and manuscript writing. Finally, we conducted single-session (n=9) and long-term (n=3) workshop studies, where novice HCI researchers used the LitWeaver to carry out literature reviews. Our results show that the scaffolded workflow and LLM-driven assistance positively influenced participants in initiating writing, organizing research narratives, and refining clarity and comprehensiveness. Furthermore, long-term participants developed dynamic partnerships with LLM, gaining agency and confidence and receiving emotional support from their interaction with LLM.

*CHI EA '24, May 11–16, 2024, Honolulu, HI, USA*

ACM ISBN 979-8-4007-0331-7/24/05
https://doi.org/10.1145/3613905.3650787

We emphasize that our research goal is not to automate literature reviews but rather to encourage novice researchers to actively engage with literature, enabling them to develop *"original perspectives"* from literature. Our study reveals new opportunities for long-term interactions between novice researchers and LLM to foster researchers' development of agency and confidence in writing literature reviews. We discuss the implications for designing systems to support this growth.

## 2 RELATED WORK

### 2.1 Process-driven Education of Literature Review

Novice researchers are expected to develop certain skills in reviewing literature as they advance in academia, including critical evaluation, reconciliation, and synthesis of knowledge to incorporate their original perspectives into the field [22]. This is a complex knowledge process that requires extensive training, yet many novices (and even experienced students [4]) hold misconceptions about literature reviews, often viewing them as mechanical summaries [16]. Boote strongly criticizes that the common literature review guidelines over-simplify the process as *"identifying key terms, locating papers, and reading and organizing them,"* leading to a perception that writing a literature review is *"no more complicated than writing a high school term paper"* [5].

Paradoxically, to effectively learn literature review techniques, novice researchers must actively engage in the literature review process despite potential barriers. Conducting a literature review involves a series of recursive and iterative steps, such as searching, reading, organizing, and writing [29, 43]. This complexity necessitates "process-oriented literature review education" [10], where researchers learn by doing with their own research topics. Novices, however, often face multiple obstacles such as language barriers, lack of methodological experience, diminished confidence, and misconceptions [10]. Personalized mentorship can offer diverse supports [20] and reduce psychological stress [7, 8], but is not always accessible due to mentors' time constraints or lack of specific expertise [31].

Literature review supporting systems can play a complementary role by providing automated support for various tasks involved in the process. However, as we will elaborate in the next section, few systems offer comprehensive support covering the complete iteration from understanding papers to organizing and writing. Our study provides design implications for such systems, supported by empirical findings from novice researchers who engaged with our prototype system over a two-week period.

### 2.2 Literature Review Supporting Systems

Systems designed to manage and organize literature range from discovering relevant research work [11, 25, 37] and locating and navigating papers [21, 36], to flexibly creating and organizing review notes [3, 28, 46], as well as offering predefined analysis procedures [13, 15]. Another line of research focuses on facilitating the consumption of large volumes of papers, such as navigating through multiple related work sections, by providing enhanced browsing features like prioritizing unvisited content [9, 24, 40].

The rise of LLMs enabled the systems to directly provide content-related support, such as paper comprehension [14], critical reading [49], and feedback on users' writing in terms of fluency, coherence, word choice, and structure [23, 26, 33, 50]. Additionally, a range of commercial services is rapidly growing (e.g., [1, 2, 47]) offering intelligent searching and curation of literature and in-depth question-answering based on paper content [41].

While existing systems help reduce burdens and provide support for novice researchers in writing literature reviews, few bridge the gap from paper review to organizing and writing. Language model-based services mainly focus on delivering accurate content from papers, which is not optimal for supporting expository writing like literature reviews, where knowledge is generated through organizing evidence [45]. We introduce a scaffolded workflow, complemented by language-model-based actionable feedback, designed to assist users in constructing narratives for their writing based on literature.

## 3 FORMATIVE STUDY

We conducted semi-structured interviews with 11 novice HCI researchers to identify where they need system-based support during the literature review process. The results indicated that these novice researchers encounter several barriers, especially in the stages of initiating the literature review narrative, organizing the structure of the narrative, and reflecting on their drafts. From these findings, we have identified three key design requirements:

**DR1: Support initiating writing.** Before writing, participants encountered the challenge of initiating the writing process for their literature reviews. Our participants felt barriers moving on to writing literature reviews, feeling daunted by the numerous potential discussions, which in turn intensified their burden. P2 described, *"there are so many directions to explore. [...] Some days I think, 'Oh, that's important,' but the next day, I change my mind by saying, 'No, that's not related.' This pattern keeps repeating, and I am at a loss for what to do."* Activities that continued without tangible results increasingly pressured participants. P9 said, *"Reading ten papers doesn't reveal the results, but I have to write an introduction and related work at some point, so that's the stressful part".*

**DR2: Support organization of research narrative.** During the writing process, organizing research narratives became a central concern. When our participants began writing their literature reviews, they faced difficulties and recognized the necessity of developing a clear and coherent research narrative. P6 expressed difficulty in interpreting literature relevant to their study, stating, *"I read a lot, but applying it to my research was a totally different challenge. I mean the specific position of the study."* Although various tools and methods were used for organizing literature, such as spreadsheets, they often found these methods insufficient for developing a narrative. P7, for example, used a mind map to organize paper relationships but eventually had to reinterpret it with his own thoughts for documentation.

The key to this challenge appeared to be the active interpretation and integration of the literature in relation to one's own research project. P5 reported that discussing the relationship between the literature and her research early in the reading phase made it *"rather less difficult to cite when writing."* However, many participants had the opposite tendency when reading papers; they skimmed through papers to identify the most important sections as quickly as possible. P1 highlighted this approach, noting, *"You can say in just one or two lines what this research did [...] Is there an AI that just roughly tells you what it is about?"* (P1).

**DR3. Support reflection on clarity and comprehensiveness.** After writing, participants faced uncertainties regarding the clarity and depth of their reviews. They often worried about the clarity of their initial drafts, including issues like appropriate word choice (P3) and logical coherence (P2). In terms of comprehensiveness, they were anxious about whether they had sufficiently explored and discussed the literature. One participant shared their experience of overlooking key discussion points due to a narrow focus, saying, *"Initially, reading others' papers made us anxious. We feared they had already covered our ideas, leading us to focus less on their work and more on ours. This subjective view changed as I continued reading and started noticing aspects we initially missed"* (P1).

