ReviewFlow: Intelligent Scaffolding to Support Academic Peer Reviewing

Lu Sun
University of California San Diego
La Jolla, CA, USA

Aaron Chan
University of California San Diego
La Jolla, CA, USA

Yun Seo Chang
University of California San Diego
La Jolla, CA, USA

Steven P. Dow
University of California San Diego
La Jolla, CA, USA

ABSTRACT

Peer review is a cornerstone of science. Research communities conduct peer reviews to assess contributions and to improve the overall quality of science work. Every year, new community members are recruited as peer reviewers for the first time. How could technology help novices adhere to their community's practices and standards for peer reviewing? To better understand peer review practices and challenges, we conducted a formative study with 10 novices and 10 experts. We found that many experts adopt a workflow of annotating, note-taking, and synthesizing notes into well-justified reviews that align with community standards. Novices lack timely guidance on how to read and assess submissions and how to structure paper reviews. To support the peer review process, we developed ReviewFlow – an AI-driven workflow that scaffolds novices with contextual reflections to critique and annotate submissions, in-situ knowledge support to assess novelty, and notes-to-outline synthesis to help align peer reviews with community expectations. In a within-subjects experiment, 16 inexperienced reviewers wrote reviews in two conditions: using ReviewFlow and using a baseline environment with minimal guidance. With ReviewFlow, participants produced more comprehensive reviews, identifying more pros and cons. However, they still struggled to provide actionable suggestions to address the weaknesses. While participants appreciated the streamlined process support from ReviewFlow, they also expressed concerns about using AI as part of the scientific review process. We discuss the implications of using AI to scaffold the peer review process on scientific work and beyond.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**.

KEYWORDS

intelligent scaffolding, academic peer review, Large Language Models LLMs)

ACM Reference Format:
Lu Sun, Aaron Chan, Yun Seo Chang, and Steven P. Dow. 2024. ReviewFlow: Intelligent Scaffolding to Support Academic Peer Reviewing. In *29th International Conference on Intelligent User Interfaces (IUI '24), March 18–21, 2024, Greenville, SC, USA.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3640543.3645159

1 INTRODUCTION

Peer review is a cornerstone of academic research, ensuring the quality, credibility, and reliability of scientific research [70]. The peer review process seeks to assess whether submissions contribute new knowledge to a research community and generate feedback that helps authors improve the quality of their work [43, 76]. Many communities are seeing a rapid increase in submissions [4, 60, 76, 79, 80]; while this could be seen as an indicator of scientific progress, it also has increased the pressure on reviewers. To meet increased demand, many research communities recruit a significant number of first-time reviewers or ACs (Associate Chairs) for each review cycle. For example, in 2023, the ACM CHI Conference on Human Factors in Computing Systems reported that more than 50% of ACs were first-time ACs [2].

Peer reviewing is a complicated and challenging task. Reviewers need to understand the paper, evaluate the scientific content using domain knowledge, make a fair decision, and compose a comprehensive review to communicate their assessments and recommendations [76]. It is a task that requires critical thinking, a deep understanding of the subject, and the ability to provide constructive feedback. Conferences and journals need to ensure that first-time reviewers meet the standards of the research community.

One approach to preparing novices for any complex task is to provide scaffolding, a strategy used in educational settings to support learning and mastery [19, 73]. Scaffolding takes many forms, including examples of prior work, templates, or hints to help novices think and potentially perform on par with experts [36, 75]. Previous studies found that scaffolded examples and templates can even help learners perform similarly to experts in terms of feedback quality [94]. For instance, the LetterSmith project developed an approach to aid writing called "scaffolded annotation" which provided key components of the writing tasks and annotated expert examples, and helped professional writing students improve the quality of their early-stage drafts [39].

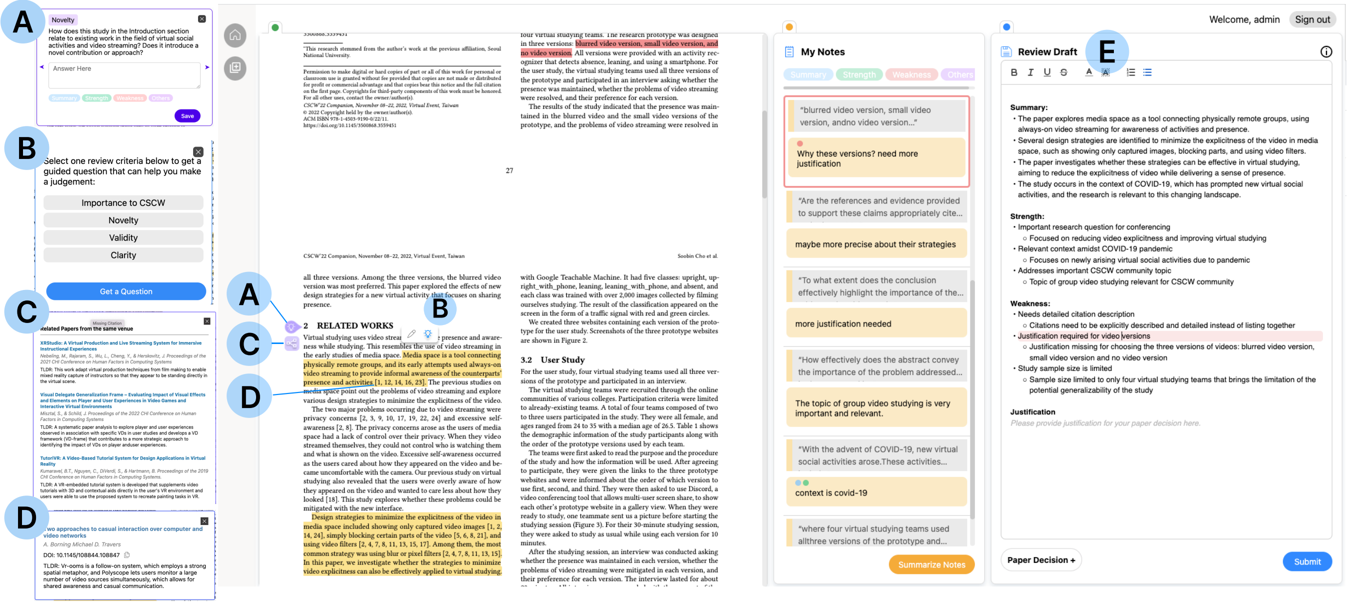
Advances in AI and large-language models (LLMs) in particular, have the potential to make scaffolding even more effective because they can detect and adapt to the user's work context [71]. For example, the CReBot project leverages LLMs to generate and place

Permission to make digital or hard copies of all or part 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 components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

IUI '24, March 18–21, 2024, Greenville, SC, USA

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0508-3/24/03
https://doi.org/10.1145/3640543.3645159



Novelty

How does this study in the Introduction section relate to existing work in the field of virtual social activities and video streaming? Does it introduce a novel contribution or approach?

Answer Here

Select one review criteria below to get a guided question that can help you make a judgement:

Related Papers from the same venue

XRStudio: A Virtual Production and Live Streaming System for Immersive Instructional Experiences
Nebeling, M., Rajaram, S., Wu, L., Cheng, Y., & Herskovitz, J. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems
TLDR: This work adapt virtual production techniques from film making to enable mixed reality capture of instructors so that they appear to be standing directly in the virtual scene.

Visual Delegate Generalization Frame – Evaluating Impact of Visual Effects and Elements on Player and User Experiences in Video Games and Interactive Virtual Environments
Misztal, S., & Schild, J. Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems
TLDR: A systematic paper analysis to explore player and user experiences observed in association with specific VDs in user studies and develops a VD framework (VD-frame) that contributes to a more strategic approach to identifying the impact of VDs on player anduser experiences.

TutoriVR: A Video-Based Tutorial System for Design Applications in Virtual Reality
Kumaravel, B.T., Nguyen, C., DiVerdi, S., & Hartmann, B. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.
TLDR: A VR-embedded tutorial system is developed that supplements video tutorials with 3D and contextual aids directly in the user's VR environment and users were able to use the proposed system to recreate painting tasks in VR.

Two approaches to casual interaction over computer and video networks
A. Borning Michael D. Travers
DOI: 10.1145/108844.108847
TLDR: Vr-ooms is a follow-on system, which employs a strong spatial metaphor, and Polyscope lets users monitor a large number of video sources simultaneously, which allows for shared awareness and casual communication.

My Notes

"blurred video version, small video version, andno video version..."
Why these versions? need more justification

"Are the references and evidence provided to support these claims appropriately cite...
maybe more precise about their strategies

"To what extent does the conclusion effectively highlight the importance of the...
more justification needed

"How effectively does the abstract convey the importance of the problem addressed...
The topic of group video studying is very important and relevant.

"With the advent of COVID-19, new virtual social activities arose.These activities...
context is covid-19

"where four virtual studying teams used allthree versions of the prototype and...

Review Draft

Summary:
- The paper explores media space as a tool connecting physically remote groups, using always-on video streaming for awareness of activities and presence.
- Several design strategies are identified to minimize the explicitness of the video in media space, such as showing only captured images, blocking parts, and using video filters.
- The paper investigates whether these strategies can be effective in virtual studying, aiming to reduce the explicitness of video while delivering a sense of presence.
- The study occurs in the context of COVID-19, which has prompted new virtual social activities, and the research is relevant to this changing landscape.

Strength:
- Important research question for conferencing
  - Focused on reducing video explicitness and improving virtual studying
- Relevant context amidst COVID-19 pandemic
  - Focuses on newly arising virtual social activities due to pandemic
- Addresses important CSCW community topic
  - Topic of group video studying relevant for CSCW community

Weakness:
- Needs detailed citation description
  - Citations need to be explicitly described and detailed instead of listing together
- Justification required for video versions
  - Justification missing for choosing the three versions of videos: blurred video version, small video version and no video version
- Study sample size is limited
  - Sample size limited to only four virtual studying teams that brings the limitation of the potential generalizability of the study

Justification
Please provide justification for your paper decision here.

Figure 1: ReviewFlow interface on an example paper. Users can (A) receive section-level contextual cues guided by community criteria; (B) request phrase-level contextual cues adapted to highlight paper content; (C) click the citation to get an in-situ summarization; (D) check the recommended citations not currently cited by the paper; and (E) click to summarize the notes into a high-level outline or expand into a detailed outline

questions within an academic paper to help induce critical thinking while reading [68]; results showed that CReBot helped novices read and comprehend the paper content. While LLMs show great potential, they are also known to hallucinate [7] and perpetuate systemic biases [38, 42] that could negatively impact the effectiveness of scaffolding. While intelligent scaffolding has proven valuable for well-scoped tasks, academic peer review involves reading submissions with a critical eye, synthesizing knowledge, and making a well-justified judgment for acceptance or rejection. Our research investigates how intelligent scaffolds can guide a complex, multi-faceted workflow for academic peer reviewing without biasing the ultimate decisions. This paper explores two key research questions: What are the challenges faced by novices and the strategies adopted by experts during the peer review process? How can we intelligently scaffold the peer review process?

To explore how we might use AI scaffolding to support the peer review process, we first conducted a formative study to understand what novices see as key challenges and how experts approach this task. We interviewed 10 novices with limited prior experience writing peer reviews to articulate their obstacles during the process. They expressed challenges around the lack of sufficient guidance on how to write a well-structured peer review peer review and how to make judgments about the paper's quality. Furthermore, we conducted observational studies with 10 experts to ask them to write a peer review for a short paper and to think aloud as they complete their workflow. Then, we invited experts to provide their perspectives on using AI to support the tasks and express their concerns. We found that many experts adopt a workflow involving critical reading, annotating, note-taking, and synthesizing a well-justified review that conforms to community guidelines.

Based on these insights, we developed a prototype – "ReviewFlow"[1], a platform for writing peer reviews that incorporates intelligent scaffolding to support a workflow for inexperienced reviewers. ReviewFlow incorporates a range of features to facilitate the review process: (1) **Contextual cues** are embedded questions that help reviewers reflect on the paper. ReviewFlow includes section-level cues guided by the community's review criteria as well as phrase-level cues that adapt to the paper's content. (2) **In-situ citation recommendations** show relevant but non-cited papers that may help reviews assess the novelty compared to existing work. (3) **Notes-to-outline synthesis** guides reviewers to organize notes and structure reviews to align with community standards. ReviewFlow gathers all the notes left by the reviewer and leverages an LLM to summarize notes into a high-level outline. Reviewers can revise and add detail to the outline while adhering to community standards.

We conducted a within-subjects study to evaluate ReviewFlow where (N=16) participants — with little to no experience as reviewers — wrote reviews for two short papers in a counterbalanced manner: one using the ReviewFlow with all scaffolding features and one using the baseline interface with only traditional forms of guidance (e.g., review rubrics and an example review). We found that novice reviewers wrote significantly more structured and more comprehensive reviews in the ReviewFlow system than in the Baseline system, as evaluated by experts. Novice reviewers wrote slightly more constructive reviews in the ReviewFlow system, but the difference was not significant. Reviewers called out more weaknesses

[1] ReviewFlow Code Repository: https://github.com/LusunHCI/ReviewFlow.git

in the paper using ReviewFlow, but they still struggled to provide actionable suggestions for the authors to address the weaknesses.

