ReviewFlow: Intelligent Scaffolding to Support a Workflow for Academic Peer Reviews

ANONYMOUS AUTHOR(S)

Peer review is essential to the scientific process. Research communities conduct peer reviews to assess contributions and to improve the overall quality of science work. As research communities grow, and review volunteers spread thin, less experienced researchers are often recruited as peer reviewers for the first time. We sought to explore how technology could be designed to help novices adopt a research 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 streamline the writing of reviews that align with community expectations. In a within-subjects experiment, 16 inexperienced reviewers wrote reviews using ReviewFlow and a baseline environment with minimal guidance. We found that reviewers provided more comprehensive and justified reviews using ReviewFlow than the baseline, as rated by experts. They called out more paper weaknesses using ReviewFlow, but still struggled to provide constructive suggestions on each issue. Participants found that ReviewFlow provided more streamlined support in the process of writing reviews and expressed current concerns about using AI in the scientific review process. We discussed the implications of using AI to scaffold complex peer review processes.

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

Additional Key Words and Phrases: intelligent scaffolding, academic peer review, LLM

ACM Reference Format:

Anonymous Author(s). 2018. ReviewFlow: Intelligent Scaffolding to Support a Workflow for Academic Peer Reviews. 1, 1 (November 2018), 25 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Peer review is a cornerstone of academic research, ensuring the quality, credibility, and reliability of scientific research [52]. The peer review process seeks to assess whether submissions contribute new knowledge to a research community and also generate feedback that helps authors improve the quality of their work [34, 58]. Many communities are seeing a rapid increase in submissions; while this could be seen as an indicator of scientific progress, it also has increased the pressure on peer review ecosystems [1, 41, 58, 61, 62]. To deal with increased demand, many research communities recruit a significant number of first-time reviewers for each review cycle.

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 draft to communicate their assessments and recommendations [58]. It is a task that requires a deep understanding of the subject matter, critical

Manuscript submitted to ACM

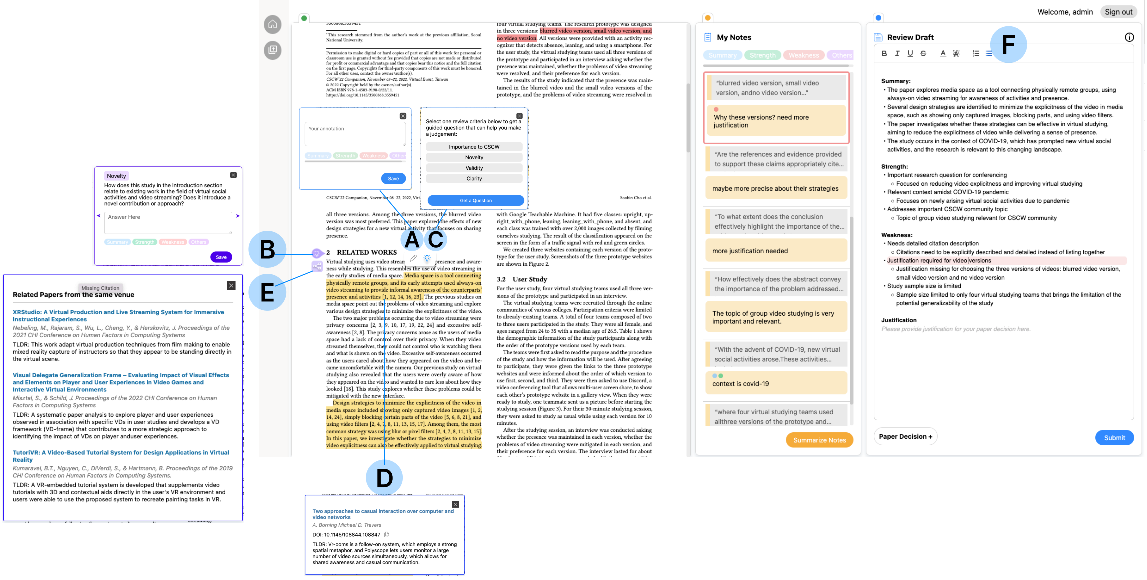
thinking, and the ability to provide constructive feedback. Given the growing number of novice reviewers, conferences need to ensure that newly added novice reviewers meet the standards and expectations of the research community. Towards supporting the growing number of novice reviewers, our research explores the research question: How can we intelligently scaffold the peer review process for novices while avoiding potential biases?

Scaffolding is commonly used in educational settings to enhance learning and aid in the mastery of tasks [16, 56]. Previous studies found that scaffolded examples and templates can even help learners perform similarly to experts in terms of feedback quality [75]. Existing research showed that AI scaffolding and machine intelligence offer tremendous potential to collaborate with humans by providing inspiration or generating diverse prompts [24, 25, 38]. Recently, LetterSmith's instructional approach to aid writing named "scaffolded annotation" helped professional writing students improve the quality of their early-stage drafting in a new genre [31]. Likewise, researchers created a chatbot–CReBot– that can provide guided questions and hints in real time to scaffold the critical reading process. Experimental results showed that CReBot can engage novices in the reading process and help them comprehend the paper content. Additionally, existing research showed that AI scaffolding and machine intelligence offer tremendous potential to collaborate with humans by providing inspiration or generating diverse prompts [24, 25, 38]. While interactive scaffolding has proven valuable for relatively well-scoped tasks, academic peer review involves critically reading, synthesizing knowledge, and making a well-justified judgment for acceptance or rejection. Our research explores how interactive scaffolds can guide a complex, multi-faceted workflow for academic peer reviewing without biasing decisions.