## 4 LITWEAVER

From the design requirements discovered in the formative study, we designed and developed a prototype system, LitWeaver, to support novice researchers in writing literature reviews. LitWeaver is implemented as a Chrome Extension for Notion, a web-based document application. It detects content and user focus in Notion, displaying a custom widget on the side to provide instructions and support features based on LLMs.

LitWeaver defines three stages of literature review: paper review, topic finding, and paragraph writing, to help users organize a narrative with clear milestones (**DR2: support for organization**). At each stage, LitWeaver offers three types of LLM-based support. The *Get it started* feature suggests starting examples to assist users in their writing (**DR1: support for initiation**). Next, *Polish up* provides feedback on the clarity of writing in two ways: (1) by highlighting vague expressions in user-written sentences and (2) by presenting a paraphrased summary to help users examine whether their sentences convey the intended meaning (**DR3: support for clarity**). Lastly, the *Look around* feature supports users in enhancing comprehensiveness by displaying questions that could lead to further discussion, thus offering possible missing perspectives (**DR3: support for comprehensiveness**).

Below, we describe through walkthrough examples how users produce outcomes at each stage and how LitWeaver assists them in the process.

### 4.1 A Walk-through Example

In the Paper Review Stage, Riley found interesting content in a paper relevant to their work but struggled to write review comments explaining why they wanted to quote it (Figure 1A1). Using the *Get it started* feature, Riley explored example comments and, inspired by the example review comments, decided to generalize and expand the quotation to relate it to their research project, crafting an initial review comment (A2-4). Next, Riley used the *Polish up* feature to get feedback on the clarity of their review comment. LitWeaver highlighted vague expressions in the sentences Riley had written and presented a paraphrase as how readers might understand the sentence as it is (A5-6). Realizing the paraphrase didn't convey what they intended, Riley revised the ambiguous expressions (A7). Lastly, using the *Look around* feature, Riley received feedback on the comprehensiveness of their review comment. This feature suggested several questions on perspectives potentially missing in Riley's review comment (A8). For instance, Riley thought the question about how the content of the quotation could be applied to other academic fields was important, realizing the need to find more use cases to strengthen their argument in future writings, and pinned this for later reference (A9).

In the Topic Finding Stage, Riley aimed to organize notes (i.e., quotations and review comments) into topic groups and name them but found it challenging due to the high volume of notes (B1). Upon using the *Get it started* feature, LitWeaver provided several example topics along with relevant notes (B2). Each suggested topic acted as a seed for starting to write one or more paragraphs, with the associated notes serving as writing materials. Riley accepted a suggested topic but refined it for more detail, checked for irrelevant notes, and sought missing relevant ones. This process was repeated until the topic was sufficiently concrete and comprehensive.

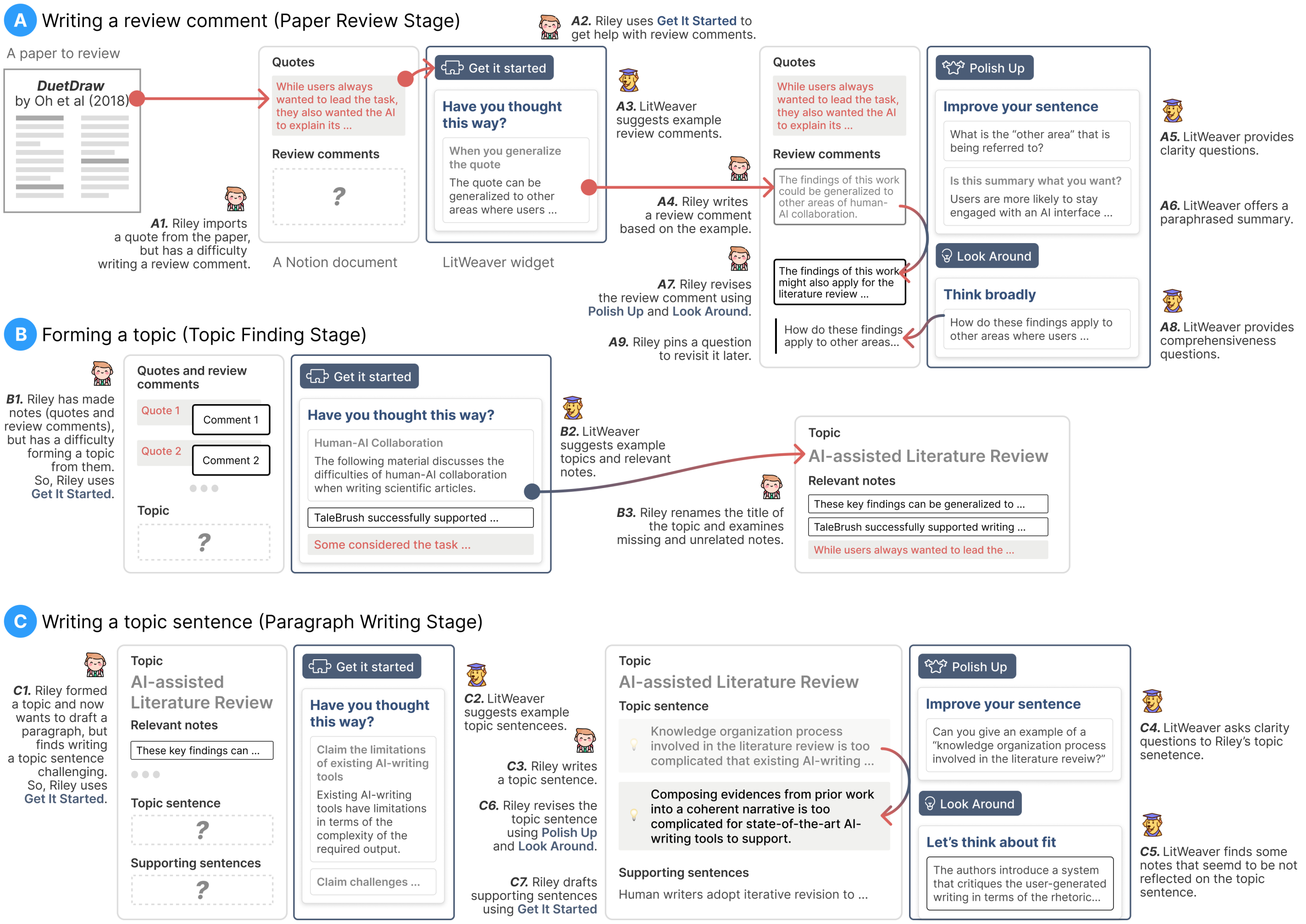
In the Paragraph Writing Stage Riley aimed to write paragraphs for a topic. To effectively structure the content, LitWeaver provided milestones for first thinking of a topic sentence and then adding supporting sentences, allowing the use of the *Get it started*, *Polish up*, and *Look around* features either for a topic sentence or supporting sentence. Riley first used the *Get it started* feature for help with writing a topic sentence. Based on the various claims and suggested example sentences generated by LitWeaver, Riley crafted a topic sentence and then explored supporting sentences using the *Get it started* feature again. The supporting sentences were provided along with roles such as *"Background of the topic sentence," "A possible solution to the problem,"* and *"Evidence to support the claim."* Inspired by the suggested supporting sentences, Riley completed an initial paragraph on the topic. Similarly to the Paper Review Stage, Riley used the *Polish up* feature to request feedback on the clarity of the paragraph (C4) and the *Look around* feature to get feedback on the comprehensiveness of the paragraph. Finding questions that inspired potential discussion points, Riley decided to develop a discussion in a new direction. Therefore, Riley returned to the Paper Review Stage to search for quotations about this point.