Our paper offers several contributions: First, a formative study revealed that novices lack opportune guidance on key considerations and expectations and uncovered common practices adopted by experts in the review process. Second, we developed ReviewFlow to model experts' workflow for peer reviewing while also leveraging LLMs to provide contextual cues, in-situ knowledge recommendations, and notes-to-outline synthesis. Third, we gained empirical insights from a within-subjects study with 16 participants which revealed how intelligent scaffolding can help novices write well-structured and comprehensive reviews.

2 RELATED WORK

2.1 Practices and challenges related to academic peer reviewing

As an important step for ensuring the scientific quality of work, peer review has been adopted by most journals and conferences [70]. In a typical conference review process, each reviewer needs to evaluate the paper's quality, make an acceptance decision and provide reviews for their assigned papers [76, 77]. The reviewers are usually experts in the area who have fruitful experience and knowledge to assess or evaluate the quality and contribution of the paper. After reviewers submit the review, a discussion takes place between reviewers and a meta-reviewer, who will carry out the final decision on the acceptance of the paper.

The number of papers in research communities has increased exponentially in recent years [85]. While this may be positively viewed as an acceleration of scientific progress, the disparity between growth rates of the submission and reviewer pools also creates more burden for reviewers [60, 76, 79, 80]. To avoid overloading reviewers, conferences need to find new sources of reviewers as there are not enough experienced reviewers to review all papers [80]. These novice and junior reviewers constitute a large fraction of the reviewer pool in computer science conferences [81]. For example, in 2023, the ACM CHI Conference on Human Factors in Computing Systems reported that more than 50% of ACs who reviewed papers are first-time ACs [2]. Given this large fraction, conferences need to ensure that newly added junior reviewers do not compromise the quality of the process, that is, are able to write reviews of quality comparable to the experienced reviewers.

A previous study explored and compared the reviews written by experienced reviewers versus junior reviewers and the study showed that junior reviewers were slightly harsher in scoring the clarity of the submissions [77, 80]. In the meantime, other works provide empirical evidence that junior reviewers are more critical than their senior counterparts and reveal that graduate students' review comments are not very useful [62, 67]. Faced with these doubts and challenges, helping novice junior reviewers to write a high-quality review becomes crucial.

Reviewing is a time-consuming and mentally demanding task [76, 85]. A constructive and comprehensive review can improve the quality of the paper, while a bad, random, dismissive, or biased review brings frustration and anger to authors [85]. To provide a high-quality review, reviewers need to go through a multi-step workflow, including understanding the contribution of the paper, accessing the merit of scientific contribution and providing an evaluation together with a comprehensive written review [16, 25, 37, 43, 53, 78, 85, 89, 96]. To make a fair judgment, reviewers need to have enough background knowledge to grasp the main idea of the paper and evaluate its contribution. More importantly, reviewers need to equip critical thinking skills to think deeply about the author's judgments, like whether the claims are reasonable and why the approaches are chosen [21, 63]. A typical peer review not only contains the paper summary and its contribution but also raises weaknesses from different aspects together with constructive feedback or thought-provoking questions. During this process, it requires readers to actively analyze, synthesize, and evaluate the paper content [63].

Researchers are typically trained extensively in conducting the research itself, but they often lack formal instructions in the peer review process. Hence, it becomes even more challenging for inexperienced junior reviewers to gain expertise quickly [76]. Existing research explored instructional methods to "teach" or "train" junior reviewers. A previous study provided a training video and found that it increased the inter-reviewer agreement, alignment with the scoring rubric, and the amount of time reading the review criteria [74]. Another study offers novice reviewers a more guided introduction to the different stages of the reviewing process, such as how to lead a discussion among reviewers, to help novices write better reviews. The results showed that with this guidance on the reviewing stages, novice reviewers could deliver more "above expectation" reviews [80]. However, their guidance is only limited to introducing the different parts of the reviewing process, such as rebuttal and discussion, and providing novices opportunities to ask expert questions on the general process [80]. Outside of the general review process, academic peer review is a complex activity that involves both understanding a paper submission's stated contributions and evaluating whether the paper crosses an acceptable threshold for the research community. To explore how we might support the peer review process, we start with a formative study to understand how experts approach this task and what novices see as key challenges.

Previous research has used computational methods to provide support to streamline several parts of the peer review process, such as matching submissions with appropriate reviewers or assessing review quality [4, 8, 37, 77, 85, 96]. However, fewer empirical studies that attempt to scaffold the entire workflow for reviewing academic papers, which includes reading, note-taking, evaluating, decision making, and synthesizing this into a written review.

2.2 Scaffolding strategies for complex cognitive tasks

To help novices improve problem-solving skills in complex cognitive tasks, cognitive apprenticeship introduces several strategies, including modeling, coaching, scaffolding, and reflection [19, 83]. Scaffolding is instructional support provided by experts to promote learning, especially when concepts and skills are being first introduced to novice students [19, 73]. These supports include advanced organizers, modeling, worked examples, concept maps, explanations, handouts, and prompts [5, 12, 14, 19, 64, 65]. Previous research shows that effective scaffolding can help novices perform work nearly as well as experts [39, 40, 94]. When scaffolding is mediated by technology, including AI-based methods, it creates more opportunities for instructors and learners but also brings more challenges in making the scaffolding contextualized, adaptive, and effective [32].

Researchers used prompts and guided questions to scaffold learners in the paper reading process [15, 68, 95]. To facilitate critical paper reading, researchers developed CReBot which interactively asks section-level critical thinking questions for routine paper readers and novice readers. Results showed that the interactive question prompts CReBot provided might not be better than static guidelines for beginners to conduct critical thinking. Furthermore, researchers developed CriTrainer which can adaptively provide questions in the reading process together with hints and feedback to help readers critically think and comprehend the paper content. Interestingly, on the opposite of CReBot, their result showed that CriTrainer can improve learners' ability to raise understandable, relevant, and critical questions after the training sessions. Their results highlighted the benefits of its text-specific critical thinking questions provided by the system. However, guided questions used in CReBot and CriTrainer are both template-based and did not fully use the user-selected content on the paper.

Existing research also used existing examples to scaffold complex writing processes [39, 40, 94]. Scholars found that scaffolding can help students learn about form and organization by analyzing the examples and templates [19, 20]. In the context of writing introductory help requests, providing high-quality examples and expert-informed templates can increase learning and writing quality [40]. Another writing support system used "scaffolded annotation" that broke down examples into each component to help professional writing students improve their early-stage drafts [39]. In traditional instruction scenarios, experts take a large amount of time to curate examples, create guidance, or author rubrics [73]. However, experts who created examples still faced the challenges of effectively adapting and contextualizing into the current learning step.

2.3 Leveraging AI scaffolding to support writing

Advances in AI and LLMs provide opportunities to provide efficient and context-specific scaffolding in the learning process [17, 18, 28–30, 41, 41, 54, 61, 72]. For example, TaleBrush allowed users to create a story with AI through sketching to aid the planning of writing. Then it enabled writers to generate diverse storylines and interactively refine them [17]. Another system Wordcraft explored how to support users collaborating with generative language models to co-write a story [93]. Spark used a language model to generate prompts related to a scientific concept to facilitate scientific writing [30]. However, as far as we know, none of the studies that used scaffolding strategies focused on the context of conference peer review writing. Our research explores how we might guide a complex, multi-faceted workflow like peer reviewing and the potential role of AI in creating contextually adaptive scaffolds.

Writing is a complex, iterative process [22]. The Hayes model describes the cognitive processes an individual writer engages in during the process of writing [22, 27, 35]. In the cognitive process of writing, there are three major components: planning, translating, and reviewing, as shown in Figure 5. Several systems are developed to facilitate different stages of writing [9, 97]. For example, VISAR is an AI-enabled writing assistant that helps writers brainstorm and revise hierarchical goals and organize argument structures in the planning stage. To facilitate the iterative planning and revising process in writing, intelligent systems further use the chain of thoughts (COT) prompting method to break down the large problem into step-by-step prompts [87]. Specifically, the Re3 framework and DOC framework used the COT approach to decompose the writing tasks where they first generate an outline and then automatically turn the outline into the story generation [90, 91]. Their evaluation demonstrated that the decomposition approach can improve the coherence of long story generation and is highly controllable where humans can control the story generation by modifying the outlines.

Control and agency are extremely important in the conference peer review writing process, as human reviewers should play the role of driving the writing process. While the development of LLMs can provide scaffolding opportunities for this writing process, it is important to address the concerns and limitations of LLMs in this context [24]. The prior work that aligns most closely with the concept of applying AI techniques to reviewing academic papers that a machine model that automatically generates feedback using LLMs. Results showed that LLM feedback could benefit researchers in earlier stages of manuscript preparation while researchers struggle with an in-depth critique of study methods [55]. Instead of using an automatic method, humans should be in the loop to drive and control the writing process [24]. Drawing on these insights, we designed the ReviewFlow system not to automate any parts of the process, but rather to scaffold key considerations and to give novices agency over how to apply machine-generated language suggestions.

3 FORMATIVE STUDIES

Before we developed our system to support academic peer reviewing, we conducted two formative studies. First, we interviewed ten novice reviewers to understand how they approached this task for the first time and their perceived challenges (represented as C). Second, we conducted observational studies with ten experienced reviewers where we invited them to write a peer review on a selected paper to capture their common practices and workflows. Last, based on the findings, we proposed design goals (represented as DG) for supporting novice reviewers.

3.1 Methods

3.1.1 *Novice Interview Study.*

We conducted semi-structured interviews with 10 novice reviewers (N1-N10) with relatively little experience with writing academic peer reviews (ranging from only reviewing once before and having less than 2 years of experience). Participants were recruited through mailing lists and social media posts. The participants (4 female and 6 male, average age of 25.5 years) came from diverse research fields including Human-Computer Interaction, AI, Cognitive Science, and Computer Security. Interviews were conducted remotely by the lead author and lasted around 30 minutes.

In the interview, we first asked open-ended questions about their perceived obstacles and challenges of conducting academic peer review. We then provided scenarios of potential features for peer review scaffolding to elicit reactions from novice reviewers. These scenarios described common potential situations faced by first-time reviewers and were designed to prompt the participants to share their needs in a real-world situation [26]. For example, "Mary is a first-year graduate student doing Human-Computer Interaction (HCI) research. While she has submitted a couple of papers before she received reviews, this will be her first time as a reviewer. She struggled to make the review constructive for the authors". Interviews were recorded with participants' permission and were transcribed. Two researchers from the team went through the transcripts and coded themes using thematic analysis [10]. Through multiple iterations along with periodic discussions, the coding led to the major themes of challenges below.

3.1.2 *Experts Observational Study.*

We conducted observational studies with 10 experts reviewers(E1-E10) with at least 5 years of experience in writing academic peer reviews, to learn about their best practices [45]. Participants were recruited through in-person invitations, email lists, and social media posts. Participants (4 female and 6 male, average age of 30.1 years) came from different fields of computer science, including HCI, AI, programming languages, learning science, computer security, and accessibility. Observational studies were conducted remotely by the lead author and lasted around 90 minutes.

During the observational study, we first spent 20 minutes asking about their experiences and challenges with peer reviewing. Then, the team asked the participant to write a peer review in a Google document for one of 5 different short papers (less than 4 pages) in 60 minutes. To find suitable short papers, the team first collected the participants' research interest descriptions from their websites and used these keywords to search on Semantic Scholar [3]. We filtered out the papers longer than 5 pages, ranked them by relevance, and selected the top 5 as the participant's options for the study.

In the end, we provided a series of design probes for the novices to capture their reactions. Drawing insights from previous literature around the intelligent support on reading and writing, three researchers on our team took multiple rounds of iteration and group discussions to develop the six design probes [59, 97]. These design problems are visualized in Figma. The design probes included: (1) scaffold annotation with community curated tags and filters, (2) reflection questions (expert-authored generic questions versus AI-generated specific reflection questions on each section or highlighted sentences), (3) extractive summarization and generated explanations to facilitate paper reading, (4) in-situ citation recommendation where the system provides a summary from the cited paper and recommend potential missing citations, (5) review draft generation, (6) mapping back the source of review draft from the paper content and visualize the location of the source to help reviewers revise their draft. The interviews were recorded and then transcribed using a machine transcription service. The research team took the same analysis procedure as the novice study.