To explore how we might use AI scaffolding to support the peer review process, we first conducted a formative study to understand how experts approach this task and what novices see as key challenges. We conducted observational studies with 10 experts to ask them wrote a peer review for a short paper and think aloud about their workflows. 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 of critical reading, annotating, note-taking, and synthesizing a well-justified review that conforms to community guidelines. To understand novice reviewers' challenges, we interviewed 10 novices to dive into their obstacles in the process. They raised the challenges of lacking sufficient guidance on the procedure of writing a conference peer review and making judgments on the paper's quality.

We developed a prototype – "ReviewFlow", an AI-scaffolding system that supports inexperienced conference reviewers in the workflow of writing scientific peer reviews. ReviewFlow incorporates a range of features to facilitate the review process: (1) **Contextual cues** that provides section-specific cues guided by community criteria or contextual cues adapted to paper content. (2) **In-situ citation recommendations** support reviewers to assess the novelty compared with other work. (3) **Notes-to-outline synthesis** guides reviewers to organize notes and restructure reviews to align with community standards. It first summarizes notes together with paper content into high-level outline with only broad topics and reviewers have the option to expand it into a detailed outline that is structured based on community standards.

We conducted a within-subjects study (n=16) to evaluate ReviewFlow where all participants who are inexperienced reviewers wrote reviews for two HCI short papers in a counterbalanced manner: one using the ReviewFlow with full scaffolding features and one using the baseline interface that only has minimal guidance including presenting a review guideline and providing an example review template. We found that novice reviewers wrote significantly more structured and more justified reviews in the ReviewFlow system than in the Baseline system, evaluated by experts. Novice reviewers can write slightly more constructive reviews in the ReviewFlow system, but the difference is not



Your annotation

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

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

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...

Welcome, admin

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.

Fig. 1. ReviewFlow interface.(A) users can highlight, annotate and tag the content; (B) users are can conduct section-specific reflections guided by community criteria; (C) users can conduct contextual cues adapted to the highlight paper content; (D) users can click the citation to check in-situ summarization; (E) users can check the recommended citations that are potential missed by the paper; (F) users can click to summarize the notes into high-level outline then expand to a detailed outline that is structured based on community standard.

significant. Reviewers can call out more weaknesses in the paper using ReviewFlow, but they still struggle to provide constructive and actionable suggestions for the weaknesses.

Our paper offers several contributions: First, a formative study uncovered common practices adopted by experts in the review process and found that novices lack timely guidance on key considerations and expectations. Second, we developed ReviewFlow to model an expert workflow for peer reviewing while also leveraging LLMs to provide contextual reflection, 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 AI-driven scaffolding can help novices better structure and justify their review.

2 RELATED WORK

2.1 Practices and challenges related to academic peer reviewing

Peer review is an essential step of academic research. As an important step for ensuring the scientific quality of work, peer review has been adopted by most journals and conferences [52]. 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 [58, 59]. 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. Note that conferences may have some differences in their peer-review process.

The number of papers in research communities has increased exponentially in recent years [68]. 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 [41, 58, 61, 62]. To avoid overloading reviewers, conferences need to find new sources of reviewers as there are not enough experienced reviewers to review all papers [**?**]. These novice and junior reviewers constitute a large fraction of the reviewer pool in computer science conferences [63]. For example, using data on the structure of the reviewer pool of the ICML 2020 conference, researchers estimated that approximately 35% of the reviewers were novice reviewers who are junior individuals who self-nominated and satisfied the screening requirements of having one or two papers published in some top ML venues [63]. 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 [59, 62]. 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 [44, 50]. Faced with these doubts and challenges, helping novice junior reviewers to write a high-quality review becomes crucial.

Reviewing is a complex, time-consuming, and mentally demanding task [58, 68]. 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 reviewers [68]. To provide a high-quality review, reviewers need to go through a multi-step workflow: first, they need to read and fully understand the contribution of the paper; second, they need to use the domain knowledge to assess the merit of scientific contributions and make a decision on the submission; last, they need to provide an evaluation that covers multiple aspects, including the originality, clarity, validity, etc [13, 21, 30, 34, 36, 60, 68, 71, 77]. 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 [18, 45]. 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 [45].

Researchers are mainly trained to do research while there is little training for the peer review process. Hence, it becomes even more challenging for inexperienced junior reviewers to gain expertise quickly [58]. 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 [57]. 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, with the goal of helping novices to write better reviews. The results showed that with this guidance on the reviewing stages, novice reviewers can deliver more "above expectation" reviews [62]. 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, for instance, how to initiate a discussion among reviewers [62]. 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, our research starts 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 peer reviewers, authoring more comprehensive and decisive reviews, as well as review quality assessment [1, 4, 30, 59, 68, 77]. However, there is less empirical study that attempts 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 [16, 64]. Scaffolding is instructional support provided by experts to promote learning, especially when concepts and skills are being first introduced to novice students [16, 56]. These supports include advanced organizers, modeling, worked examples, concept maps, explanations, handouts and prompts [2, 7, 11, 16, 46, 49]. Previous research shows that effective scaffolding can help novices perform work nearly as well as experts [31, 32, 75]. 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 [27].

Researchers used prompts and guided questions to scaffold learners in the paper reading process [12, 51, 76]. 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 read. Similarly, 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, which ignored the contextual information provided by the paper.

Existing research also uses existing examples and templates to scaffold complex writing process [31, 32, 75]. In the writing scenario, scholars found that scaffolding can help students learn about form and organization from analyzing the examples and templates [16, 17]. In the context of writing introductory help requests, providing high-quality examples and expert informed templates can increase learning and writing quality [32]. Another writing support system used scaffolded annotation and examples as an instruction approach to professional writing students improve the quality of their early-stage drafting in a new genre [31].