## 5 WORKSHOP EVALUATION

Using LitWeaver, we conducted a workshop study with 12 novice researchers. Nine participants joined a single session in the format of a group workshop to simulate a remote study group, while the other three participants went for a longer-term evaluation to observe how their usage patterns and perceptions changed over time.

### 5.1 Method

We recruited 12 novice researchers (9 female, 3 male) in HCI who had at most one publication in a major publication venue (e.g., ACM CHI, IEEE VIS). Nine participants—denoted as S1 to S9—joined a single-session group workshop (two or three in each group), and the other three participants—denoted as L1 to L3—took part in four individual sessions over two weeks (long-term). Each session was composed of 1.5 hour-long literature review activity with


A
Writing a review comment (Paper Review Stage)
A paper to review
DuetDraw
by Oh et al (2018)
Quotes
While users always wanted to lead the task, they also wanted the AI to explain its ...
Review comments
?
A Notion document
Get it started
Have you thought this way?
When you generalize the quote
The quote can be generalized to other areas where users ...
LitWeaver widget
Quotes
While users always wanted to lead the task, they also wanted the AI to explain its ...
Review comments
The findings of this work could be generalized to other areas of human-AI collaboration.
The findings of this work might also apply for the literature review ...
How do these findings apply to other areas...
Polish Up
Improve your sentence
What is the "other area" that is being referred to?
Is this summary what you want?
Users are more likely to stay engaged with an AI interface ...
Look Around
Think broadly
How do these findings apply to other areas where users ...
B
Forming a topic (Topic Finding Stage)
Quotes and review comments
Quote 1
Comment 1
Quote 2
Comment 2
Topic
?
Get it started
Have you thought this way?
Human-AI Collaboration
The following material discusses the difficulties of human-AI collaboration when writing scientific articles.
TaleBrush successfully supported ...
Some considered the task ...
Topic
AI-assisted Literature Review
Relevant notes
These key findings can be generalized to ...
TaleBrush successfully supported writing ...
While users always wanted to lead the ...
C
Writing a topic sentence (Paragraph Writing Stage)
Topic
AI-assisted Literature Review
Relevant notes
These key findings can ...
Topic sentence
?
Supporting sentences
?
Get it started
Have you thought this way?
Claim the limitations of existing AI-writing tools
Existing AI-writing tools have limitations in terms of the complexity of the required output.
Claim challenges ...
Topic
AI-assisted Literature Review
Topic sentence
Knowledge organization process involved in the literature review is too complicated that existing AI-writing ...
Composing evidences from prior work into a coherent narrative is too complicated for state-of-the-art AI-writing tools to support.
Supporting sentences
Human writers adopt iterative revision to ...
Polish Up
Improve your sentence
Can you give an example of a "knowledge organization process involved in the literature reveiw?"
Look Around
Let's think about fit
The authors introduce a system that critiques the user-generated writing in terms of the rhetoric...

*A1.* Riley imports a quote from the paper, but has a difficulty writing a review comment.

*A2.* Riley uses Get It Started to get help with review comments.

*A3.* LitWeaver suggests example review comments.

*A4.* Riley writes a review comment based on the example.

*A5.* LitWeaver provides clarity questions.

*A6.* LitWeaver offers a paraphrased summary.

*A7.* Riley revises the review comment using Polish Up and Look Around.

*A8.* LitWeaver provides comprehensiveness questions.

*A9.* Riley pins a question to revisit it later.

*B1.* Riley has made notes (quotes and review comments), but has a difficulty forming a topic from them. So, Riley uses Get It Started.

*B2.* LitWeaver suggests example topics and relevant notes.

*B3.* Riley renames the title of the topic and examines missing and unrelated notes.

*C1.* Riley formed a topic and now wants to draft a paragraph, but finds writing a topic sentence challenging. So, Riley uses Get It Started.

*C2.* LitWeaver suggests example topic sentencees.

*C3.* Riley writes a topic sentence.

*C4.* LitWeaver asks clarity questions to Riley's topic senetence.

*C5.* LitWeaver finds some notes that seemd to be not reflected on the topic sentence.

*C6.* Riley revises the topic sentence using Polish Up and Look Around.

*C7.* Riley drafts supporting sentences using Get It Started

**Figure 1: Riley (a novice researcher) writes a paragraph for the "AI-assisted Literature Review" topic. (A) In Paper Review Stage, Riley makes imports quotes from a paper (DuetDraw [38]) and writes a review comment inspired by examples from the *Get it started* feature and revises it using the *Polish up* and *Look around* features. (B) In Topic Finding Stage, Riley forms a topic based on the notes (quotes and review comments), assisted by the *Get it started* feature. (C) In Paragraph Writing Stage, Riley starts writing a topic sentence after exploring examples suggested by the *Get it started* feature and revises it with *Polish up* and *Look around.***

LitWeaver and one hour-long interview. In the interview, we asked about their overall experience of using the system and further discussed how they perceived and utilized the LLM-based support features. Detailed procedures for the workshop are elaborated in the appendix.