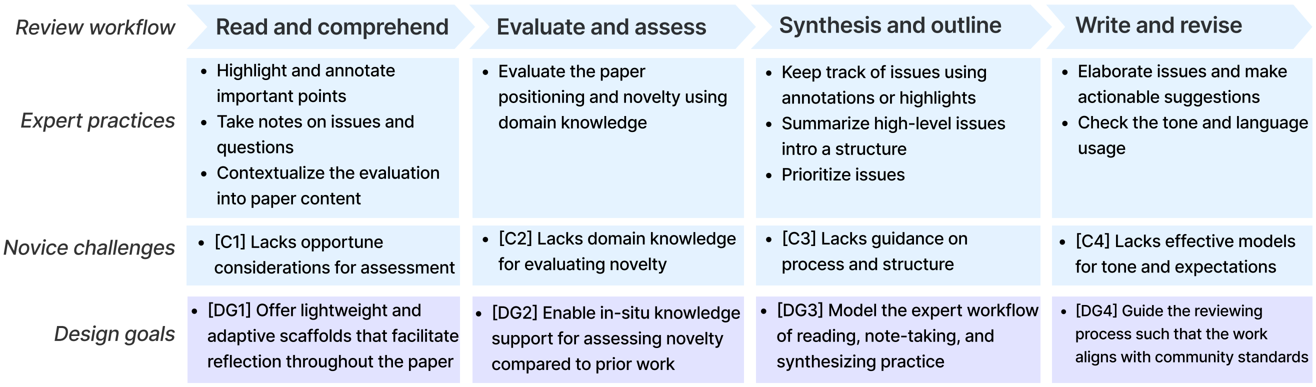
3.2 Findings

3.2.1 Novice reviewers felt they lacked guidance in evaluating the paper and writing structured and constructive reviews. The novice participants reported that their prior attempts at reviewing papers were cognitively demanding and took an average of 6.4 hours. When asked about challenges, 6/10 mentioned that they *struggled to fully understand the background knowledge and existing work and did not feel confident about assessing the paper's novelty (C2).* 4/10 mentioned that they *lack opportune considerations for assessment (C1).* Specifically, one participant mentioned that they would love to have "co-reviewers who have already read the paper and known the specifics to guide me through the evaluation process"(N5). In addition, 4/10 highlighted that they *need more guidance on how to structure issues during the writing process(C3).* 3/10 explicitly mentioned that they want to receive some feedback from experts, other reviewers, or authors, to *make sure the review is in the right tone and meets the expectation (C4).* Specifically, one participant mentioned "I am not sure whether I covered all the necessary points or whether authors will perceive the reviews as useful"[N7]. Another participant is worried about being "impolite or harsh"[N9].

We presented novices with the design scenarios one by one and asked them to rate the degree to which they resonate with each scenario (1 = Not resonate at all to 5 = Strongly resonate with the situation). The top two situations that resonated with novices the most were: "the novice reviewer struggles to write a high-quality review" (3.8/5) and "the novice reviewer spends a lot of time reading a paper but doesn't know how to evaluate the paper" (3.8/5).

3.2.2 Experienced reviewers adopt a workflow of sense-making, annotating, and synthesizing notes. Experienced reviewers reported that they spent 4.75 hours to review one paper. In the observational study with 10 experts, we synthesized the common workflow adapted by experts in the reviewing process. As an initial step, reviewers read and comprehend the paper's content. While reading through the paper, all experts (10/10) highlighted some content, annotated sentences or paragraphs, and took some notes. The format and style of notes varied between different reviewers. We observed some reviewers using symbols while others used phrases or short sentences. After reading and annotating the content, reviewers start to re-read the notes to evaluate the paper's quality. 3 out of 10 experts went back to check the introduction and related work to assess the novelty of the paper.

After finishing reading the paper, instead of directly editing the review draft, experts created high-level headers or topics, such as "lack clarity on study design", that summarized the paper's weaknesses and strengths as well as prioritized the concerns (7/10). For instance, one expert reflected on the review process as "I will make a lot of annotations and then I will just review them section by section. At the same time, thinking about what is the biggest concern in terms of the overall research novelty"(E6). We observed that 7 participants wrote the review in the following structure: summary of the paper, strengths or contribution of the paper, two to three weaknesses of the paper, and end the review with decision justification and general recommendations. Some experts will then list bullet points under each topic together with questions that they would like to ask the author. Last, experts compiled these comments and bullet points into a complete draft and revised them two to three times to offer suggestions and make it more constructive. Figure 2 represents the expert's review workflow together with their practices.



Review workflow	Read and comprehend	Evaluate and assess	Synthesis and outline	Write and revise
Expert practices	• Highlight and annotate important points • Take notes on issues and questions • Contextualize the evaluation into paper content	• Evaluate the paper positioning and novelty using domain knowledge	• Keep track of issues using annotations or highlights • Summarize high-level issues intro a structure • Prioritize issues	• Elaborate issues and make actionable suggestions • Check the tone and language usage
Novice challenges	• [C1] Lacks opportune considerations for assessment	• [C2] Lacks domain knowledge for evaluating novelty	• [C3] Lacks guidance on process and structure	• [C4] Lacks effective models for tone and expectations
Design goals	• [DG1] Offer lightweight and adaptive scaffolds that facilitate reflection throughout the paper	• [DG2] Enable in-situ knowledge support for assessing novelty compared to prior work	• [DG3] Model the expert workflow of reading, note-taking, and synthesizing practice	• [DG4] Guide the reviewing process such that the work aligns with community standards

Figure 2: Experts' workflow in academic peer review, along with experts' practices, novices challenges, and design goals for each stage

3.2.3 Experienced reviewers stressed the importance of specific and contextual guidance for scaffolding. Among these features, 5 experts reflected that the contextual cues can be helpful, especially for novice reviewers. One participant preferred the AI-generated cues on the selected paper content and explained that "I like to have some capability of freedom, but I think this will be super useful for novice reviewers who don't know how to review. But for people who have reviewed for so many years, those guidances for conferences are pretty much the same" [E6]. We provided experts with two sets of tags to select. The first set of tags is designed based on the review structure that includes "summary of the paper", "strength", "weakness" and "others". The second set of tags is designed using community review criteria that include "relevance", "novelty", "validity", "clarity". Most experts preferred the first set of tags (7/10). 2 participants mentioned that they were concerned the experts' authored cues might be too similar to the existing guidelines. Hence, they suggested the contextual question can provide better guidance for novices.

9 out of 10 experts mentioned their preference for the in-situ citation support to provide summary and recommendations. They reflected that this in-situ knowledge support can "raise awareness on unknown work" [E5]. 9 out of 10 participants expressed their concerns about the summarization feature, as they don't trust the AI's ability to identify important information. Instead, they think reviewers should have control over the reading process. Specifically, one expert mentioned "I think we are also reviewing the style of writing, or how something is communicated and logically connected between each paragraph or each sentence. So I think there is value with actually reading everything to get the message behind the paragraphs" [E3]. Participants shared their opinions on using AI in the review process and all of them agreed that LLMs should not generate the review on the fly, and instead, human experts should drive the process since the limitation of LLMs can bias human experts and lead to over-reliance on the use of AI.

Based on the interview on the design probes, we summarize the following design considerations mentioned by novices and experts. The support should be lightweight and not distract from the current review flow. While involving AI in the review process, AI should not go too far to bias or lead the thinking process. Human reviewers should still preserve agency in reading, writing, and decision-making. Faced with the limitations of current LLMs, intelligent systems should try to avoid hallucinations and provide enough opportunities for fact-checking.

3.3 Formative Study Discussion

From the interview study with novices and the observational study with experts, we identified a range of challenges perceived by novices as well as insights on expert practices (see Figure 2). Novices lack confidence in identifying the novelty of research based on prior research and in knowing how to structure a peer review that meets community standards. Experienced reviewers tend to avoid biasing the decision-making and preserve agency in reviewing. The juxtaposition of these novice challenges and the expert practices suggests four core design goals for intelligent scaffolding:

DG1: Offer lightweight and adaptive scaffolds that facilitate reflection throughout the paper Novices highlighted their need for guidance in critically evaluating papers from various perspectives, expressing uncertainty about the key points to focus on (C1). In educational settings, instructors typically serve as scaffolds when introducing new concepts or knowledge to novice learners [19]. Prior studies in scientific paper reading developed methods to offer contextual hints or guided questions that encouraged reflection and critical thinking [6, 71]. For example, Paper Plain provided a collection of key questions that guide readers to answering passages and plain language summaries of those passages [6]. Similarly, in the context of our study, the objective is to provide guidance comparable to that of expert peer reviewers. The emphasis is on delivering locally relevant questions that help users think critically about each section of the paper.

DG2: Enable in-situ knowledge support for assessing novelty compared to prior work Novices struggled with insufficient background knowledge for evaluating papers (C2). Providing knowledge support in situ can externalize the user's working memory, aid in sense-making, and facilitate a swift reviewing and resumption of task contexts [50]. Prior studies in scientific literature review highlight the importance of in-situ knowledge support [13, 44]. To help novice reviewers who lack background knowledge to evaluate the novelty while reading the paper, our goal is to enable in-situ knowledge support by extracting the abstract of the cited paper.

DG3: Model the expert workflow of reading, evaluating, and synthesizing practiced by experienced reviewers Novices expressed their need for more specific guidance from experts in terms of their process and structure (C3). Prior work showed that surfaced expert practices can better structure and scaffold the process for novices [47]. Motify illustrates that storytelling patterns extracted from expert stories can be used to effectively scaffold novices to create video stories [47]. In the context of writing introductory help requests, providing high-quality examples and expert-informed templates can increase learning and writing quality [40]. Our approach involves the modeling of the expert workflow encompassing reading, evaluating, synthesizing, and revising. This model is then employed to structure the peer review process for novice reviewers.

DG4: Guide the reviewing process such that the work aligns with community standards Novices reported that they lacked effective models for tone and expectation (C4). More guidance on the expectation of the review can help reviewers to reflect on their tone and structure. In addition, experts raised valid concerns about the potential for intelligent tools to introduce biases or influence the thinking process. Existing studies have highlighted apprehensions regarding AI exhibiting biases in human-AI collaboration tasks [46, 69]. Notably, research has shown that providing information, including explanations, generated by AI has the potential to mislead users in decision-making [11, 31, 52]. Therefore, a key design objective is to steer clear of biasing the decision-making process and focus solely on encouraging justifications. This approach aims to assist novice reviewers in producing high-quality and original reviews without compromising the integrity of the evaluation process.

4 REVIEWFLOW

We developed ReviewFlow which employed AI-driven scaffolding strategies naturally into the review workflow to support novice reviewers to gain expertise in conference peer reviewing.

4.1 Key features

4.1.1 ***DG1: Contextual cues.*** Researchers used prompts and guided questions to scaffold learners in the paper reading process [15, 68, 95]. Understanding the paper and critically reflecting on the content is essential for decision-making and review writing in the later stage. Prior research showed that providing questions can increase user engagement in actively searching for information. Prior systems mostly used template-based guided questions to engage readers [68]. To *offer lightweight and adaptive scaffolds that facilitate reflection throughout the paper [DG1]*, ReviewFlow provides two types of contextual cues for novices, as shown in Figure 3. One type is **section-level cues guided by community criteria** which are adapted based on each paper section's content together with the review criteria. For section-level reflection cues in Figure 3-A), it takes the entire section, the paper abstract, and the community criteria into account to generate questions. Users can either read these questions before they read the section to obtain contextual guidance or after to reflect on the section's content. Another type is **phrase-level cues adapted to paper content**. As shown in Figure 3-B, users have the freedom to highlight the content that they would like to reflect on at the phrase level and then select the review criteria. Then, it generates cues for participants to reflect on in real-time.

4.1.2 ***DG2: In-situ citation recommendation as knowledge scaffolding.*** Prior research in scientific literature review highlights the importance of in-situ knowledge support [6, 13, 44, 71]. To enable in-situ knowledge scaffolding for assessing novelty compared to prior work, when the user clicks on each reference, ReviewFlow in-situ presents the title and a TLDR summary, as shown in Figure 4-D. The TLDR is queried using Semantic Scholar [3]. To raise awareness of the existing references in the reading workflow, ReviewFlow provides a popup window with potential missing citations from the same venues, as shown in Figure 4-C.

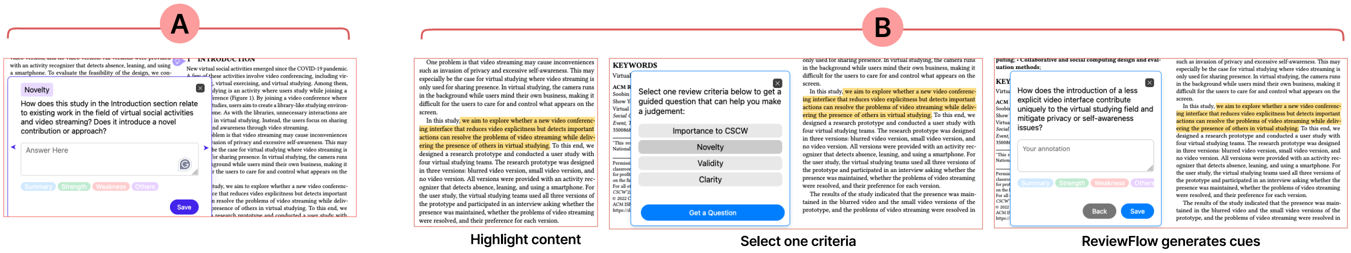
4.1.3 ***DG3: Notes-to-outline synthesis.*** In the review process, we observed expert reviewers synthesize notes to make a plan on how to draft a review. By modeling the cognitive processes of writing proposed by Flower and Hayes, we designed the **notes-to-outline synthesis** feature [22]. As shown in Figure 5, we decomposed the cognitive processes which include planning, translating, and reviewing, and provided scaffolding features for each stage. In the planning stage, to facilitate writers to generate ideas, set goals, and organize information, ReviewFlow provides a notes-to-topic synthesis that summarizes the user notes together with the highlighted text into general topics, such as "needs more detailed citation description". These topics are organized into a structure the aligned with the community practices including summary, strength, and weakness. In the translating stage, to facilitate drafting, ReviewFlow can expand broad topics into detailed bullets if the user clicks the expand button. These bullet points provided a more specific summary based on users' notes. In the reviewing stage, to help users evaluate their written review and revise the text, ReviewFlow pops up a self-reflection card and asks users to self-reflect on their written review using the review criteria, including tone, comprehensive, constructive, justified, and accurate. This encourages users to revise the text if they would like to improve the review quality.