In traditional instruction scenarios, experts take a huge amount of time to curate examples, create guidance, or author rubrics [56]. However, experts who created examples still faced the challenges of effectively adapting and contextualizing into the correct learning step. Thus, this brings AI and machine intelligence opportunities to provide efficient and context-specific scaffolding in the learning process [65, 66]. Previous research has developed multiple methods for using machine intelligence to scaffold complex tasks, including writing and researching [25, 48, 54, 74]. For example, researchers developed ArgueTutor, a conversational agent that tutors students with adaptive argumentation feedback in their learning journey using Natural Language Processing [65]. 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.

2.3 Leveraging AI to support writing

Recent works explored the potential of using AI to support people's workflow on writing tasks [14, 15, 23–25, 33, 33, 37, 43, 55]. For example, TaleBrush allows users to create a story with AI through sketching to aid the planning of writing. Then it enables writers to generate diverse storylines and interactively refine them [14]. Another system Wordcraft explores how to support users collaborating with generative language models to co-write a story [74]. Spark used a language model to generate prompts related to a scientific concept to facilitate scientific writing [25]. Research showed that writing is a complex, iterative process [19]. The Hayes model describes the cognitive processes an individual writer engages in during the process of writing [19, 22, 29]. In the writing process, there are three major components: the planning component, the translating component, and the reviewing component, as shown in Figure 5. To facilitate the cognitive process of writing, several systems are developed to support different stages [5, 78]. For example, VISAR is an AI-enabled writing assistant to 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 [69]. Specifically, 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. The evaluation demonstrated that this decomposition approach can improve the coherence of long story generation [72, 73]. However, this approach could be 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 where human reviewers should play the role of driving the writing process. The development of LLM brings opportunities along with concerns on what LLMs shouldn't do in the academic writing area [20]. 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 [39]. 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. Instead of using an automatic method, humans should be in the loop to drive and control the writing process [20]. 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 a series of formative studies. First, we interviewed ten novice reviewers to understand how they approached this task for the first time and their perceived challenges. Second, we observed ten experienced reviewers as they conducted a peer review on a couple of selected papers in order to understand common practices and workflows.

3.1 Formative Study Methods

3.1.1 Novice Interview Study. We conducted 30-minute semi-structured interviews with 10 novice reviewers who have only experience in writing academic peer reviews for once but less than 2 years (4 female and 6 male, average age of 25.5 years). Participants were recruited through email lists and social media posts. Participants came from diverse domains: Human-Computer Interaction, Artificial Intelligence, Cognitive Science, and Computer Systems.

In the interview, we focused on the current obstacles and challenges of conducting conference peer review. We further provided scenarios to the participants and asked them to rate how much they resonated with the novice reviewers based on their own experiences. These scenarios described common novices' situations when they review papers. For example, "Mary is a first-year graduate student in the research area of Human-Computer Interaction (HCI). She conducted research for two years and submitted two papers on the creativity support tool for HCI conferences. This is the first time that she was invited as a conference peer reviewer to write a peer review and make a decision for another paper on the topic of using LLM to support creativity. While she wrote the review, she struggled to make the review constructive to the author.". This approach can help to elicit participants' needs in a real-world situation [51]. Based on these answers, our research team conducted multiple rounds of designs to come up with ideas and design probes targeting these challenges.

In the end, we provided a series of design probes to the novices at the end of the study to elicit their needs. Drawing insights from previous literature around the intelligent support on reading and writing, the research team took multiple rounds of iteration and group discussions to develop the six design probes. Two researchers in the team visualized the design probes in Figma and provided the prototypes to the user. The design probes include: (1) scaffold annotation with community curated tags and filters, (2) contextual reflection questions (expert-authored generic questions versus AI-generated specific 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 with structures, (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.

3.1.2 Experts Observational Study. Cognitive psychology uses the think-aloud technique [26, 47] to model learners' thinking processes and improve instructions. To learn from experts' best practices that are gained through years of learning-by-doing, we further conducted an observational study with experts to observe their real practices and extract their common review workflow. We recruited 10 experts who have experience in writing academic peer reviews at least 5 years (4 female and 6 male, average age of 30.1 years). Participants were recruited through in-person invitations, email lists, and social media posting. Participants came from diverse domains: Human-Computer Interaction, Artificial Intelligence, programming language, learning science, security, and accessibility.

During the observational study, we first spent 20 minutes asking for experience and challenges of the current peer review practice. Then, the team asked the participant to write a peer review on a Google doc for a 2-4-page short paper in 60 minutes. The team provided a list of papers that were selected based on the participant's research interest. To find a suitable paper, the team first collected the participants' research interest descriptions online and used these keywords to search on the Semantic Scholar platform, and then we ranked the paper by relevance and recency. Then we filtered out the papers that were longer than 5 pages.

Each expert was asked to write a conference peer review in a document. Then, we provided the same design probes to the experts to elicit what scaffolding strategies could benefit the novices the most. Observational studies were conducted remotely by the lead author and lasted around 90 minutes. They were recorded and then transcribed using a machine transcription service. Then, three researchers from the team went through the transcripts and coded them for themes using an open coding approach [10]. Through multiple iterations along with periodic discussions with the rest of the research team, the coding led to the major themes below.