## 5.2 Results

### 5.2.1 *Supporting Initiation, Organization, and Reflection in Writing.*

The scaffolded workflow, combined with LLM-driven assistance, positively influenced our participants in initiating writing (DR1), organizing research narratives (DR2), and refining clarity and comprehensiveness (DR3).

For initiating writing, the scaffolded milestones enabled participants to begin drafting, while language-model-generated examples aided the rapid development of drafts. S5 found that the milestones made the writing process more manageable and helped in regaining focus to start actively writing. S6 used the suggested sentences for brainstorming, stating it eased the challenge of starting from scratch.

In organizing narratives, single-session participants saw the scaffolded workflow as a clear, structured guide for writing literature reviews. Over time, long-term participants L3 and L2 further integrated this workflow into their existing methods. L2 transitioned from spreadsheet templates to more extensive free-form comments, deepening her engagement with the literature. L3, who initially believed that a few keywords were enough for paper reviews, later discovered that writing detailed comments was more effective in improving her narrative construction and organization.

Regarding feedback, participants valued the clarity and comprehensive feedback for identifying underdeveloped areas. L1, who typically reflected on his writing independently, found the feedback useful for filling gaps and inspiring what to discuss in his write-ups.

*5.2.2 Calibrating Expectations with LLM Assistance.* Single-session participants had specific expectations regarding the role of LLM, desiring it to align with their individual needs. Some participants viewed LLM as an *assistant* that efficiently performs tedious tasks with precision. For instance, P5 *"kept reloading until the sentence that perfectly captured the intended meaning appeared."* Conversely, other participants envisioned LLM as a mentor, providing guidance and expertise while emphasizing that users should lead the learning process. They appreciated that LitWeaver, rather than offering direct corrections, allowed users to choose from suggestions or reflect clarity through paraphrased summaries. Participants exhibited different preferences over LLM-generated feedback depending on how determined a narrative they had beforehand. Those with a less congruent or no narrative perceived the feedback from *Polish up* and *Look around* to be useful. In contrast, people who already had concrete narratives in mind perceived the LLM-generated feedback as diverging and out-of-scope.

Long-term participants experienced shifts in their expectations of LLM over time, indicating a dynamic relationship with the technology as they became more familiar with its capabilities and limitations. Long-term participants generally perceived that the LLM had limited capabilities for suggesting novel and detailed narratives. For example, L3 perceived that "*Examples for the topic and supporting sentences were what I had already thought of, so they were not very useful. It organized stuff well but did not help me with difficult tasks.*" However, they tried to calibrate their strategy with the LLM features. For instance, L2 stated during the second session, "*I thought the AI would help me easily write a topic sentence, but . . . I had to come up with my own words. ... I felt betrayed.*" In the third session, however, she said, "*I had a clear message to deliver, so I could have just picked some relevant notes that matched my claim. Why was I so foolish to think I had to compose narratives from the system suggestion?*" and started to build narratives on her own. L1 found that he should have brought quotations more carefully (i.e., not just getting phrases from abstracts but actually reading the paper's content) for the similar reasons of L2 and L3, he tried to compensate using LLM features.

As a result, in the Paragraph Writing Stage, the three long-term participants showed varying levels of engagement with LLM. For example, while L1 regained trust and actively collaborated with the LLM, L2 still thought that it was a researcher's role to write narratives, so she used minimal LLM support. In particular, L1 found that the quality of the LLM-generated topic sentences and supporting sentences improved after he added more refined notes to the topic. On the other hand, L2 claimed that because LLM features, even paraphrasing, would prevent users from learning skills, she would allow for it to suggest similar content at most. Alternatively, L3 used LLM to validate her outcome, so she said, "*I think it's okay to let AI paraphrase things.*" Although L3 believed that she had to build narratives and did not refer to any brainstorming examples for topics and paragraphs, she thought it was okay to reduce repetitive tasks with the help of LLM. She noted, "*We finally collaborated as I understood exactly what role AI could play.*"

*5.2.3 Developing Agency and Confidence through LLM Interaction.* Our long-term participants established dynamic partnerships with LLM, moving beyond an instrumental perspective. Through calibrating their expectations with LLM, they developed agency and confidence in literature review writing and experienced emotional support that encouraged their continued engagement. This is contrary to single-session participants who perceived the LLM either as an instrumental assistant for task efficiency or as a mentor for guidance.

L1 stated that although it was AI, he felt like his colleagues were there to advise him. Similarly, L3 perceived that the system was guiding her step by step because once she wrote something, LLM provided hints regarding the next step, like a one-on-one review session. She expressed pride in having written *"this much"* as a result of her continued engagement.

Even L2, who emphasized the leading role of humans, developed a sense of agency from experiencing LLM's performance, which deviated from her initial expectations. "*I think the pressure [I had] gradually disappeared. At first, I was under pressure, and I thought I wouldn't be good at this. So, I had expectations for AI to provide some complete sentences, but that wasn't the case. So I tried to do it for myself, and then I started to think that, 'I'm good enough!'*"

# 6 DISCUSSION

## 6.1 Balancing Trade-offs in Literature Review Writing

In our scaffolded workflow designed for novice researchers, from reviewing papers and identifying topics to crafting paragraphs, we incorporated LLM-driven support that either guides them to proceed to the next step or helps them reflect on and refine their current work. We encouraged participants to revisit and enhance their work by offering feedback on the clarity and completeness of their written content, while also facilitating their transition to the next stage by providing example sentences. These two forms of support had complementary effects. On the one hand, participants benefited from enhancing the quality of their intermediary outcomes, which eventually helped them in subsequent stages. For instance, L2 and L3 found that writing more detailed review comments later served as a foundation for developing their narratives. L1, after receiving unsatisfactory LLM output due to the quality of collected quotes, was motivated to refine their input, leading to better results. On the other hand, progressing to the next stage without spending too much time on intermediary phases allowed participants to gain a sense of efficacy and agency. L3, initially overwhelmed by the abundance of information, gained confidence after deciding to focus on the main argument and structure their paragraphs accordingly. L2 also gained confidence and independence by deciding to take control of the argument-building process rather than relying too heavily on LLM. These results suggest that managing the trade-off between improving intermediary outcomes and advancing to the next stage is a critical design decision in supporting novice researchers during the literature review writing process. Systems designed to support literature reviews would benefit from maintaining a balance between different levels of writing activities, acting as a pace-setter in the writing process.