4.1.4 ***DG4: Fact-checking between outline and source notes and self-reflections.*** When the research team introduced the idea of using AI to facilitate the reading and writing process, participants expressed the need for fact-checking and providing explanations. To avoid biasing the decision to either accept or reject the paper but encourage justifications, ReviewFlow provided the feature that for each synthesized outline bullet, when the reviewers click it, the notes in the middle column will be highlighted. The PDF on the left column will also scroll automatically and highlight the corresponding location of the notes, as shown in Figure 6.

4.2 System implementation

ReviewFlow's front-end is a React web application built on top of an existing PDF highlighting library [2]. The front end is responsible for displaying the PDF, managing the annotation data, and displaying the text input used to write the review draft. The back-end uses a

[2] https://github.com/agentcooper/react-pdf-highlighter



Novelty

How does this study in the Introduction section relate to existing work in the field of virtual social activities and video streaming? Does it introduce a novel contribution or approach?

Answer Here

Select one review criteria below to get a guided question that can help you make a judgement:

Importance to CSCW

Novelty

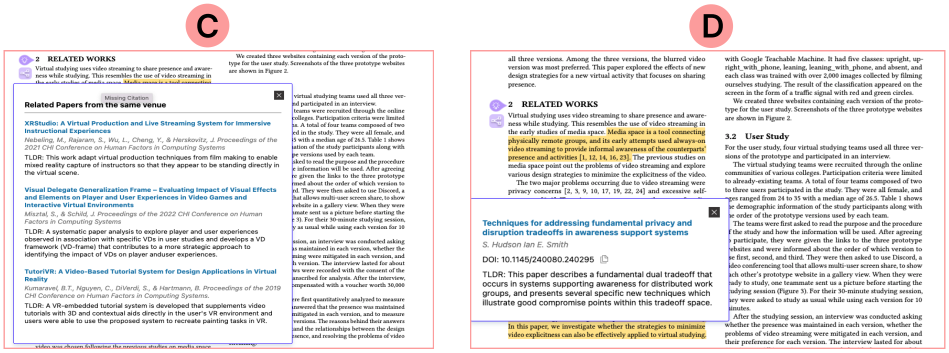
Validity

Clarity

How does the introduction of a less explicit video interface contribute uniquely to the virtual studying field and mitigate privacy or self-awareness issues?

Your annotation

Figure 3: Contextual cues. ReviewFlow provides (A) section-level cues guided by community criteria and (B) phrase-level cues adapted to user highlighted content and selected criteria.



Related Papers from the same venue

XRStudio: A Virtual Production and Live Streaming System for Immersive Instructional Experiences

Nebeling, M., Rajaram, S., Wu, L., Cheng, Y., & Herskovitz, J. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems

TLDR: This work adapt virtual production techniques from film making to enable mixed reality capture of instructors so that they appear to be standing directly in the virtual scene.

Visual Delegate Generalization Frame – Evaluating Impact of Visual Effects and Elements on Player and User Experiences in Video Games and Interactive Virtual Environments

Misztal, S., & Schild, J. Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems

TLDR: A systematic paper analysis to explore player and user experiences observed in association with specific VDs in user studies and develops a VD framework (VD-frame) that contributes to a more strategic approach to identifying the impact of VDs on player anduser experiences.

TutoriVR: A Video-Based Tutorial System for Design Applications in Virtual Reality

Kumaravel, B.T., Nguyen, C., DiVerdi, S., & Hartmann, B. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.

TLDR: A VR-embedded tutorial system is developed that supplements video tutorials with 3D and contextual aids directly in the user's VR environment and users were able to use the proposed system to recreate painting tasks in VR.

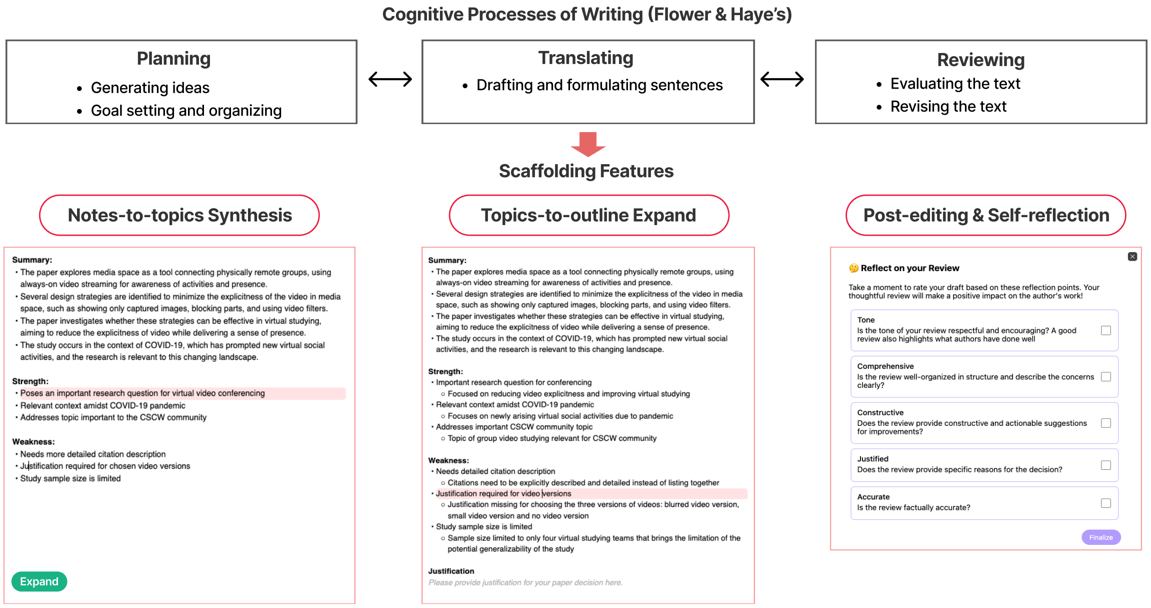
Techniques for addressing fundamental privacy and disruption tradeoffs in awareness support systems

S. Hudson Ian E. Smith

DOI: 10.1145/240080.240295

TLDR: This paper describes a fundamental dual tradeoff that occurs in systems supporting awareness for distributed work groups, and presents several specific new techniques which illustrate good compromise points within this tradeoff space.

Figure 4: In-situ knowledge scaffolding. ReviewFlow provides (C) a list of relevant papers from the same venue that are not cited and (D) an in-situ citation summary including title, author, and the TLDR summary.


Cognitive Processes of Writing (Flower & Haye's)
Planning
Generating ideas
Goal setting and organizing
Translating
Drafting and formulating sentences
Reviewing
Evaluating the text
Revising the text
Scaffolding Features
Notes-to-topics Synthesis
Topics-to-outline Expand
Post-editing & Self-reflection
Expand
Finalize

Summary:
- The paper explores media space as a tool connecting physically remote groups, using always-on video streaming for awareness of activities and presence.
- Several design strategies are identified to minimize the explicitness of the video in media space, such as showing only captured images, blocking parts, and using video filters.
- The paper investigates whether these strategies can be effective in virtual studying, aiming to reduce the explicitness of video while delivering a sense of presence.
- The study occurs in the context of COVID-19, which has prompted new virtual social activities, and the research is relevant to this changing landscape.

Strength:
- Poses an important research question for virtual video conferencing
- Relevant context amidst COVID-19 pandemic
- Addresses topic important to the CSCW community

Weakness:
- Needs more detailed citation description
- Justification required for chosen video versions
- Study sample size is limited

Summary:
- The paper explores media space as a tool connecting physically remote groups, using always-on video streaming for awareness of activities and presence.
- Several design strategies are identified to minimize the explicitness of the video in media space, such as showing only captured images, blocking parts, and using video filters.
- The paper investigates whether these strategies can be effective in virtual studying, aiming to reduce the explicitness of video while delivering a sense of presence.
- The study occurs in the context of COVID-19, which has prompted new virtual social activities, and the research is relevant to this changing landscape.

Strength:
- Important research question for conferencing
  - Focused on reducing video explicitness and improving virtual studying
- Relevant context amidst COVID-19 pandemic
  - Focuses on newly arising virtual social activities due to pandemic
- Addresses important CSCW community topic
  - Topic of group video studying relevant for CSCW community

Weakness:
- Needs detailed citation description
  - Citations need to be explicitly described and detailed instead of listing together
- Justification required for video versions
  - Justification missing for choosing the three versions of videos: blurred video version, small video version and no video version
- Study sample size is limited
  - Sample size limited to only four virtual studying teams that brings the limitation of the potential generalizability of the study

Justification

Please provide justification for your paper decision here.

🤔 **Reflect on your Review**

Take a moment to rate your draft based on these reflection points. Your thoughtful review will make a positive impact on the author's work!

Tone
Is the tone of your review respectful and encouraging? A good review also highlights what authors have done well ☐

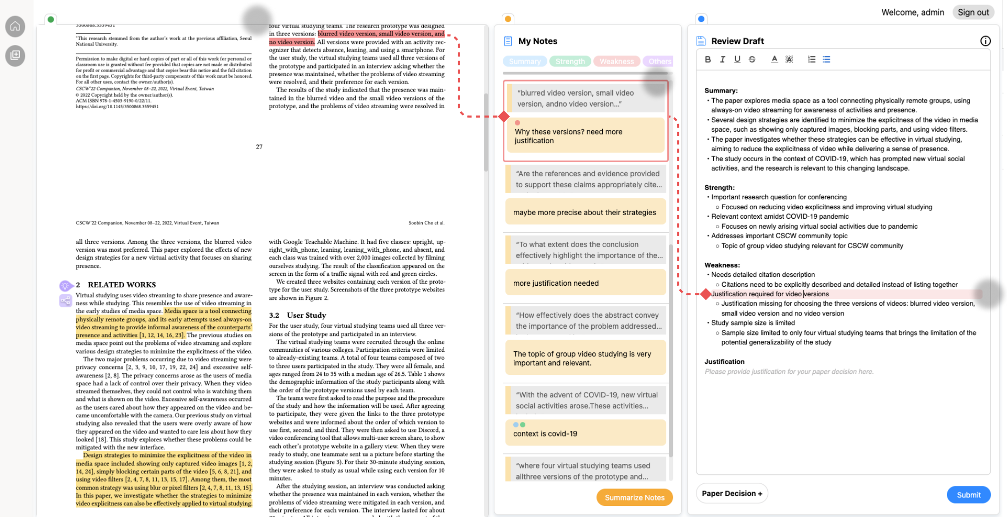
Comprehensive
Is the review well-organized in structure and describe the concerns clearly? ☐

Constructive
Does the review provide constructive and actionable suggestions for improvements? ☐

Justified
Does the review provide specific reasons for the decision? ☐

Accurate
Is the review factually accurate? ☐

Figure 5: Scaffolding features support the review writing process. The top part shows Flower and Hayes cognitive process of writing [22]. The bottom part shows the ReviewFlow features that support each writing step. On the left, ReviewFlow summarized notes into broad topics under strengths and weaknesses to facilitate planning. In the middle, use can click to expand the topics to a detailed outline. On the right, the pop up window encourages self-reflection and post-editing.



"blurred video version, small video version, andno video version..."

Why these versions? need more justification

"Are the references and evidence provided to support these claims appropriately cite...

maybe more precise about their strategies

"To what extent does the conclusion effectively highlight the importance of the...

more justification needed

"How effectively does the abstract convey the importance of the problem addressed...

The topic of group video studying is very important and relevant.

"With the advent of COVID-19, new virtual social activities arose.These activities...

context is covid-19

"where four virtual studying teams used allthree versions of the prototype and...

Summary:

- The paper explores media space as a tool connecting physically remote groups, using always-on video streaming for awareness of activities and presence.
- Several design strategies are identified to minimize the explicitness of the video in media space, such as showing only captured images, blocking parts, and using video filters.
- The paper investigates whether these strategies can be effective in virtual studying, aiming to reduce the explicitness of video while delivering a sense of presence.
- The study occurs in the context of COVID-19, which has prompted new virtual social activities, and the research is relevant to this changing landscape.