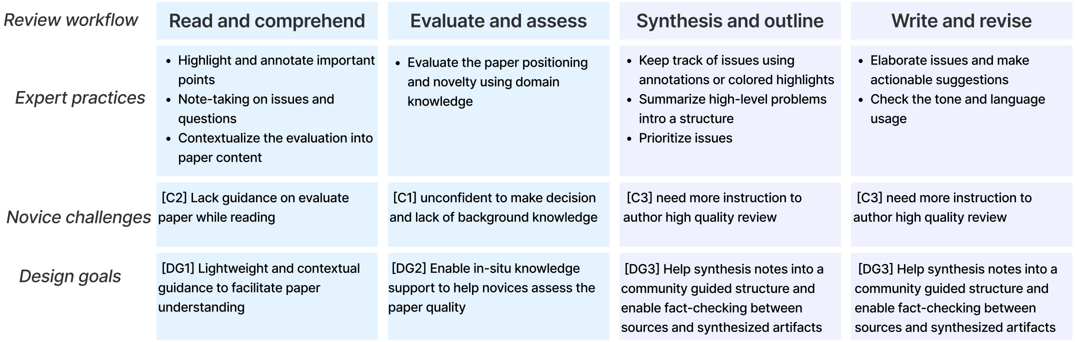
3.2 Formative Study Findings

3.2.1 Novice reviewers' highlights challenges of lacking guidance on evaluating and writing. Novice reviewers report that their review process took on average 6.4 hours. All participants reviewed once or more than once but have less than two years of review experience in their research field. All participants mentioned that reviewing papers is a cognitively demanding task. When asked about challenges, 6/10 mentioned that they *struggle to fully understand the background knowledge and existing work and feel unconfident about assessing the paper's novelty (C1).* "If I'm not super familiar with the topic, making sure the work is novel and all prior work is included can be hard for me because I don't know if they might have missed or not included related things or references"(N3). 4/10 mentioned that they *need more guidance on evaluating the paper critically from different aspects(C2).* Specifically, one participant mentioned that "I need some co-reviewers who have already read the paper and know the specifics to guide me through the evaluation process"(N5). Another participant said "I think I need experts to tell me what are some of the main things you should consider while evaluating the paper"(N1). In addition, 4/10 said that they *need more instructions on writing a high-quality peer review(C3).* Specifically, one participant is worried about "I don't want to be impolite or harsh when I point out issues in the paper(N9)" and "I am not sure whether I covered all the necessary points or whether authors will perceive the reviews as useful"(N7).

When asked about the ratings on the degree of resonate with the novice reviewer's scenarios (Not resonate at all 1 - 5 Strongly resonate with the situation), the top two situations that resonate with novice reviewers the most are that "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 don't know how to evaluate the paper"(3.8/5). In addition, participants also resonate with the situation that "the novice reviewer takes several hours to conduct extensive background research in order to understand the paper" (3.6/5) and "the novice reviewer finds it hard to provide evaluations on the paper given the general conference review guideline"(3.6/5).

3.2.2 Experienced reviewers adopt a workflow of sensemaking, 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 will 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 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 have to make a lot of highlighted a lot of annotations, and then I will probably just review it again. And then, just like section by section and say, Okay, so like, thinking about, like the biggest concern in terms of the overall direction, like a novelty of things"(E6). We observed that 6 out of 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



Review workflow	Read and comprehend	Evaluate and assess	Synthesis and outline	Write and revise
Expert practices	• Highlight and annotate important points • Note-taking 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 colored highlights • Summarize high-level problems intro a structure • Prioritize issues	• Elaborate issues and make actionable suggestions • Check the tone and language usage
Novice challenges	[C2] Lack guidance on evaluate paper while reading	[C1] unconfident to make decision and lack of background knowledge	[C3] need more instruction to author high quality review	[C3] need more instruction to author high quality review
Design goals	[DG1] Lightweight and contextual guidance to facilitate paper understanding	[DG2] Enable in-situ knowledge support to help novices assess the paper quality	[DG3] Help synthesis notes into a community guided structure and enable fact-checking between sources and synthesized artifacts	[DG3] Help synthesis notes into a community guided structure and enable fact-checking between sources and synthesized artifacts

Fig. 2. Expert workflow in conference peer reviewing, along with experts practices, novices challenges, and design goals for each stage.

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.

3.2.3 Experienced reviewers stressed the importance of specific and contextual guidance for the design of scaffolding. Among these features, 5 experts reflected that the guided questions can be helpful, especially for novice reviewers. One participant preferred the AI-generated question 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 review. So now they do. They don't really know how to review, and I think some guided question is important. But for people who have reviewed for so many years, I just see if you keep in review all the conference, those guidance are pretty much 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 questions 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 where they don't trust the AI 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 LLM should not generate the review on the fly while human experts should drive the process since the limitation of LLM can bias human experts and bring 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 don't distract the current review flow. While involving AI in the review process, AI should not go too far to basis or lead the thinking process. Human reviewers should still preserve agency in reading, writing, and decision-making. Faced with the limitations of current LLM, 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, to facilitate each stage of the review workflow, we identify the following 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 practiced by experienced reviewers
- DG4: Guide the reviewing process such that the work aligns with community standards
- DG5: Avoid biasing the decision to either accept or reject the paper, but encourage justifications

.