## 6.2 Enriching Literature Review Writing Experience

We identified that companionship and the sense of accomplishment attributed to making literature reviews motivating, plausible, and pleasurable. Our long-term participants perceived the LLM features as a writing companion; it made participants feel as if they had been working together with the system, and it motivated them to continue engaging with the literature review. Also, participants reflected on their own achievements, such as completing the first manuscript, to feel self-agency through the achievement.

We note that these positive changes were not solely dependent on the LLM performance. Instead, the interplay between user expectations of LLM and its actual performance shaped novice researchers' experiences. This implies that to enhance companionship with LLM, systems can exploit a variety of personas with different performance levels, allowing users to choose the one they find most relatable. For example, an agent might make some mistakes or produce random output so that users perceive the system as having a similar level of capacities 'like me' [19, 27]. This could encourage users to take more initiative in correcting the LLM's output or help them feel they are not alone in their literature review tasks. Alternatively, by offering more direct guidance, the system could simulate learning from a teacher, allowing users to observe their own progress and experience a greater sense of achievement.

## 6.3 Coping with Diverse Context

We found that providing structured milestones and inter-level support effectively aided in initiating, organizing, and reflecting on literature reviews. While we required users to perform various literature review activities in a single Notion interface, in reality, users operate within complex ecosystems of tools [42]. For example, some users may wish to highlight and make notes directly on the PDF reader interface. While accommodating users' preferred toolchains is ideal, our study demonstrated that novice researchers can effectively develop literature review skills through a process-oriented approach, benefiting from a unified, complete iteration cycle. Implementing a holistic literature review writing experience in an ecologically valid environment presents an opportunity for future research. In this context, we highlight the Semantic Reader Project [35] as a notable example, where a range of features designed to improve the reading and organization of academic papers are archived and integrated into a cohesive interface.

## REFERENCES

[1] 2024. Consensus: AI Search Engine for Research. https://consensus.app Last accessed January 24, 2024.

[2] 2024. ScholarAI | Home. Retrieved January 24, 2024 from https://scholarai.io

[3] Alexis Allot, Qingyu Chen, Sun Kim, Roberto Vera Alvarez, Donald C Comeau, W John Wilbur, and Zhiyong Lu. 2019. LitSense: making sense of biomedical literature at sentence level. *Nucleic acids research* 47, W1 (2019), W594–W599.

[4] John Bitchener and Madeline Banda. 2007. Postgraduate students' understanding of the functions of thesis sub-genres: The case of the Literature Review. *New Zealand Studies in Applied Linguistics* 13, 2 (2007), 61–68.

[5] David N Boote and Penny Beile. 2005. Scholars before researchers: On the centrality of the dissertation literature review in research preparation. *Educational researcher* 34, 6 (2005), 3–15.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[7] Rosemary S Caffarella and Bruce G Barnett. 2000. Teaching doctoral students to become scholarly writers: The importance of giving and receiving critiques. *Studies in Higher Education* 25, 1 (2000), 39–52.

[8] Gulfidan Can and Andrew Walker. 2011. A model for doctoral students' perceptions and attitudes toward written feedback for academic writing. *Research in Higher Education* 52, 5 (2011), 508–536.

[9] Joseph Chee Chang, Amy X Zhang, Jonathan Bragg, Andrew Head, Kyle Lo, Doug Downey, and Daniel S Weld. 2023. CiteSee: Augmenting Citations in Scientific Papers with Persistent and Personalized Historical Context. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15.

[10] Der-Thanq "Victor" Chen, Yu-Mei Wang, and Wei Ching Lee. 2016. Challenges confronting beginning researchers in conducting literature reviews. *Studies in Continuing Education* 38, 1 (2016), 47–60.

[11] Kiroong Choe, Seokweon Jung, Seokhyeon Park, Hwajung Hong, and Jinwook Seo. 2021. Papers101: Supporting the discovery process in the literature review workflow for novice researchers. In *2021 IEEE 14th Pacific Visualization Symposium (PacificVis)*. IEEE, 176–180.

[12] John Joon Young Chung, Wooseok Kim, Kang Min Yoo, Hwaran Lee, Eytan Adar, and Minsuk Chang. 2022. TaleBrush: Sketching Stories with Generative Pretrained Language Models. In *CHI Conference on Human Factors in Computing Systems*. 1–19.

[13] Kerry Dhakal. 2022. NVivo. *Journal of the Medical Library Association* 110, 2 (2022), 270–272.

[14] Raymond Fok, Joseph Chee Chang, Tal August, Amy X Zhang, and Daniel S Weld. 2023. Qlarify: Bridging scholarly abstracts and papers with recursively expandable summaries. *arXiv preprint arXiv:2310.07581* (2023).

[15] Susanne Friese. 2019. *Qualitative data analysis with ATLAS. ti.* Sage.

[16] Arnold D Froese, Brandon S Gantz, and Amanda L Henry. 1998. Teaching students to write literature reviews: A meta-analytic model. *Teaching of Psychology* 25, 2 (1998), 102–105.

[17] Katy Ilonka Gero and Lydia B Chilton. 2019. Metaphoria: An algorithmic companion for metaphor creation. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.

[18] 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.

[19] Matthew Guzdial, Nicholas Liao, Jonathan Chen, Shao-Yu Chen, Shukan Shah, Vishwa Shah, Joshua Reno, Gillian Smith, and Mark O Riedl. 2019. Friend, collaborator, student, manager: How design of an ai-driven game level editor affects creators. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.

[20] J Haves. 1980. Identifying the organization of writing processes. *Cognitive processes in writing* (1980).

[21] Victor Henning and Jan Reichelt. 2008. Mendeley-a last. fm for research?. In *2008 IEEE fourth international conference on eScience*. IEEE, 327–328.

[22] Allyson Holbrook. 2007. 'Levels' of success in the use of the literature in a doctorate. *South African Journal of Higher Education* 21, 8 (2007), 1020–1041.

[23] Grammarly Inc. 2022. *Grammarly: Free Online Writing Assistant.* Retrieved September 16, 2022 from https://www.grammarly.com/

[24] Hyeonsu B Kang, Joseph Chee Chang, Yongsung Kim, and Aniket Kittur. 2022. Threddy: An Interactive System for Personalized Thread-based Exploration and Organization of Scientific Literature. *arXiv preprint arXiv:2208.03455* (2022).