Strength:

- Important research question for conferencing
  - Focused on reducing video explicitness and improving virtual studying
- Relevant context amidst COVID-19 pandemic
  - Focuses on newly arising virtual social activities due to pandemic
- Addresses important CSCW community topic
  - Topic of group video studying relevant for CSCW community

Weakness:

- Needs detailed citation description
  - Citations need to be explicitly described and detailed instead of listing together
- Justification required for video versions
  - Justification missing for choosing the three versions of videos: blurred video version, small video version and no video version
- Study sample size is limited
  - Sample size limited to only four virtual studying teams that brings the limitation of the potential generalizability of the study

Justification

Please provide justification for your paper decision here.

Figure 6: When the participants click on each outline element, ReviewFlow shows the visual mapping among the summarized outline bullet (a), the note in the middle (b) and original pdf with highlights (c).

Flask server which handles the GPT-4 [3] API endpoints, MongoDB endpoints for storing data, and GROBID [1] for PDF data extraction. This data includes all of the parsed text of the PDF, its PDF coordinates, and citations linked to their references. Detailed prompts are provided in the Appendix.

4.2.1 Contextual cues. After the user uploads a paper PDF on the front end, a request is sent to ReviewFlow's back-end server where GROBID is used to parse and extract the paper's content, creating an XML/TEI structured document with the coordinates and content of the section titles, text body sentences, and inline citations. This extracted data is then sent to the front-end client where ReviewFlow combines GROBID's section data with PDF.js's section data to make GPT-4 API calls to generate the contextual cues for each of the paper's sections.

Each GPT-4 API call utilizes the section's text as a prompt to generate a contextual reflection question for each of the following critical review aspects: importance, novelty, validity, and clarity. The GPT-4 response is formatted as a JSON and is streamed to the front-end client. To generate phrase-level cues, when the user highlights text on the PDF and clicks on the button with the light bulb icon, a pop-up appears with the review criteria. When the user selects one of the aspects and clicks "Get a Question", GPT-4 will use the selected aspect, highlighted phrases, paragraph of the highlighted phrases, and paper abstract to generate cues.

4.2.2 In-situ citation recommendation. When the PDF is initialized, ReviewFlow uses GROBID's XML/TEI data to create a citation layer that overlays all in-line citations. When the user clicks on an in-text citation, a popup shows the paper title, publication date, DOI link, and a short description that is searched using Semantic Scholar API [3]. To recommend corresponding papers, the system calls the Semantic Scholar Recommendation API[4] and uses the keywords of the paper together with the venue to retrieve the most similar paper. After filtering out the retrieved papers that have already been included in the current paper, the top three papers are added as a citation pop-up.

4.2.3 Notes-to-outline Synthesis. After the user creates multiple notes, a "Summarize Notes" button will appear at the bottom of the notes panel. Upon clicking this button, the front-end client will organize all the user's notes data including the corresponding highlighted paper content, selected tags, note text, and paper abstract. Subsequently, the system will use this data as prompt. We use the few-shot learning paradigm with the corresponding prompts [58]: "Please create three important topics on the paper's [strengths] and [weaknesses] with less than ten words that combine and summarize the user notes." The output is then formatted in JSON and streamed directly into the front-end's text-editable draft panel with a summary and topics organized on strengths and weaknesses. When the user clicks on the "Expand" button at the bottom of the draft panel, all of the text in the draft text input box, notes, and paper abstract is sent to the back-end server and requested a GPT-4 API call. The GPT-4 API call will then respond with a streamed JSON-formatted output with more details based on the user's notes and topics specifically for the strength and weakness sections.

5 METHOD

We designed a within-subjects experiment to answer the questions below:

- RQ1: How does ReviewFlow affect participants' final written review quality, compared to the same system with no intelligent scaffolding?
- RQ2: How does ReviewFlow affect participants' workflow, in terms of time and engagement?
- RQ3: What benefits and challenges do participants perceive with ReviewFlow's intelligent scaffolding?

[3] https://openai.com/research/gpt-4
[4] https://api.semanticscholar.org/api-docs/recommendations

5.1 Study Design

We conducted a within-subjects experiment with 16 novice participants where each participant experienced both the ReviewFlow condition and the Baseline condition in two sessions separately. We counterbalanced the order of two conditions and papers using a Latin Square design to minimize the potential order effects. To reduce knowledge transfer and minimize fatigue, we scheduled the two study sessions of each participant at least 24 hours apart. The ReviewFlow condition includes all scaffolding features, while the Baseline condition includes the minimal guideline. In the baseline interface, users can still highlight, take notes, and tag notes. Community guideline is also provided.

5.1.1 Participants. Before the study, we sent out a pre-study survey to participants to collect information about their expertise, research topics, review experience, knowledge level on AI, and demographics (e.g. age, gender, race). We recruited 16 participants who have experience in conducting academic research for at least two years and have zero to twice conference peer review experience to make sure that they are novice reviewers who lack review experience but are equipped with domain knowledge. We advertised recruitment messages to colleagues, mailing lists, communication channels, and social media and recruited participants from four universities across the US. Using a snowball sampling approach, we asked participants to refer their friends and colleagues. Participants have on average 1.2 years of writing and submitting academic papers. Each study section takes around 90 minutes and all participants were compensated for $20 per hour. The study is IRB-approved.

5.1.2 Procedure. Before the user study, participants filled out a pre-survey that captured their previous experience in reviewing and reviewing papers on HCI/CSCW conferences and their expertise in HCI/CSCW. We asked participants whether they had read the paper before to make sure they all were seeing the work for the first time. Participants conducted reviews on two papers using different versions of our interface (ReviewFlow or Baseline). To counterbalance the order effect, we randomized the order of the control condition and the experiment condition for each participant, so half participants encountered ReviewFlow in their first session, and the other half experienced it in their second session. For each session, we followed the review process used at the CSCW conference, where we provided the paper draft and the review guidelines.

For both sessions, participants were told to spend around 60 minutes – and no less than 30 minutes – on the review, but they could take as much time as needed. In the pilot study with 2 participants who had very little review experience, we found that participants could finish reviewing within 45 minutes. Before each condition session, we gave the participants a quick 2-minute demo of the interface, while in the experiment session demo, we also introduced the functionality of ReviewFlow. After each session, the research team asked the participants to fill out a post-survey to evaluate the system and assess their self-efficacy. After the second session, we conducted a 15-minute semi-structured post-interview to ask open-ended questions about their experience, perceptions, and feedback. The post-interviews were video recorded with participants' permission and were transcribed into text for later analysis.

5.1.3 Paper Selection Process. We collected recent one-year papers from HCI-related conferences including CHI, CSCW, UIST, UBICOMP, IUI, DIS using the list provided by [5]. To make our study time short so that people have the energy to finish the task, we filtered papers that had fewer than 3000 words and further filtered out papers that had technical terms and jargon. The two papers we selected need to have similar lengths, similar difficulties, and from the same conference venues. Combining all the criteria above, two papers from CSCW Companion were selected for the study. The first paper includes the keywords "virtual studying, video streaming, awareness" and contains 2853 words. The second paper includes the keywords "virtual environment, cross-lingual collaboration, team formation" and contains 2910 words.

5.1.4 Measures. We collected a mix of quantitative and qualitative data, including each participant's log data that captured their interactive behaviors with the system, the final review written for each paper (N=32), the post-survey, and the interview transcripts. The research team analyzed these combined sources of data to reveal insights.

Quality Ratings on the Final Peer Review . To measure the quality of the review, we recruited two experts who have conducted research for more than three years of review experience in HCI or CSCW conferences. They counted the number of strengths and weakness in the review (a proxy for coverage) and rated the quality of all final reviews (N=32) with a five-dimension rubric based on reviewer guidelines for CSCW conference and previous research [96]:

- Tone: The tone of a peer review is always encouraging and respectful. A good review also highlights what authors have done well.
- Comprehensive: A good review is always well-organized in structure, which includes a summary, strengths, weaknesses, and a clear description of concerns.
- Constructive: A good review usually provides constructive suggestions. Following the weakness, reviewers usually will provide actionable items that the author can work on to improve the paper's quality.
- Justified: A good review justifies specific reasons for their decisions. Avoid providing a decision without any supporting evidence.

Two experts rated these five dimensions on a simple seven-point scale (1-7). Each expert first read one paper and wrote a peer review of the paper. After this process, the research team provided provided instructions and two examples for them to rate and discuss until they reached a consensus on ratings based on the instructions. Then, they rated each dimension of the review on the paper independently. The inter-rater reliability between two experts on all 32 data is moderate where Krippendorff's alpha is higher than 0.50 [49]. The research team then used the average scores by two experts for each dimension.

To construct the proxy of review quality, we further asked the experts to count the number of strengths and weaknesses raised by the participants in each review. Another proxy of the review quality we used is the participants' self-rated satisfaction with their reviews. After participants submitted the reviews, we asked them

[5] http://www.conferenceranks.com/

to rate their satisfaction with their reviews on a scale of 1 -7 (1 as not satisfied at all and 7 as very satisfied).

User Interaction Data. To measure participants' interaction with the tool, we instrumented the interface to collect a range of user activity log data. We collected two timing measures – how long each participant took to finish the review session and how long each participant spent editing the review within the text box. The entire review session time includes reading time and writing time. The ReviewFlow interface also collected interaction data to capture how much each participant interacted with each scaffolding feature, including the times they answered the contextual cues, clicked on the citation summary, checked the citation pop-ups and used the outline summarization feature.

User Preferences and Reactions. To evaluate users' preferences for the ReviewFlow experience compared to a baseline plain review editor, we asked participants to fill out a short post-study survey. The survey asked participants to directly compare the perceived usefulness, enjoyment, easiness, and sense of control between the ReviewFlow and baseline system. We collected the level of cognitive demand using NASA TLX on a scale of 1-5 [34]. We further collected the feeling of control and collaboration on a scale of 1-5. Then we specially asked participants to evaluate each of the scaffolding features. The survey collected their 5-point Likert scale ratings on perceived usefulness, and perceived accuracy for each feature.

Previous research showed that scaffolding can promote learners' self-efficacy [92]. Here, we measured whether using the scaffold system can improve novice reviewers' self-efficacy and confidence. We asked each participant to report self-efficacy after using the ReviewFlow system and the Baseline system by answering the question "How confident are you in your ability to write a conference peer review next time after using the system (1 as not confident and 7 as very confident)".

After the post-survey, we conducted a 15-minute semi-structured interview with all participants to capture their overall thoughts as well as specific perceptions of machine-generated highlights and summaries. For example, the research team asked "What do you think of the difference between the task with and without the support of ReviewFlow", "What did you learn after writing the review with the system", and "What concerns did you have when using ReviewFlow?".

Users' knowledge of each paper's topic as control variables. We measured participants' existing knowledge of the two papers as a control variable. Participants described their knowledge of each paper's topic from not familiar as 1 to very familiar as 7 in the post-survey. All participants reported that they had never read and remembered the papers before. The average rating of knowledge level on the first paper is 3.46 and the average rating for the second paper is 3.42.

5.2 Analysis

Quantitative data analysis. To measure the effect of the ReviewFlow system on each dimension of the review quality (eg. tone, comprehensive, constructive and justified), we conducted repeated measure ANCOVA tests. We used paper ID, the order of experiment conditions (whether ReviewFlow was used for the participant's first or second paper), the knowledge level of each paper topic as co-variants. To measure the effects of the experiment condition on the time they spent reading independent reviews and writing reviews taking the consideration the differences between the paper topic and order effects, we ran a repeated measure ANCOVA using the paper ID, the order of experiment condition and the knowledge level of each paper topic and the review word counts as co-variants. To compare participants' behaviors between the two conditions, we conducted Wilcoxon signed-rank test, which is a non-parametric rank test, on users' interaction data, e.g. number of notes participant taken in both conditions [88].

Qualitative data analysis. All semi-structured interviews with participants are recorded and transcribed. Two researchers conducted iterative open coding on the transcripts using Dovetail[6] and conducted thematic analysis on the transcripts [10]. They open-coded the data by identifying topics mentioned by the participants. Initial codes were combined into preliminary themes, which were discussed among the research team. After iteratively discussing the code themes, researchers identified the final themes around: participants' reactions to each feature, their overall perceptions of the ReviewFlow system and their reactions to each scaffolding feature.

6 RESULTS

We report our results from the within-subjects experiment and the post-interview. In the within-subject experiment, across both conditions, participants spent an average of 38.7 minutes writing a review with an average length of 243 words written. 60% of the participants decided to accept the paper, which is consistent with the original decisions for the two papers. Our findings revealed that ReviewFlow provided more guidance to novice reviewers and made the review process more useful and engaging.

6.1 RQ1: ReviewFlow helped participants write more comprehensive reviews

To compare the quality of written review in both conditions, we performed ANCOVA tests to examine the effect of the two conditions on each quality measure, accounting for the review length, participants' knowledge of the topic, the order of the experiment condition as co-variants. As shown in Figure 7, we found that the reviews in the ReviewFlow condition are significantly more comprehensive than the Baseline condition ($p = 0.04*$, $F = 3.55$). We found no significant interaction effect between the order of the experiment conditions and no statistically significant interaction effect between the knowledge of the two papers.