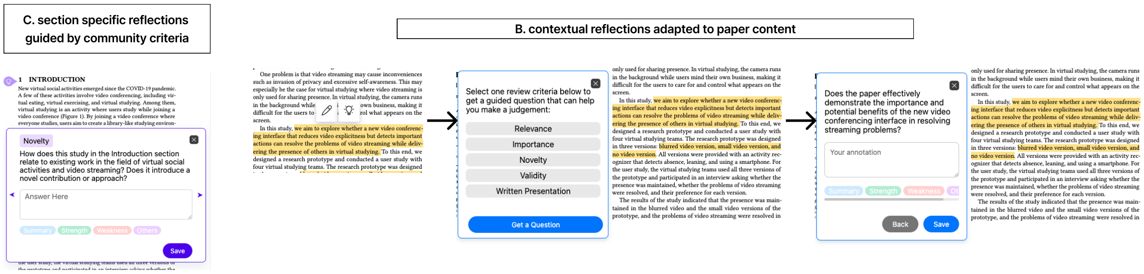
4 REVIEWFLOW

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 [12, 51, 76]. Understanding the paper and critically reflecting on the content is essential for decision-making and review writing in the later stage. Previous research showed that providing questions can increase user engagement in actively searching for information. Previous systems mostly use template-based guided questions to engage readers [51]. To *DG1: Offer lightweight and adaptive scaffolds that facilitate reflection throughout the paper*, ReviewFlow provides two types of contextual cues for novices, as shown in Figure 3. One type is **section-specific reflections cues guided by community criteria** which are adapted based on each paper section's content together with the review criteria. For section-specific reflections (C.), it took the entire section, the paper abstract, and the community criteria into account to generate cues. Readers 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 **contextual reflections cues adapted to paper content** where readers can identify the content to highlight and focus on. For contextual reflection adapted to paper content (B.), readers have the freedom to highlight the content that they would like to reflect on and select the review criteria. This provides more adaptations for reviewers to judge on different types of content.


C. section specific reflections guided by community criteria
B. contextual reflections adapted to paper content

1 INTRODUCTION

New virtual social activities emerged since the COVID-19 pandemic. A few of these activities involve video conferencing, including virtual eating, virtual exercising, and virtual studying. Among them, virtual studying is an activity where users study while joining a video conference (Figure 1). By joining a video conference where everyone studies, users aim to create a library-like studying environ-

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?

One problem is that video streaming may cause inconveniences such as invasion of privacy and excessive self-awareness. This may especially be the case for virtual studying where video streaming is only used for sharing presence. In virtual studying, the camera runs in the background while own business, making it difficult for the users to rol what appears on the screen.

In this study, we aim to explore whether a new video conferencing interface that reduces video explicitness but detects important actions can resolve the problems of video streaming while delivering the presence of others in virtual studying. To this end, we designed a research prototype and conducted a user study with four virtual studying teams. The research prototype was designed

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

only used for sharing presence. In virtual studying, the camera runs in the background while users mind their own business, making it difficult for the users to care for and control what appears on the screen.

In this study, we aim to explore whether a new video conferencing interface that reduces video explicitness but detects important actions can resolve the problems of video streaming while delivering the presence of others in virtual studying. To this end, we designed a research prototype and conducted a user study with four virtual studying teams. The research prototype was designed in three versions: blurred video version, small video version, and no video version. All versions were provided with an activity recognizer that detects absence, leaning, and using a smartphone. For the user study, the virtual studying teams used all three versions of the prototype and participated in an interview asking whether the presence was maintained, whether the problems of video streaming were resolved, and their preference for each version.

The results of the study indicated that the presence was maintained in the blurred video and the small video versions of the prototype, and the problems of video streaming were resolved in

Does the paper effectively demonstrate the importance and potential benefits of the new video conferencing interface in resolving streaming problems?

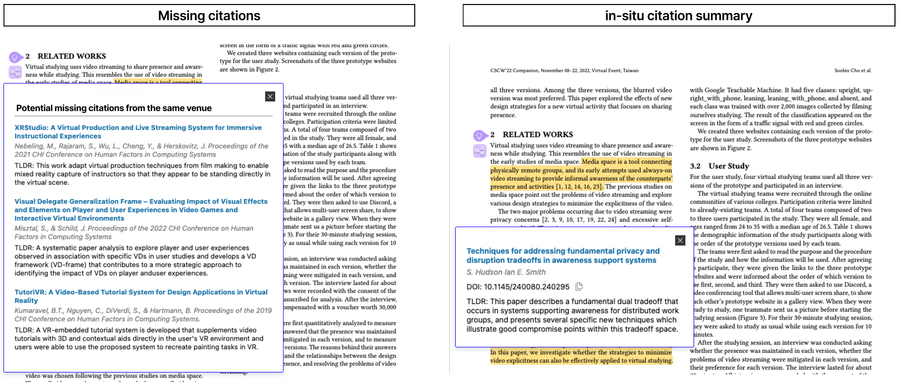
only used for sharing presence. In virtual studying, the camera runs in the background while users mind their own business, making it difficult for the users to care for and control what appears on the screen.

In this study, we aim to explore whether a new video conferencing interface that reduces video explicitness but detects important actions can resolve the problems of video streaming while delivering the presence of others in virtual studying. To this end, we designed a research prototype and conducted a user study with four virtual studying teams. The research prototype was designed in three versions: blurred video version, small video version, and no video version. All versions were provided with an activity recognizer that detects absence, leaning, and using a smartphone. For the user study, the virtual studying teams used all three versions of the prototype and participated in an interview asking whether the presence was maintained, whether the problems of video streaming were resolved, and their preference for each version.

The results of the study indicated that the presence was maintained in the blurred video and the small video versions of the prototype, and the problems of video streaming were resolved in

Fig. 3. Contextual reflections cues. ReviewFlow provided (B.)contextual reflection cues adapted to paper content and (C.) section-specific reflection cues guided by community criteria.

4.1.2 ***DG2: In-situ citation recommendation as knowledge scaffolding.*** Previous research in scientific literature reviewing highlights the importance of in-situ knowledge support [3, 8, 9, 35, 53]. For example, PaperPlain provided in-situ plain language section summaries to facilitate paper reading. To enable in-situ knowledge support for assessing novelty compared to prior work, ReviewFlow in-situ presents TLDR summary when the user hovers over each reference to quickly provide background knowledge around each paper, as shown in Figure 4.To further raise the awareness of the existing reference in the reading workflow, ReviewFlow provided a popup window with potential missing citations from the same venues, as shown in Figure 4



Potential missing citations 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.