[25] Hyeonsu B Kang, Nouran Soliman, Matt Latzke, Joseph Chee Chang, and Jonathan Bragg. 2023. ComLittee: Literature Discovery with Personal Elected Author Committees. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.

[26] Simon Knight, Antonette Shibani, Sophie Abel, Andrew Gibson, and Philippa Ryan. 2020. AcaWriter: A learning analytics tool for formative feedback on academic writing. *Journal of Writing Research* (2020).

[27] Janin Koch, Andrés Lucero, Lena Hegemann, and Antti Oulasvirta. 2019. May AI? Design ideation with cooperative contextual bandits. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[28] Andrew Kuznetsov, Joseph Chee Chang, Nathan Hahn, Napol Rachatasumrit, Bradley Breneisen, Julina Coupland, and Aniket Kittur. 2022. Fuse: In-Situ Sensemaking Support in the Browser. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–15.

[29] Becky SC Kwan. 2008. The nexus of reading, writing and researching in the doctoral undertaking of humanities and social sciences: Implications for literature reviewing. *English for Specific Purposes* 27, 1 (2008), 42–56.

[30] Notion Labs. 2022. *Notion-One workspace. Every team.* Retrieved September 14, 2022 from https://www.notion.so/product

[31] Mary R Lea and Brian V Street. 1998. Student writing in higher education: An academic literacies approach. *Studies in higher education* 23, 2 (1998), 157–172.

[32] Yuyu Lin, Jiahao Guo, Yang Chen, Cheng Yao, and Fangtian Ying. 2020. It is your turn: collaborative ideation with a co-creative robot through sketch. In *Proceedings of the 2020 CHI conference on human factors in computing systems*.

1–14.

[33] Ming Liu, Rafael A Calvo, and Vasile Rus. 2012. G-Asks: An intelligent automatic question generation system for academic writing support. *Dialogue & Discourse* 3, 2 (2012), 101–124.

[34] Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023. Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment. *arXiv preprint arXiv:2308.05374* (2023).

[35] Kyle Lo, Joseph Chee Chang, Andrew Head, Jonathan Bragg, Amy X Zhang, Cassidy Trier, Chloe Anastasiades, Tal August, Russell Authur, Danielle Bragg, et al. 2023. The Semantic Reader Project: Augmenting Scholarly Documents through AI-Powered Interactive Reading Interfaces. *arXiv preprint arXiv:2303.14334* (2023).

[36] KK Mueen Ahmed and Bandar E Al Dhubaib. 2011. Zotero: A bibliographic assistant to researcher. *Journal of Pharmacology and Pharmacotherapeutics* 2, 4 (2011), 304–305.

[37] Arpit Narechania, Alireza Karduni, Ryan Wesslen, and Emily Wall. 2021. VITALITY: Promoting Serendipitous Discovery of Academic Literature with Transformers & Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2021), 486–496.

[38] Changhoon Oh, Jungwoo Song, Jinhan Choi, Seonghyeon Kim, Sungwoo Lee, and Bongwon Suh. 2018. I lead, you help but only with enough details: Understanding user experience of co-creation with artificial intelligence. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

[39] OpenAI. 2022. *Text completion - OpenAI API*. Retrieved September 14, 2022 from https://beta.openai.com/docs/guides/completion/prompt-design

[40] 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.

[41] Robert Pinzolits. 2024. AI in academia: An overview of selected tools and their areas of application. *MAP Education and Humanities* 4 (2024), 37–50.

[42] Xin Qian, Katrina Fenlon, Wayne Lutters, and Joel Chan. 2020. Opening up the black box of scholarly synthesis: Intermediate products, processes, and tools. *Proceedings of the Association for Information Science and Technology* 57, 1 (2020), e270.

[43] Diana Ridley. 2012. The literature review: A step-by-step guide for students. (2012).

[44] Quentin Roy, Futian Zhang, and Daniel Vogel. 2019. Automation accuracy is good, but high controllability may be better. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–8.

[45] 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. *arXiv preprint arXiv:2304.02623* (2023).

[46] Craig S Tashman and W Keith Edwards. 2011. LiquidText: A flexible, multitouch environment to support active reading. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 3285–3294.

[47] Typeset.io. 2024. AI chat for scientific PDFs | SciSpace. Retrieved January 24, 2024 from https://www.typeset.io

[48] Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. 2023. A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica* 10, 5 (2023), 1122–1136.

[49] Kangyu Yuan, Hehai Lin, Shilei Cao, Zhenhui Peng, Qingyu Guo, and Xiaojuan Ma. 2023. CriTrainer: An Adaptive Training Tool for Critical Paper Reading. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–17.

[50] Xin Zhao. 2022. Leveraging Artificial Intelligence (AI) Technology for English Writing: Introducing Wordtune as a Digital Writing Assistant for EFL Writers. *RELC Journal* (2022), 00336882221094089.

[51] David Zhou and Sarah Sterman. [n. d.]. Creative Struggle: Arguing for the Value of Difficulty in Supporting Ownership and Self-Expression in Creative Writing. ([n. d.]).

# A SYSTEM DETAIL

We implemented LitWeaver as a Chrome Extension for Notion [30], a Web-based document application. LitWeaver provides its features in a custom widget on the side. After detecting which stage a user is in from their focus in a document, LitWeaver displays instructions and its features accordingly in the *Control* view (Figure 2A). Users can discover the results by using a feature in the *Outcome* view (Figure 2B).

## A.1 Structuring Data in Notion

Notion allows for the flexible sectioning of a document for different purposes. There are two parts in a document: (1) a paper database for the Paper Review Stage and (2) topics and paragraphs for the Topic Finding Stage and Paragraph Writing Stage. The paper database uses the Notion-native database feature in a table format where each row represents a paper that a user has reviewed. By clicking each row, users can see the quotations and review comments of the corresponding paper in a pop-up.