Furthermore, we evaluated several proxies of review quality including the number of strengths and weaknesses and participants' self-rated satisfaction with the written review. As shown in Table 1, participants wrote longer reviews and called out more strengths and weaknesses in the paper in the ReviewFlow condition. However, there is no significant improvement in the constructiveness of the review. This indicates that the ReviewFlow system can scaffold participants to capture more pros and cons but still cannot help participants write constructive solutions for each weakness. Also,

[6] https://dovetailapp.com/

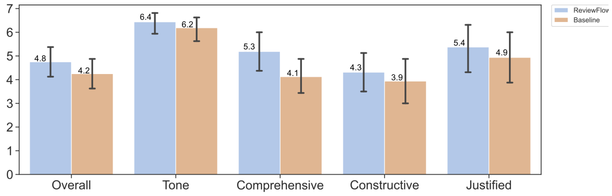


Figure 7: Experts' evaluation scores on the quality of review written by ReviewFlow versus Baseline.

participants' self-rated satisfaction with reviews written using the ReviewFlow tends to be higher than using the Baseline. ANCOVA test with proxies did not a show significant effect. Participants reflected on the reason that "I captured more pros and cons since I took more notes this time, so I feel satisfied with the review as it covered more aspects" [P3].

	ReviewFlow	Baseline	p	F
Proxy of review quality				
Count of strengths	2.38 (0.19)	1.91 (0.17)	-	-
Count of weaknesses	2.62 (0.22)	2.08 (0.27)	-	-
Self-rated satisfaction	4.35 (0.4)	4.00 (0.34)	-	-
Length of the review				
Word counts	250.8 (19.4)	235.8 (23.5)	-	-
Time on task				
Reading time (minutes)	26.9 (1.9)	19.2 (2.1)	***	19.2
Writing time (minutes)	15.5 (2.2)	15.7 (2.4)	-	-
Self-efficacy				
Self-efficacy on reviewing	4.92 (1.43)	3.92 (1.36)	*	4.2

Table 1: Proxies of review quality include the number of strengths, number of weaknesses, and satisfaction on the written review rated by participants themselves. Participants included more weaknesses in the ReviewFlow condition. Participants spent more time reading and reported higher self-efficacy after using the ReviewFlow.

6.2 RQ2: Longer interaction duration with ReviewFlow but improved participants' self-efficacy

6.2.1 Participants took a longer time to read papers with ReviewFlow but a shorter time on drafting the final review. We performed ANCOVA to examine the effect of the two conditions on writing time, accounting for the length of the review, participants' knowledge of the topic, and the order of the experiment condition as covariates. As shown in Table 1, participants spent significantly more time in reading and sense-making with the ReviewFlow than with the Baseline. However, they spent less time on writing, while the result was not significant. 68.8% of them still reported that "ReviewFlow saved me more time on writing reviews". P11 reflected the reasons that "Answering pop-up questions made me spend more time on reading the paper, judging it from different aspects, and taking notes, but I feel it did save me time in the end since I don't need to go through all the notes again"[P11]. Similarly, P6 also mentioned the reason for saving time as "having that[my notes] summarized and organized at the end meant that I didn't have to go back through each section and think about where I should put the strengths and weaknesses"[P6]. On the contrary, P5 reflected that they spent more time considering different aspects, such as validity, novelty, clarity, etc. P5 highlighted that "I can write the review fast, but that is not my goal. As a reviewer, it is more important to spend enough time carefully evaluating the paper"[P5].

6.2.2 Participants reported to have higher self-efficacy after experiencing ReviewFlow than the Baseline. We asked participants to rate their self-efficacy on the ability to conduct a conference peer review and conducted the ANCOVA test between each participant's ratings, accounting for covariates. Table 1 showed that self-efficacy ratings in the ReviewFlow are significantly higher than the Baseline. This result indicated that participants built more confidence in learning after using the ReviewFlow. Specifically, participants described their learning process –"I feel I gradually built confidence. At the beginning, I would try to answer every question. But later on, I started to remember which aspect I should think about while reading that section, so I didn't need to frequently check guided questions"[P13].

6.2.3 Participants took more notes in the ReviewFlow and used the features actively. Participants took significantly more notes in the ReviewFlow condition (M=12.9, STD = 3.7) than in the Baseline condition (M = 9.1, STD = 3.9, Wilcoxon signed-rank test, Z= 103.0, p <= 0.01, ∗∗). This indicates that participants are more engaged in the reading process with ReviewFlow and not only reading text but also critically reflecting on the content. Participants actively answered the guided questions. All participants used the notes-to-outline features which helped them summarize these notes into an outline. Three participants did not expand the high-level outline to a detailed outline.

6.3 RQ3: Participants perceived ReviewFlow as useful in the review workflow

After participants had experienced the two systems, we asked them to compare two conditions. As shown in Figure 8, participants highly preferred ReviewFlow over the Baseline. 93.7% of participants perceived the interface as enjoyable to use. Participants mentioned that the scaffolding features make the review process more engaging and less boring (N=5). All participants think that the system is useful and they like the guidance provided in reviewing. For instance, P11 described the guidance they received as similar to experts – "I feel like some experts, such as my advisor, were sitting next to me to prompt me in particular ways and show me how to write the review"[P11].

We further measured participants' perceptions of the tasks in each condition. In the post-survey, we asked participants to rate cognitive workload, such as distraction and engagement, the feeling of control and collaboration with the interface, and perceived learning gains. As shown in Figure 9, 75% of the participants reported that they are engaged in the process, but the add-on scaffolding features bring in some distraction for 68.8% of the participants. However, participants highlighted that it is not a bad distraction, but served as a staging process that motivated them to think. For

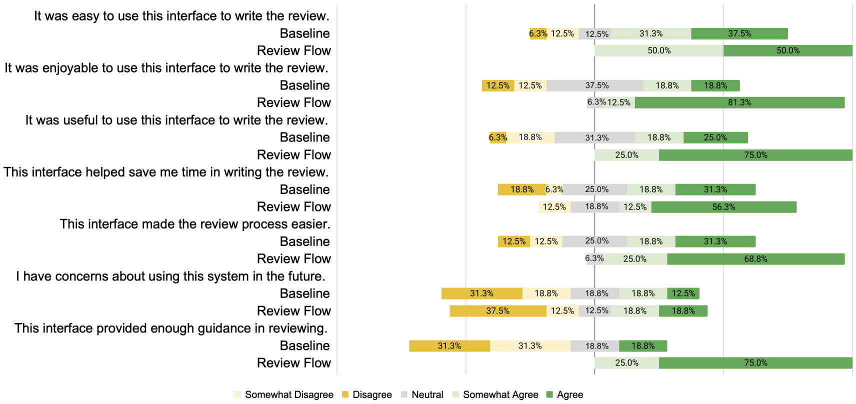


Figure 8: Participants' reactions to the system in two conditions in terms of easiness, enjoyment, and efficiency.

instance, P8 explained that "Since you don't want to just keep going through the paper without getting anything from it. The distraction is like a pop-up that keeps prompting you about review criteria from different places"[P8].

All participants believe that the ReviewFlow helped them learn the peer review process. 68.8% of participants have the feeling of collaborating with the interface. Participants identified the collaboration mainly happened during the process of answering contextual cues in each section and using the note-to-outline synthesis feature. The two-step process where participants can first summarize notes into a high-level outline and then expand it provided them the feeling of "iterating with an assistant"[P9]. Even though the current system uses AI-generated outlines, all participants agreed that they still have the control over writing process. One participant mentioned the reason is that "It just synthesized what I have already written in the notes and I can still write it by myself"[P9]. 37.5% of the participants described that the notes synthesis did not bias them in the current session, but they still worried that "busy reviewers might randomly create notes and use the outline to make a decision"[P14].

6.3.1 Participants have different preferences on the scaffolding strategies. As shown in Table 2, participants perceived that section-level cues were more useful and more accurate than the phrase-level cues on the highlighted text. Two participants reflected that the uncertainty on the phrase-level cues is high since the questions' quality depends on the number of words they highlighted. P15 also said "I have already had a question in my mind when I highlighted certain parts of the text. So if the question did not match with the question I was thinking, I felt a bit disappointed"[P15]. We also found that many people did not pay attention to the missing citation support, since they did not focus on evaluating the related work section. The most useful feature perceived by participants is the notes-to-outline synthesis feature where participants reflected that it helped them "easily get a sense on top of my notes"[P3]. However, three participants did not expand the topic to a detailed outline. They reflected that the detailed outline bullets contained similar content as the high-level outline topic which made it less useful. We further conducted an in-depth analysis of how participants used and perceived different types of scaffolding strategies:

	#(p)	freq	useful	accuracy
Section-level cues	14	3.8	5.4	4.9
phrase-level cues	10	1.6	3.8	4.3
In-situ citation pop-up	10	1.0	4.6	6.2
Missing citation pop-up	9	1.0	2.5	3.4
Summarize notes to high-level topics	16	1.12	6.3	5.2
Expanded topics to detailed outline	13	1	5.2	4.3

Table 2: Usage of scaffolding features shows the number of participants, average frequency, and participants' rating on the usefulness and accuracy. All participants used the features that synthesize notes to high-level topics, while three participants did not expand it to a detailed outline.

Contextual cues guided novices to evaluate from different criteria. Participants described that the contextual questions are "more specific and tailored to the paper compared with the general community instruction"(N=3). Instead of providing the review criteria with general questions, such as "how well does the paper execute their contribution?", participants preferred to have more contextually guided questions according to the paper content, such as "does the method section provide clear and well-justified explanations of the research prototype's design including the choice of videos and the method to set up the activity recognizer?". Participants described that their process of using the section-level cues is slightly different from the phrase-level cues. P10 described that "I quickly checked out the questions next to each section before I dived into the reading. After I finished reading the section, I came back to these questions and reflected on how I felt"[P10]. Correspondingly, phrase-level questions are used when "I feel confused or not sure what I should think about"[P11]. These contextual cues reminded them or prompted them to evaluate different aspects of each section.

Notes-to-outline synthesis helped participants structure their notes according to community practices. All participants used the notes-to-outline synthesis. Participants preferred having notes in a "structured version" and indicated that they didn't need to summarize

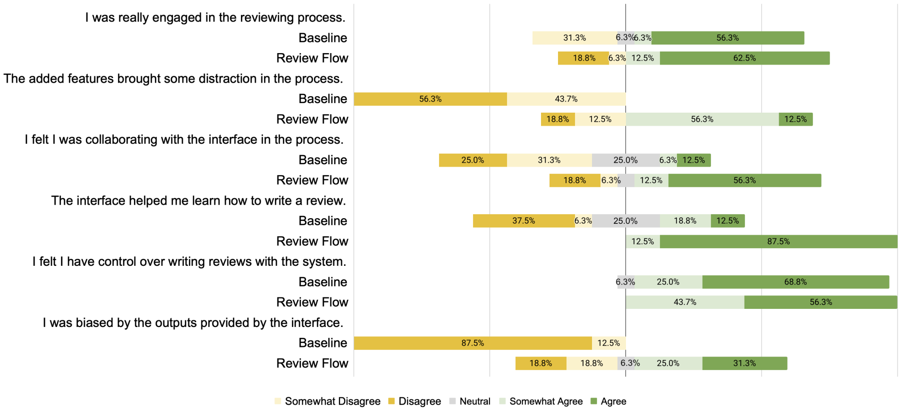


Figure 9: Participants' feelings of the engagement, control, and perceptions of bias in two conditions.

and map their notes by hand. The ability to summarize notes into each review section helped them "reflect on the strengths and weaknesses" (N=2). As a result, they perceived that the outline generation feature saved their time in writing (N=5). Participants liked the fact-checking function that when they can map each topic back to their notes and original pdf, as shown in Figure 6. The visual mapping helped them cross-check the notes and increased their trust in the topic. However, three participants intentionally avoided using the expand button to get a more detailed bullet. They are concerned that the expanded outline may generate content from nowhere. P3 explained that "I was worried that if I clicked to expand the themes into an outline, it would have incorporated things that I didn't want to include"[P3].

6.3.2 *Participants' concerns about using ReviewFlow.*

Participants are mostly concerned about the potential errors that AI can make in the notes synthesis process. Participants did not express severe concerns in the study session since the generated outline mainly reused or summarized their notes. However, they were still worried that the nuanced tone in the outline produced by AI may exaggerate the weakness of the paper and influence reviewers' perceptions. For instance, one participant mentioned that if you wrote the notes as "the details around study participants seems a little bit unclear", but the AI turns it into "this paper lack clarity", which may mislead the reviewer. Moreover, participants expressed their concerns of other reviewers in the community. They are worried that "last minute reviewers may randomly leave notes and then use the tool to generate an outline. If they expand the notes it might not be truly reflective of what they thought if they were not paying attention. The misuse of the tool will not be fair to the paper authors and the research community in general"[P14].