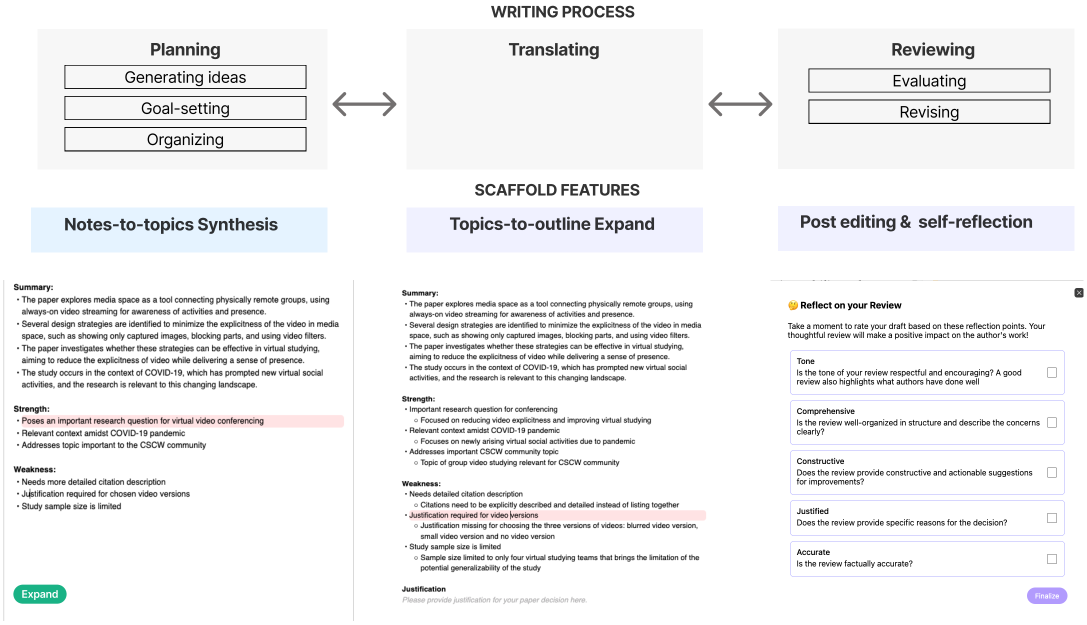
TutoriVR: A Video-Based Tutorial System for Design Applications in Virtual Reality
Kumaravel, B.T., Nguyen, C., DiVerdi, S., & Hartmann, B. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.
TLDR: A VR-embedded tutorial system is developed that supplements video tutorials with 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.

Fig. 4. Knowledge support

4.1.3 ***DG3 / DG4: Notes-to-outline synthesis.*** In the review process, we observe expert reviewers synthesize notes to make a plan on how to draft a review. 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 [69]. Specifically, 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. The evaluation demonstrated that this decomposition approach can improve the coherence of long story generation [72, 73]. However, this approach could be highly human controllable where humans can control the story generation by modifying the outlines. Recent studies explored the potential of automatically generating feedback using LLM for conference papers [39] Hence, by modeling the expert workflow of reading, note-taking, and synthesizing practiced by experienced reviewers, we designed the **notes-to-outline synthesis** feature that facilitates notes synthesis and review planing [19]. Figure 5 showed that ReviewFlow first synthesized notes to high-level outline to help reviewers organize their notes into structured topics and provide ideas for reviewers to make decisions. To further guide the reviewing process such that the work aligns with community standards, notes are summarized into broad topics and structured according to community standards, which includes aspects of summary, strengths, and weaknesses. Given these high-level outlines, reviewers can add their own points or delete any points. If reviewers would like to expand the high-level outline into a detailed outline with complete bullet points, reviewers can click the expand button that synthesized notes into a structured outline, as shown in Figure 5

4.1.4 ***DG5: Fact-checking between outline and source notes.*** 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



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? □

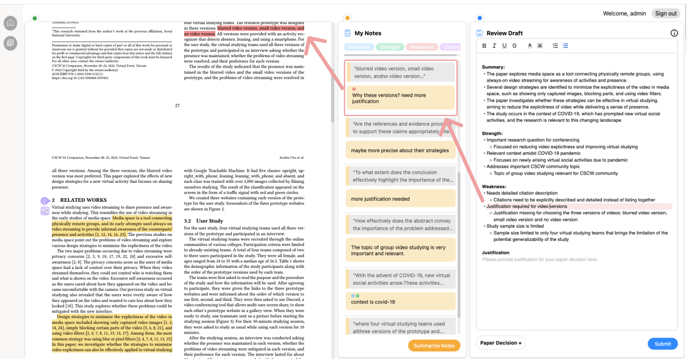
Constructive
Does the review provide constructive and actionable suggestions for improvements? □

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

Accurate
Is the review factually accurate? □

Fig. 5. Scaffolding features support the review drafting process. The top part shows the Flower and Hayes cognitive process of writing [19]. The bottom part shows the ReviewFlow features that support each stage of writing.

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 hover on it, the source notes which was used to synthesize will be highlighted in the middle, together with the previous pdf content, as shown in the Figure 6.



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

Why these versions? need more justification

"Are the references and evidence provi... to support these claims appropriately ite...

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.

Fig. 6. Visually mapping the bullet back to the notes and paper content to enable fact-checking

4.2 System implementation

ReviewFlow's front-end is a React web application 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 Flask server which handles the GPT-4 API endpoints, MongoDB endpoints for storing data, and GROBID for PDF data extraction. This data includes all of the parsed text of the PDF, its PDF coordinates, and citations linked to their references.