## A.2 Language Model-based Supporting Features

We employed GPT-3, a language model provided by OpenAI [6], to enable LitWeaver's AI features. GPT-3's text completion API takes a text-based prompt as input and suggests an outcome text that seems to best respond to the input prompt. As GPT-3 models are known to have a task-agnostic understanding of language through pre-training, we engineered prompt text templates for various tasks supported by LitWeaver. We followed the official prompt design guideline from OpenAI [39]. For example, we crafted the prompt for the *Get it started* feature in the Paper Review Stage to ask for example review comments for a given quotation (Figure 3). In particular, this prompt template begins with a task description, example cases to specify the level of details, the input and output format, and the placeholder to be replaced with the user input. When a user uses the feature, LitWeaver first reads the user-selected content (e.g., supporting sentences) and relevant context (e.g., their corresponding topic sentence) from the current Notion document. Then, it feeds the content into the predefined prompt template and requests GPT-3 to run the completed prompt. Once LitWeaver receives the outcome from GPT-3, it parses the information from the raw text response for display purposes (Figure 3).

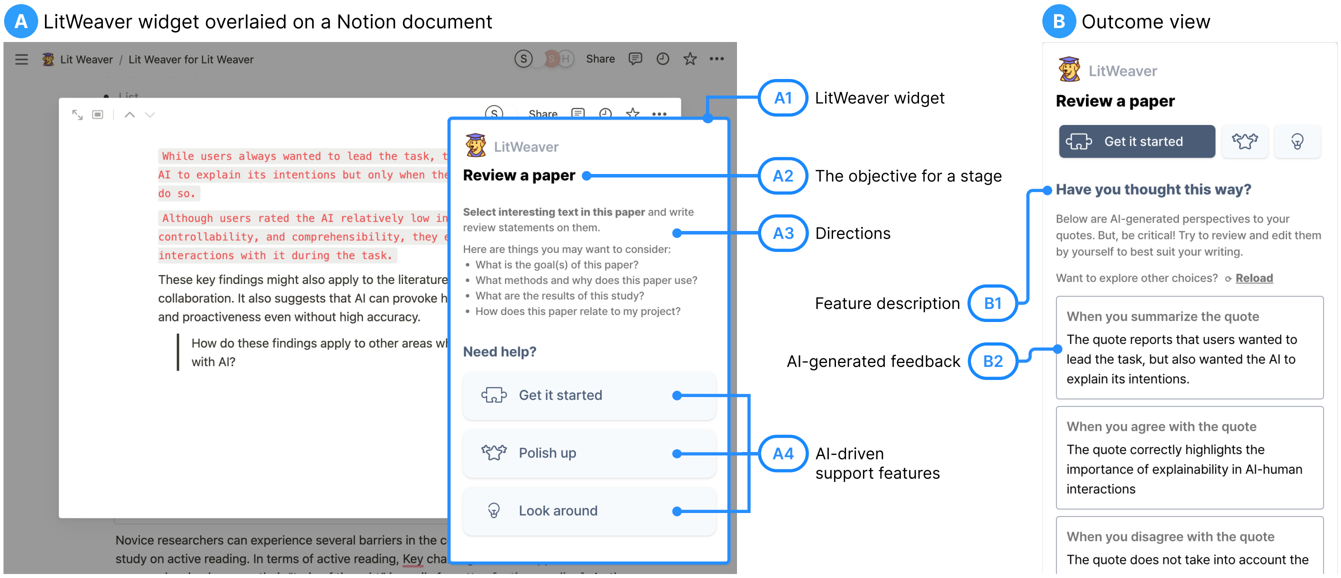
## A.3 Output Variability and Reloading

GPT-3 may generate different outcomes given the same text prompt. While this degree of randomness can be adjusted using a parameter (temperature), having variability to an extent could benefit the exploration of possible alternatives. We experimented with different temperature values to decide the final value that reliably gives quality outcomes given the model's capability. We enable reloading outcomes so that users can explore further alternatives in case they are not able to get reasonable outcomes initially.

# B WORKSHOP EVALUATION PROCEDURE

## B.1 Participants

We recruited 12 novice researchers (9 female, 3 male) in HCI who had at most one publication in a major publication venue (e.g., ACM CHI, IEEE VIS). As LitWeaver is developed in English, we invited those who were relatively fluent in written English. As shown in Table 1, groups A and C consisted of novice researchers who knew each other. All single-session and long-term participants individually performed a literature review for their current project. P10 from the formative study also joined a single-session workshop, identified as S8 in the workshop study.



While users always wanted to lead the task, t
AI to explain its intentions but only when the
do so.

Although users rated the AI relatively low in
controllability, and comprehensibility, they e
interactions with it during the task.

These key findings might also apply to the literature
collaboration. It also suggests that AI can provoke h
and proactiveness even without high accuracy.

> How do these findings apply to other areas wi
> with AI?

Novice researchers can experience several barriers in the c
study on active reading. In terms of active reading, Key cha

**Select interesting text in this paper** and write review statements on them.

Here are things you may want to consider:
- What is the goal(s) of this paper?
- What methods and why does this paper use?
- What are the results of this study?
- How does this paper relate to my project?

Below are AI-generated perspectives to your quotes. But, be critical! Try to review and edit them by yourself to best suit your writing.

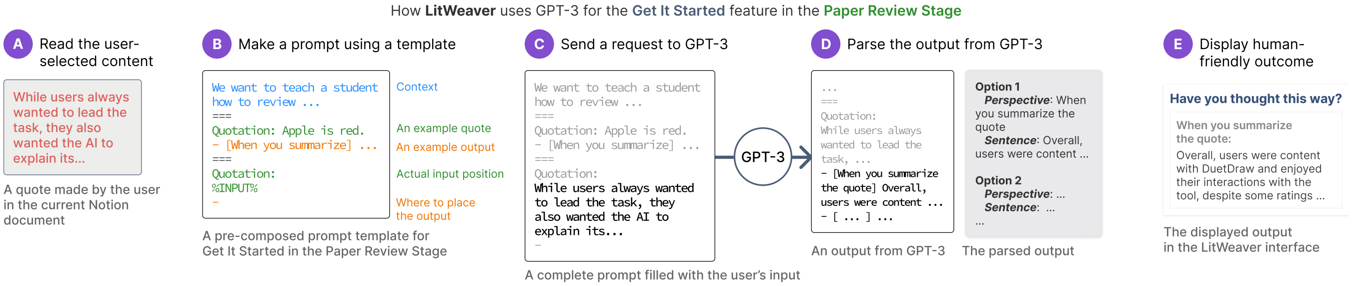
Want to explore other choices? ⟳ Reload

When you summarize the quote
The quote reports that users wanted to lead the task, but also wanted the AI to explain its intentions.