7 DISCUSSION

Advances in large language models are changing how work gets done. Our research explores how we might integrate intelligent scaffolding in a way that informs and guides novices and steers away from biasing the underlying judgment. To inform the design of intelligent scaffolding, we conducted a formative study where we learned what strategies and practices experts adopt, as well as, what challenges novices face when writing peer reviews. Our ReviewFlow system aimed to leverage LLMs to provide timely considerations while reading and assessing a submission and structural support when composing a review. Our within-subject experiment (N=16 participants) found that novice reviewers not only preferred ReviewFlow over the baseline system, but they also wrote longer and more comprehensive reviews, as judged by experts. Using ReviewFlow participants spent more time reading the paper and took more detailed notes. Those notes were aided by ReviewFlow's contextual cues, and helped participants call out more strengths and weaknesses. Using ReviewFlow, participants were more satisfied with their reviews (according to self-ratings) and attributed this to the timely cues and a useful workflow.

In comparison with general peer review practices [76, 85], participants using ReviewFlow allocated more time to reflect on each section and evaluate the paper's quality. Interestingly, 40% of participants opted for rejection decisions in both conditions, contrary to the original decisions for the two papers. A typical peer review not only contains the paper summary and its contribution but also raises weaknesses from different aspects together with constructive feedback or thought-provoking questions [63]. However, participants in both conditions still found it hard to provide constructive feedback on each weakness, even with example reviews. Perhaps future systems could incorporate additional features into the intelligence scaffolding, such as presenting examples of improvement suggestions from similar prior reviews or offering feedback to the reviewers after they have an initial draft [48].

7.1 What role should machine intelligence play in this review process?

AI has come a long way since the early and annoying attempts at supporting work (e.g. Clippy). With the development of LLMs, AI can automate or augment many aspects of knowledge work, including information discovery, sensemaking, and writing [57, 66, 97]. While LLMs have become incredibly valuable, the risk now might the risk now might be the tendency, especially among tech-focused innovators, of taking automation too far and dehumanizing work. Prior research has explored whether models can predict a paper decision or even draft a review [55]. Our ReviewFlow system was designed with the intention of balancing the use of technology with human values. Our system strikes this balance by taking cues from the learning sciences research on scaffolding [36, 75]. The goal of ReviewFlow is not to produce reviews, per se, but to convey an understanding of how to think while reading a submission and writing a review. Our values place more emphasis on training and preparing novices (perhaps for their next peer review), not just on getting to a final peer review.

Prior work indicates that well-design scaffolding can help novices operate more like experts [94]. The concept of Zone of Proximal Development (ZPD) represents the space between what a learner is capable of doing unsupported and what the learner cannot do even with support [33, 86]. Scaffolding works most effectively when it meets novices when they need support and tapering off when they have internalized the best practices. The scaffolding in ReviewFlow is intelligent and contextual. For instance, users get reflection cues that adapt specifically to the current section.

ReviewFlow was explicitly designed to support novices, but it does not adapt to someone's existing knowledge or experience with the task. A longitudinal deployment would provide insight on whether novices continue to prefer ReviewFlow, or whether it is useful for early experiences with peer reviewing. As novices become more proficient, they may eventually prefer to work with less structure (e.g. traditional text editor), although even old timers likely find value in ReviewFlow's pragmatic support for capturing and synthesizing notes. Future research could extend the current system to be more context-aware and more adaptive to users and then study its use with both novices and experienced reviewers over time to gain design insights for interactive scaffolding. Furthermore, future work could explore other properties of scaffolding, such as the timing of when it's provided, the communication modality, the format or representations used, or the strategies leveraged, such as comparing generated examples vs. guiding questions.

ReviewFlow incorporates three types of scaffolding to support cognitive activities: contextual scaffolding generates reflection questions to aid in paper reading; knowledge scaffolding focuses on citations to facilitate paper evaluation; and structural scaffolding synthesizes notes into structured outlines to assist in review writing. The current study provided insights into user perceptions of different scaffolding by observing their usage of all the features and asking them to rate each feature's usefulness and accuracy. More rigorous and well-controlled ablation studies could help us understand the underlying factors impacting the ReviewFlow experience. For instance, a study with three experimental conditions, each providing only one type of scaffolding—contextual scaffolding, knowledge scaffolding, and structural scaffolding—would allow for a more comprehensive evaluation of the quality of review writing, task efficiency, and the learning impacts for the user.

7.2 How can intelligent scaffolding support novices across the science ecosystem?

Our study of ReviewFlow indicates novice reviewers can benefit from intelligent scaffolding: it helps them evaluate the submission and write more comprehensive reviews. Participants mentioned how they gradually learned about the research community's practices and became more confident in their review writing. Beyond just writing reviews, the strategies for intelligent scaffolding built into ReviewFlow have the potential to provide value to other aspects of the peer review ecosystem [77]. For example, similar intelligent scaffolding can be used to support novice reviewers to revise their papers and write a persuasive rebuttal. The process of meta-reviewing, or summarizing independent reviews, could also benefit from scaffolding since there are new people jumping into that role each cycle [82].

Future deployments of this type of intelligent scaffolding would require careful consideration and round-table discussions within the research community. Previous research revealed that writing with an opinionated language model can affect participants' attitudes towards social topics in writing [42]. In our study, even though our scaffolding was carefully designed to allow the user to drive the process and to avoid biasing the decision, we still heard participants who were concerned about other people using such a system. Interestingly, very few participants were concerned that their own decision or writing was being biased by the AI, instead, they worried about how others would appropriate the intelligent support and how it might erode a community's trust in the peer review system. To deploy the system in the future and build trust with people in the community, we need to make sure that the system is robust and trustworthy [84]. A real-world deployment would need to have a large-scale consent process and commitment from key stakeholders.

Beyond its application in peer reviewing, intelligent scaffolding could support a range of of complex knowledge work in science, such as paper reading, literature reviewing, sense-making, and paper writing [56, 57, 66, 97]. Notable projects like the semantic reader project have developed interactive and dynamic reading interfaces to aid paper reading and citation discovery [6, 23, 44, 59]. These studies suggest broader possibilities for incorporating intelligent, process-driven scaffolding into science work.

7.3 Ethical considerations of AI scaffolding for academic review

Generative AI and LLMs introduce numerous ethical considerations in the design of human-AI collaboration systems. In the context of academic publishing and peer review, a primary concern involves the potential violation of academic integrity and harm to authorship when directly incorporating automatically generated content into writing artifacts. We try to mitigate this in ReviewFlow by constraining the use of LLMs to primarily sensemaking activities, like guided note-taking. In the final step, ReviewFlow synthesizes an outline, not paragraphs, even though LLMs are quite capable of doing so. To limit data sharing, ReviewFlow only sends the LLM the reviewer's notes and highlighted segments, not the full paper. While individual reviewers may still choose to (mis)use LLMs in these ways, ReviewFlow subtly prioritizes learning and thinking over getting the job done fast.

Another concern revolves around the tendency of LLMs to create inaccurate information or to mislead people[38, 42]. In line with our scaffolding strategy, ReviewFlow includes a checklist pop-up that allows users to engage in self-reflection, with reminders to proofread and fact-check their review. To enhance transparency in the writing process, future systems might explore the development of intelligent highlighting. For example, highlights could directly color-code the portions that were user-generated and those that were automatically generated, providing a clear visual indication and promoting a more transparent and accountable writing process.

Given the uncertainties, future LLM-powered tool designers should consider strategies for inducing users into a "reflective skepticism" around all data, especially those produced by machine models. As exemplified in prior work [51], this involves fostering a mindset of critical evaluation and thoughtful questioning to counteract the potential limitations, hallucinations, biases, or misinformation that may arise from these language models.

8 LIMITATIONS

Our study has several limitations. First, ReviewFlow combined multiple scaffolding strategies into one tool, leaving future work to understand to what extent each strategy impacted the outcomes. Ablation studies could provide more insight into the effectiveness of each scaffold. For instance, a study with three experimental conditions, each providing only one type of scaffolding would allow for a more comprehensive evaluation. Second, we simulated a mock peer review scenario where users had about one hour to read and write a review for a short paper. In practice, as we learned in our preliminary interviews, writing peer reviews takes hours. A longer study could give insights into the enduring value of scaffolding on a longer paper. Third, we selected two papers in HCI to use in our experiment, but different research domains and communities have different guidelines and standards for reviewing. Deploying the system across research communities may bring new insights into the generalization of scaffolding strategies. Lastly, the underlying machine models and LLMs will keep improving, which can impact the performance of ReviewFlow, for better or worse. Continuous updates and adaptations to the latest AI models would be required to maintain the tool's effectiveness and relevance.

9 CONCLUSION

This research explores techniques for integrating LLMs into intelligent scaffolding for academic peer reviews. Our formative studies found that expert reviewers adopted a workflow of annotating, note-taking, and synthesizing notes before writing, while novices lacked perspective on the prior work in the domain and reviewing standards. Modeling the expert workflow, we developed ReviewFlow –LLM-supported workflow that scaffolds novices using contextual reflection cues, in-situ knowledge support, and notes-to-outline synthesis. In a within-subject experiment with 16 novice reviewers, we found that ReviewFlow led to more comprehensive peer review and higher self-efficacy on the task. We further discuss the implication of intelligent scaffolding in knowledge work.

REFERENCES

[1] 2008. GROBID. https://github.com/kermitt2/grobid

[2] 2015. Investigating the Quality of Reviews, Reviewers, and their Expertise for CHI2023. https://chi2023.acm.org/2023/01/05/investigating-the-quality-of-reviews-reviewers-and-their-expertise-for-chi2023/

[3] 2015. Semantic Scholar. https://www.semanticscholar.org/

[4] Ines Arous, Jie Yang, Mourad Khayati, and Philippe Cudré-Mauroux. 2021. Peer grading the peer reviews: a dual-role approach for lightening the scholarly paper review process. In *Proceedings of the Web Conference 2021*. 1916–1927.

[5] Robert K Atkinson, Sharon J Derry, Alexander Renkl, and Donald Wortham. 2000. Learning from examples: Instructional principles from the worked examples research. *Review of educational research* 70, 2 (2000), 181–214.

[6] Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A Hearst, Andrew Head, and Kyle Lo. 2022. Paper plain: Making medical research papers approachable to healthcare consumers with natural language processing. *ACM Transactions on Computer-Human Interaction* (2022).

[7] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023* (2023).

[8] Alberto Bartoli, Andrea De Lorenzo, Eric Medvet, and Fabiano Tarlao. 2016. Your paper has been accepted, rejected, or whatever: Automatic generation of scientific paper reviews. In *International conference on availability, reliability, and security*. Springer, 19–28.

[9] Advait Bhat, Saaket Agashe, Parth Oberoi, Niharika Mohile, Ravi Jangir, and Anirudha Joshi. 2023. Interacting with Next-Phrase Suggestions: How Suggestion Systems Aid and Influence the Cognitive Processes of Writing. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 436–452.

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

[11] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.

[12] Dung C Bui and Mark A McDaniel. 2015. Enhancing learning during lecture note-taking using outlines and illustrative diagrams. *Journal of Applied Research in Memory and Cognition* 4, 2 (2015), 129–135.

[13] 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.

[14] Davida H Charney and Richard A Carlson. 1995. Learning to write in a genre: What student writers take from model texts. *Research in the Teaching of English* (1995), 88–125.

[15] Xiang'Anthony' Chen, Chien-Sheng Wu, Tong Niu, Wenhao Liu, and Caiming Xiong. 2022. Marvista: A Human-AI Collaborative Reading Tool. *arXiv preprint arXiv:2207.08401* (2022).

[16] Liying Cheng, Lidong Bing, Qian Yu, Wei Lu, and Luo Si. 2020. APE: argument pair extraction from peer review and rebuttal via multi-task learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 7000–7011.

[17] 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 *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–19.

[18] Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A Smith. 2018. Creative writing with a machine in the loop: Case studies on slogans and stories. In *23rd International Conference on Intelligent User Interfaces*. 329–340.

[19] Allan Collins. 2006. Cognitive apprenticeship: The cambridge handbook of the learning sciences, R. Keith Sawyer.

[20] Sara Doan. 2021. Teaching workplace genre ecologies and pedagogical goals through résumés and cover letters. *Business and Professional Communication Quarterly* 84, 4 (2021), 294–317.

[21] Peter Facione. 1990. Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction (The Delphi Report). (1990).

[22] Linda Flower and John R Hayes. 1981. A cognitive process theory of writing. *College composition and communication* 32, 4 (1981), 365–387.

[23] Raymond Fok, Hita Kambhamettu, Luca Soldaini, Jonathan Bragg, Kyle Lo, Marti Hearst, Andrew Head, and Daniel S Weld. 2023. Scim: Intelligent Skimming Support for Scientific Papers. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 476–490.

[24] Raymond Fok and Daniel S Weld. 2023. What Can't Large Language Models Do? The Future of AI-Assisted Academic Writing. In *In2Writing Workshop at CHI*.

[25] Yang Gao, Steffen Eger, Ilia Kuznetsov, Iryna Gurevych, and Yusuke Miyao. 2019. Does my rebuttal matter? insights from a major nlp conference. *arXiv preprint arXiv:1903.11367* (2019).

[26] Bill Gaver, Tony Dunne, and Elena Pacenti. 1999. Design: cultural probes. *interactions* 6, 1 (1999), 21–29.