4.2.1 Section-specific reflection cues guided by community criteria. 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 reflection questions 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. Upon initializing the PDF, the front-end client aligns GROBID's coordinates with the scale-adjusted client PDF coordinates to create corresponding markers beside each section header. When the user clicks on this marker, a popup appears, enabling the user to cycle between the different contextual reflection questions and create tagged notes when answering the questions.

4.2.2 Contextual reflection cues adapted to paper content. When the user highlights text on the PDF and clicks on the insitu button with the light bulb icon, a pop-up appears with the 4 critical review aspects: importance, novelty, validity, and clarity. When the user selects one of the aspects and clicks "Get a Question", a POST request with the selected aspect, highlighted sentence, paragraph of the highlighted sentence, and paper abstract is sent to the back-end server. This data is then used in the prompt for a GPT-4 API call to generate a contextual reflection question that is streamed to the front-end client. Similar to the contextual reflection question, the user will then have the ability to respond to the question and create a tagged note.

4.2.3 In-situ citation recommendation. When the PDF is initialized, ReviewFlow not only utilizes GROBID's XML/TEI data to create markers for the section-specific reflection questions but also uses the data to create a citation layer that overlays all in-line citations. This citation layer is constructed by aligning GROBID's scaled coordinates to the client PDF's coordinates to create an HTML division element that overlays the in-line citation. This new element is bound with an event listener such that when the user clicks on the element, a popup appears with the corresponding citation data. The citation data includes GROBID's citation data such as the title, publication date, and DOI. Additionally, a Semantic Scholar API call is made for each citation to generate a TLDR description of each paper and to populate any missing fields that are not already provided by GROBID. To recommend corresponding papers, the system calls the Semantic Scholar Recommendation API[1] and uses the keywords of the paper together with the venue to retrieve the most similar paper. After checking the retrieved papers DUIs are not included in the current paper, the top three papers are added as a citation pop-up.

4.2.4 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

[1] https://api.semanticscholar.org/api-docs/recommendations

front end initiates a POST request to the back-end server with this data. The back-end server will use this data to prompt the GPT-4 model. We used the few shot learning paradigm with the corresponding prompts [40]: "Please create three important topics on the paper's [strengths] and [weaknesses] with less than ten words that combine and summarize the user notes." This 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 any bullet point from the generated outline, the sourced note and paper content will be highlighted so the user can see where the generated bullet point originated from. After generating this high-level outline, the user can freely make changes and further expand on the outline by generating low-level details for each of the short topics under the Strength and Weakness sections. 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 where a GPT-4 API call is requested. 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. The topic bullet points will additionally update their source and highlight the corresponding note and paper content when clicked on. [**?**]

5 METHOD

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

- RQ1: How does ReviewFlow affect participants' final written review quality?
- RQ2: How does ReviewFlow affect participants' peer review workflow, in terms of time and engagement?
- RQ2: How do participants perceive the AI scaffolding in the ReviewFlow system? What challenges do they have?

5.1 Within-subjects Experiment

We conducted a within-subjects experiment with 16 participants where each participant will experience both conditions in two sessions separately. We counterbalanced the order of two conditions and papers using Latin Square to minimize the potential order effect. To reduce knowledge transfer and minimize fatigue, we scheduled the two study sessions of each participant at least 24 hours apart.

5.1.1 Within-subjects Experiment Participants. Before the study, we sent out a pre-study survey to participants to collect information about their expertise, research topics, review experience, ML knowledge level, 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 two years of conference peer review experience. 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. All 16 participants are graduate students.

5.1.2 Within-subjects Experiment Procedure. Before the user study, participants filled out a pre-survey that captured their previous experience in reviewing and reviewing papers on ML conferences and their expertise in ML to join the study through an online teleconferencing tool. We asked participants whether they had read the paper before to make sure they all were seeing the work for the first time. In the beginning, participants read and virtually signed our IRB-approved consent form. Then they 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 some participants encountered ReviewFlow in their first session, and some others experienced it in their second session. For each session, we followed the review process used at the ICLR conference, where we provided the paper draft and all three individual reviews with their ratings and confidence scores.

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 4 participants who had reviewing experience prior to the study, we found that participants can finish writing 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 two sessions, the research team asked the participants to fill out a post-survey to evaluate the system, such as their satisfaction with each feature and their trust in the system. The research team then 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. Each session takes around 75 minutes.

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 [2]. 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. Then we filtered out papers that have technical terms and jargon. The two papers we selected need to have similar lengths, similar difficulties and from the same conference. Combining all the criteria above, we selected two comparable papers from CSCW Companion that were used for the within-subject experiment.

5.1.4 Within-subjects Experiment 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 artifacts for each paper (N=32) that recorded the final review, the post-survey (N=32 and the interview transcripts (N=16). The research team analyzed these combined sources of data to reveal insights towards our research questions.

Quality of final 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/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 [77]:

- 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 also 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 first provided two examples and provided instructions for them to rate and discuss until they reached a consensus on ratings based on the instructions. Then they rate each dimension of the review on the paper they wrote independently (N=16). The research team used the average scores by two experts for each paper dimension.

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

User interaction data. To measure participants' interaction with the tool, we instrumented the interface in order 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. We collected how much time they spent on authoring the draft from the time stamp they started editing the review in the text box as well as the entire session time, including time spent on reading and writing.

The ReviewFlow interface also collected interaction data to indicate how much each participant interacted with each machine intelligence feature including how many times a participant turned on and off the tags to highlight aspects, how many times a participant hovered over the extracted sentences, and how much content a participant cop from the generated review draft to write their own review.

