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

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/authors.

ACM Reference Format:

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.
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
we designed and developed a prototype system, 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 comprehensiveness). 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
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 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 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


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 output 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, they also wanted the AI to explain it. Therefore, they wanted to lead the task, but also asked the AI to explain its intentions. The quote report that users wanted to lead the task, also asked the AI to explain its intentions.

The AI-generated feedback feature in the Paper Review Stage, 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). The parsed 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 workshops, 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.

<table>
<thead>
<tr>
<th>ID</th>
<th>Type</th>
<th>Group</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>single-session</td>
<td>A</td>
<td>Lab colleague of S2, S3</td>
</tr>
<tr>
<td>S2</td>
<td>single-session</td>
<td>A</td>
<td>Lab colleague of S1, S3</td>
</tr>
<tr>
<td>S3</td>
<td>single-session</td>
<td>A</td>
<td>Lab colleague of S1, S2</td>
</tr>
<tr>
<td>S4</td>
<td>single-session</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>S5</td>
<td>single-session</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>S6</td>
<td>single-session</td>
<td>C</td>
<td>Co-author of S7</td>
</tr>
<tr>
<td>S7</td>
<td>single-session</td>
<td>C</td>
<td>Co-author of S6</td>
</tr>
<tr>
<td>S8</td>
<td>single-session</td>
<td>D</td>
<td>Formative study participant (P10)</td>
</tr>
<tr>
<td>S9</td>
<td>single-session</td>
<td>D</td>
<td></td>
</tr>
<tr>
<td>L1</td>
<td>long-term</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L2</td>
<td>long-term</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L3</td>
<td>long-term</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.