[27] Katy Gero, Alex Calderwood, Charlotte Li, and Lydia Chilton. 2022. A design space for writing support tools using a cognitive process model of writing. In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*. 11–24.

[28] Katy Ilonka Gero and Lydia B Chilton. 2019. How a Stylistic, Machine-Generated Thesaurus Impacts a Writer's Process. In *Proceedings of the 2019 on Creativity*

and Cognition. 597–603.

[29] 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.

[30] Katy Ilonka Gero, Vivian Liu, and Lydia B. Chilton. 2021. Sparks: Inspiration for Science Writing using Language Models. arXiv. http://arxiv.org/abs/2110.07640 arXiv:2110.07640 [cs].

[31] Navita Goyal, Eleftheria Briakou, Amanda Liu, Connor Baumler, Claire Bonial, Jeffrey Micher, Clare R Voss, Marine Carpuat, and Hal Daumé III. 2023. What Else Do I Need to Know? The Effect of Background Information on Users' Reliance on AI Systems. *arXiv preprint arXiv:2305.14331* (2023).

[32] Michael Hannafin, Susan Land, and Kevin Oliver. 1999. Open learning environments: Foundations, methods, and models. *Instructional-design theories and models: A new paradigm of instructional theory* 2 (1999), 115–140.

[33] Tony Harland. 2003. Vygotsky's zone of proximal development and problem-based learning: Linking a theoretical concept with practice through action research. *Teaching in higher education* 8, 2 (2003), 263–272.

[34] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology.* Vol. 52. Elsevier, 139–183.

[35] John R Hayes. 2012. Modeling and remodeling writing. *Written communication* 29, 3 (2012), 369–388.

[36] Derek Holton and David Clarke. 2006. Scaffolding and metacognition. *International journal of mathematical education in science and technology* 37, 2 (2006), 127–143.

[37] Xinyu Hua, Mitko Nikolov, Nikhil Badugu, and Lu Wang. 2019. Argument mining for understanding peer reviews. *arXiv preprint arXiv:1903.10104* (2019).

[38] Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. 2019. Reducing sentiment bias in language models via counterfactual evaluation. *arXiv preprint arXiv:1911.03064* (2019).

[39] Julie Hui and Michelle L Sprouse. 2023. Lettersmith: Scaffolding Written Professional Communication Among College Students. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems.* 1–17.

[40] Julie S Hui, Darren Gergle, and Elizabeth M Gerber. 2018. Introassist: A tool to support writing introductory help requests. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.* 1–13.

[41] Daphne Ippolito, Ann Yuan, Andy Coenen, and Sehmon Burnam. 2022. Creative Writing with an AI-Powered Writing Assistant: Perspectives from Professional Writers. *arXiv preprint arXiv:2211.05030* (2022).

[42] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems.* 1–15.

[43] Tom Jefferson, Philip Alderson, Elizabeth Wager, and Frank Davidoff. 2002. Effects of editorial peer review: a systematic review. *Jama* 287, 21 (2002), 2784–2786.

[44] Hyeonsu Kang, Joseph Chee Chang, Yongsung Kim, and Aniket Kittur. 2022. Threddy: An Interactive System for Personalized Thread-based Exploration and Organization of Scientific Literature. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology.* 1–15.

[45] Holtzblatt Karen and Jones Sandra. 2017. Contextual inquiry: A participatory technique for system design. In *Participatory design.* CRC Press, 177–210.

[46] Anna Kawakami, Venkatesh Sivaraman, Logan Stapleton, Hao-Fei Cheng, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. 2022. "Why Do I Care What's Similar?" Probing Challenges in AI-Assisted Child Welfare Decision-Making through Worker-AI Interface Design Concepts. In *Designing Interactive Systems Conference.* 454–470.

[47] Joy Kim, Mira Dontcheva, Wilmot Li, Michael S Bernstein, and Daniela Steinsapir. 2015. Motif: Supporting novice creativity through expert patterns. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems.* 1211–1220.

[48] Markus Krause, Tom Garncarz, JiaoJiao Song, Elizabeth M Gerber, Brian P Bailey, and Steven P Dow. 2017. Critique style guide: Improving crowdsourced design feedback with a natural language model. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems.* 4627–4639.

[49] Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology.* Sage publications.

[50] 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.

[51] Emily R Lai. 2011. Critical thinking: A literature review. *Pearson's Research Reports* 6, 1 (2011), 40–41.

[52] Himabindu Lakkaraju and Osbert Bastani. 2020. " How do I fool you?" Manipulating User Trust via Misleading Black Box Explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.* 79–85.

[53] John Langford and Mark Guzdial. 2015. The arbitrariness of reviews, and advice for school administrators. *Commun. ACM* 58, 4 (2015), 12–13.

[54] Mina Lee, Percy Liang, and Qian Yang. 2022. CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. In *CHI Conference on Human Factors in Computing Systems.* ACM, New Orleans LA USA, 1–19. https://doi.org/10.1145/3491102.3502030

[55] Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Smith, Yian Yin, Daniel McFarland, and James Zou. 2023. Can large language models provide useful feedback on research papers? A large-scale empirical analysis. arXiv:2310.01783 [cs.LG]

[56] Susan Lin, Jeremy Warner, JD Zamfirescu-Pereira, Matthew G Lee, Sauhard Jain, Michael Xuelin Huang, Piyawat Lertvittayakumjorn, Shanqing Cai, Shumin Zhai, Björn Hartmann, et al. 2024. Rambler: Supporting Writing With Speech via LLM-Assisted Gist Manipulation. *arXiv preprint arXiv:2401.10838* (2024).

[57] Michael Xieyang Liu, Tongshuang Wu, Tianying Chen, Franklin Mingzhe Li, Aniket Kittur, and Brad A Myers. 2023. Selenite: Scaffolding decision making with comprehensive overviews elicited from large language models. *arXiv preprint arXiv:2310.02161* (2023).

[58] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *Comput. Surveys* 55, 9 (2023), 1–35.

[59] 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).

[60] Alison McCook. 2006. Is peer review broken? Submissions are up, reviewers are overtaxed, and authors are lodging complaint after complaint about the process at top-tier journals. What's wrong with peer review? *The scientist* 20, 2 (2006), 26–35.

[61] Piotr Mirowski, Kory W. Mathewson, Jaylen Pittman, and Richard Evans. 2022. Co-Writing Screenplays and Theatre Scripts with Language Models: An Evaluation by Industry Professionals. arXiv. http://arxiv.org/abs/2209.14958 arXiv:2209.14958 [cs].

[62] Jeffrey C Mogul. 2013. Towards more constructive reviewing of SIGCOMM papers. , 90–94 pages.

[63] Tim Moore. 2013. Critical thinking: Seven definitions in search of a concept. *Studies in Higher Education* 38, 4 (2013), 506–522.

[64] John C Nesbit and Olusola O Adesope. 2013. Concept maps for learning. *Learning through visual displays. Charlotte, NC: Information Age Publishing* (2013), 303–328.

[65] Wendy Peia Oakes, Kathleen Lynne Lane, Holly M Menzies, and Mark Matthew Buckman. 2018. Instructional feedback: An effective, efficient, low-intensity strategy to support student success. *Beyond Behavior* 27, 3 (2018), 168–174.

[66] 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.

[67] Ferdinando Patat, Wolfgang Kerzendorf, Dominic Bordelon, Glen Van de Ven, and Tyler Pritchard. 2019. The distributed peer review experiment. *The Messenger* 177 (2019), 3–13.

[68] Zhenhui Peng, Yuzhi Liu, Hanqi Zhou, Zuyu Xu, and Xiaojuan Ma. 2022. CReBot: Exploring interactive question prompts for critical paper reading. *International Journal of Human-Computer Studies* 167 (2022), 102898.

[69] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems.* 1–52.

[70] Simon Price and Peter A Flach. 2017. Computational support for academic peer review: A perspective from artificial intelligence. *Commun. ACM* 60, 3 (2017), 70–79.

[71] Napol Rachatasumrit, Jonathan Bragg, Amy X Zhang, and Daniel S Weld. 2022. Citeread: Integrating localized citation contexts into scientific paper reading. In *27th International Conference on Intelligent User Interfaces.* 707–719.

[72] Sajjadur Rahman, Pao Siangliulue, and Adam Marcus. 2020. MixTAPE: Mixed-initiative Team Action Plan Creation Through Semi-structured Notes, Automatic Task Generation, and Task Classification. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2, 1–26. https://doi.org/10.1145/3415240

[73] Brian J Reiser. 2004. Scaffolding complex learning: The mechanisms of structuring and problematizing student work. *The Journal of the Learning sciences* 13, 3 (2004), 273–304.

[74] David N Sattler, Patrick E McKnight, Linda Naney, and Randy Mathis. 2015. Grant peer review: improving inter-rater reliability with training. *PloS one* 10, 6 (2015), e0130450.

[75] John W Saye and Thomas Brush. 2002. Scaffolding critical reasoning about history and social issues in multimedia-supported learning environments. *Educational Technology Research and Development* 50, 3 (2002), 77–96.

[76] Nihar B Shah. 2022. An overview of challenges, experiments, and computational solutions in peer review (extended version). *Commun. ACM* (2022).

[77] Nihar B Shah, Behzad Tabibian, Krikamol Muandet, Isabelle Guyon, and Ulrike Von Luxburg. 2018. Design and analysis of the NIPS 2016 review process. *Journal of machine learning research* (2018).

[78] Richard Smith. 2006. Peer review: a flawed process at the heart of science and journals. *Journal of the royal society of medicine* 99, 4 (2006), 178–182.

[79] Ivan Stelmakh, Nihar B Shah, and Aarti Singh. 2019. PeerReview4All: Fair and accurate reviewer assignment in peer review. In *Algorithmic Learning Theory*. PMLR, 828–856.

[80] Ivan Stelmakh, Nihar B Shah, Aarti Singh, and Hal Daumé III. 2021. A novice-reviewer experiment to address scarcity of qualified reviewers in large conferences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 4785–4793.

[81] Ivan Stelmakh, Nihar B Shah, Aarti Singh, and Hal Daumé III. 2021. Prior and prejudice: The novice reviewers' bias against resubmissions in conference peer review. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–17.

[82] Lu Sun, Stone Tao, Junjie Hu, and Steven P. Dow. 2024. MetaWriter: Exploring the Potential and Perils of AI Writing Support in Scientific Peer Review. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW (2024). https://doi.org/10.1145/3637371

[83] Lev Semenovich Vygotsky and Michael Cole. 1978. *Mind in society: Development of higher psychological processes*. Harvard university press.

[84] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. 2023. DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. *arXiv preprint arXiv:2306.11698* (2023).

[85] Qingyun Wang, Qi Zeng, Lifu Huang, Kevin Knight, Heng Ji, and Nazneen Fatema Rajani. 2020. ReviewRobot: Explainable Paper Review Generation based on Knowledge Synthesis. http://arxiv.org/abs/2010.06119 arXiv:2010.06119 [cs].

[86] Rob Wass, Tony Harland, and Alison Mercer. 2011. Scaffolding critical thinking in the zone of proximal development. *Higher Education Research & Development* 30, 3 (2011), 317–328.

[87] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.

[88] Robert F Woolson. 2007. Wilcoxon signed-rank test. *Wiley encyclopedia of clinical trials* (2007), 1–3.

[89] Wenting Xiong and Diane Litman. 2011. Automatically Predicting Peer-Review Helpfulness. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Portland, Oregon, USA, 502–507. https://aclanthology.org/P11-2088

[90] Kevin Yang, Dan Klein, Nanyun Peng, and Yuandong Tian. 2022. Doc: Improving long story coherence with detailed outline control. *arXiv preprint arXiv:2212.10077* (2022).

[91] Kevin Yang, Nanyun Peng, Yuandong Tian, and Dan Klein. 2022. Re3: Generating longer stories with recursive reprompting and revision. *arXiv preprint arXiv:2210.06774* (2022).

[92] Phuttharaksa Yantraprakorn, P Darasawang, and P Wiriyakarun. 2013. Enhancing self-efficacy through scaffolding. *Proceedings from FLLT* (2013).

[93] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: Story Writing With Large Language Models. In *27th International Conference on Intelligent User Interfaces*. 841–852.

[94] Alvin Yuan, Kurt Luther, Markus Krause, Sophie Isabel Vennix, Steven P Dow, and Bjorn Hartmann. 2016. Almost an expert: The effects of rubrics and expertise on perceived value of crowdsourced design critiques. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 1005–1017.

[95] Kangyu Yuan, Hehai Lin, Shilei Cao, Zhenhui Peng, Qingyu Guo, and Xiaojuan Ma. 2023. CriTrainer: An Adaptive Training Tool for Critical Paper Reading. (2023).

[96] Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2021. Can we automate scientific reviewing? *arXiv preprint arXiv:2102.00176* (2021).

[97] Zheng Zhang, Jie Gao, Ranjodh Singh Dhaliwal, and Toby Jia-Jun Li. 2023. VISAR: A Human-AI Argumentative Writing Assistant with Visual Programming and Rapid Draft Prototyping. *arXiv preprint arXiv:2304.07810* (2023).