When you agree with the quote
The quote correctly highlights the importance of explainability in AI-human interactions

When you disagree with the quote
The quote does not take into account the

**Figure 2: LitWeaver operates as a single-panel widget (A1) overlaid on a Notion document. LitWeaver determines the stage of the literature review based on the content in the Notion document to provide an objective (A2), directions (A3), and AI-driven support features (A4) accordingly. (A) Users can click the buttons appearing at the bottom of the widget to use the AI-driven support features. (B) The outcome view shows the AI-generated feedback to the user’s request (B2).**


How LitWeaver uses GPT-3 for the Get It Started feature in the Paper Review Stage
A Read the user-selected content
A quote made by the user in the current Notion document
B Make a prompt using a template
Context
An example quote
An example output
Actual input position
Where to place the output
A pre-composed prompt template for Get It Started in the Paper Review Stage
C Send a request to GPT-3
A complete prompt filled with the user’s input
GPT-3
D Parse the output from GPT-3
An output from GPT-3
The parsed output
E Display human-friendly outcome
The displayed output in the LitWeaver interface

While users always wanted to lead the task, they also wanted the AI to explain its...

```
We want to teach a student
how to review ...
===
Quotation: Apple is red.
- [When you summarize] ...
===
Quotation:
%INPUT%
-
```

```
We want to teach a student
how to review ...
===
Quotation: Apple is red.
- [When you summarize] ...
===
Quotation:
While users always wanted
to lead the task, they
also wanted the AI to
explain its...
-
```

```
...
===
Quotation:
While users always
wanted to lead the
task, ...
- [When you summarize
the quote] Overall,
users were content ...
- [ ... ] ...
```

**Option 1**
*Perspective*: When you summarize the quote
*Sentence*: Overall, users were content ...

**Option 2**
*Perspective*: ...
*Sentence*: ...
...

**Have you thought this way?**

When you summarize the quote:
Overall, users were content with DuetDraw and enjoyed their interactions with the tool, despite some ratings ...

**Figure 3: To retrieve example review comments with the *Get it started* feature in the Paper Review Stage, LitWeaver takes as input a quote from a Notion document (A) and fills the pre-composed template with the quote (B) and sends it to GPT-3 (C). Then, LitWeaver parses the raw-text output from GPT-3 (D) and display human-friendly results to the user (E).**

## B.2 Procedures

Regardless of the workshop type (single-session or long-term), each session was composed of 1.5 hour-long literature review activity with LitWeaver and one hour-long interview. Before each session, participants attended an hour-long individual tutorial to make sure that LitWeaver was running on their own devices and that they were familiar with the system. We asked them to prepare 10–15 papers that were relevant to their own research topics and to share those papers with us at the end of the tutorial session.

For the single-session group workshop, participants individually (i.e., without interacting with other participants) performed a literature review, and then responded to a semi-structured focus-group interview together. In the interview, we asked about their overall experience of using the system and further discussed how they perceived and utilized the three stages of LitWeaver and language model-based support features. Lastly, we asked how they would use the system in the future. Each long-term session followed the same procedure except that they joined individually (i.e., one person per session). In the interview session of each long-term session, we further asked how they used the material from the previous session and how their perspectives to LitWeaver changed over time. Upon the completion, participants received compensation of 35,000 KRW (single-session) and 110,000 KRW (long-term) (approximately 26 USD and 80 USD, respectively)

## B.3 Pre-populated Material for Single-session Workshops

In LitWeaver, the Topic Finding Stage and Paragraph Writing Stage assume that users have produced a sufficient number of quotations and review comments (from about 10–15 papers) in the Paper Review Stage, so that GPT-3 can produce more than trivial

**Table 1: Participants of the single-session and long-term workshop studies. Single-session participants jointly took part as a group, while long-term participants joined individually.**

| ID | Type | Group | Note |
|---|---|---|---|
| S1 | single-session | A | Lab colleague of S2, S3 |
| S2 | single-session | A | Lab colleague of S1, S3 |
| S3 | single-session | A | Lab colleague of S1, S2 |
| S4 | single-session | B | |
| S5 | single-session | B | |
| S6 | single-session | C | Co-author of S7 |
| S7 | single-session | C | Co-author of S6 |
| S8 | single-session | D | Formative study participant (P10) |
| S9 | single-session | D | |
| L1 | long-term | - | |
| L2 | long-term | - | |
| L3 | long-term | - | |

outcomes. Because an 1.5 hour-long session was too short to produce that many notes, we provided pre-populated material for the single-session workshops. They were four quotes per paper: the title and three GPT-3-selected sentences from the abstract (prompt: “what are the most important sentences?”). In addition, we asked participants to remain in the Paper Review Stage for the first 50 minutes so that they could use their own notes as well in the later stages.

# C ADDITIONAL DISCUSSION

## C.1 Performance of Language Model

We clarify that the prototype system was built with GPT-3, and at the time of the study—when GPT-3 was the latest—participants were generally unfamiliar with language models. This situation contrasts with the current environment, where language models like ChatGPT have enhanced performance, and there is a growing expectation among the general public for such improved AI capabilities. Future research can delve deeper into the relationship between AI performance and its educational influence. This could also include examining how ethical issues, such as plagiarism, and potential side effects, like hallucination [34], interact with the use of language models. That being said, the improved performance of AI can play a role in reducing undesirable challenges. Research on human-AI interaction indicates that users often prefer to lead and maintain a significant level of control over an AI model [38, 44]. Zhou introduces the concept of "creative struggle" [51], suggesting that it's beneficial for AI to offload tasks where users are less motivated, allowing them to focus on areas where they have ownership and are more engaged.

## C.2 Supplementing the Evaluation

We could observe the long-term lived experience of novice researchers conducting their own literature review. Yet, the findings of our study can be further validated by quantitative analyses. Even though we included some insights derived from log analysis in the supplementary material, we found that the request logs were not sufficient to fully observe the impact of AI features on users' content generation processes. By analyzing videos with appropriate definitions of usage patterns, for example, more insights can be gained from the compound interactions of users. A comparison of the system's effectiveness with other baselines and conditions will also be beneficial. We can compare the literature review experience with or without a system or compare the experience of novices and experts. We can add more conditions based on each user's status (e.g., stage of literature review, language competency). With the proper subdivision of user characteristics, we would be able to identify more diverse support needs based on quantitative observations.