User preferences and task perceptions. To evaluate users' preferences for the ReviewFlow experience compared to a 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. Then we specially asked participants to evaluate each of the machine-generated features. The survey collected their 7-point Likert scale ratings on perceived usefulness, perceived accuracy for each feature. We collected the level of cognitive demand using NASA TLX on a scale of 1-7 [28]. We further collected the feeling of control, collaboration on a scale of 1-7 [70].

User reactions. 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", and "Which function did you use the most? Which one did you like the most?", and "What concerns did you have when using ReviewFlow?".

Control variables - Users' knowledge of each paper's topic. After the post-survey, participants described their knowledge of each paper's topic from not familiar as 1 to very familiar as 7. All participants reported that they had never read or remembered the papers before.

5.2 Analysis

Quantitative data anlaysis. To measure the effect of the ReviewFlow system on each dimension of the review quality (eg. summative, coverage, 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, and the length of each review 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 of 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 as co-variants.

Qualitative data anlaysis. All semi-structured interviews with participants are recorded and transcribed. Two researchers conducted iterative open coding on the transcripts using Dovetail[3] following the grounded theory approach [10]. They open-coded the data by identifying topics mentioned by the participants. Initial codes were combined

[3] https://dovetailapp.com/

into preliminary themes, which were discussed among the research team. Finally, after iteratively discussing the code themes, researchers derived the final themes around: participants' reactions to each feature, their overall perceptions of the ReviewFlow system, and their concerns about using the system.

6 RESULTS

We report our results from the within-subjects experiment and the interview study. In the within-subject experiments, across both conditions, participants spent an average of 38 minutes writing a review with an average length of 243 words written. 59.3% of participants chose to accept decisions, which is consistent with the original decisions for the two papers. Our findings suggested that ReviewFlow provides more guidance to novice reviewers and can make the review process more useful and engaging.

6.1 RQ1: ReviewFlow helps improve the review draft quality

6.1.1 ReviewFlow helps participants write more comprehensive reviews, as rated by experts. To compare the quality of written review in both conditions, we performed repeated measures ANCOVA to examine the effect of the two conditions on the review length as well as each quality measure. We control the order of experiment condition and the knowledge level of each paper topic as co-variates. As shown in Figure 7, we found that the written reviews in the ReviewFlow condition are statistically significantly more comprehensive reviews than the baseline condition(F = 3.55, p = 0.04 ∗). We found no significant interaction effect between the order of the experiment conditions and no statistically significant differences between the knowledge of the two papers. Because we randomly assigned participants to one of the groups.

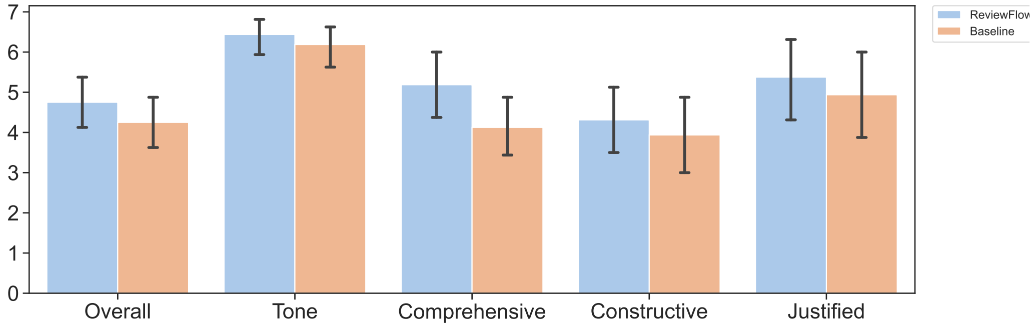


Fig. 7. Experts evaluation on the quality of review written by ReviewFlow versus Baseline.

We also found that 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 items towards each weakness. In addition, we asked the participants to self-rate their satisfaction with the review quality. Participants are more satisfied with the reviews written in the ReviewFlow condition.

	ReviewFlow	Baseline
Num of strengths	2.38(0.19)	1.91 (0.17)
Num of weaknesses	2.62(0.22))	2.08(0.27)
Length(words)	250.8(19.4)	235.8(23.5)
Self-rated satisfaction on the review	4.35(0.4)	4.0(0.34)

Table 1. Proxy of written review quality includes number of strengths, number of weaknesses, and length of the review. Participants self-perceived review quality

6.2 RQ2: Longer interaction duration with ReviewFlow but improves users' self-efficacy

6.2.1 Participants took a longer time to read papers with ReviewFlow but a shorter time on drafting the final review. Our experiment showed that ReviewFlow reduced review writing time but the reading time is higher compared to the baseline. We performed repeated measures of ANCOVA to examine the effect of the two conditions on writing time and control the length of the review as a co-variate. The results did not show a significant difference. However, even though participants spent more time on reading, they reported that "ReviewFlow saved me more time on writing review". One participant reflected on the reasons that "I feel actively receiving these pop-up questions make me spend more time on reading the paper, judging it from multiple dimensions and taking notes, but I think it did save me time on the review drafting since I don't need to go through all the notes again"(P2).

	ReviewFlow	Baseline
reading time (minutes)	26.9 (1.9)	19.2(2.1)
writing time (minutes)	15.5(2.2)	15.7(2.4)

Table 2. Time on tasks in both conditions.

6.2.2 Participants reported to have higher self-efficacy after experiencing ReviewFlow than the Baseline. We further measure whether using the scaffold system can improve novice reviewers' self-efficacy on writing conference peer review and knowledge. 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). Given that the study had a very small sample size, we conducted the Mann- Whitney U Test between each participant's ratings on self-efficacy [42]. Results showed that self-efficacy ratings in the ReviewFlow (M = 4.92, STD = 0.40) is significantly higher than the Baseline (M = 3.92, STD= 0.36, p = 0.05 $*$)

6.2.3 Participants' took more notes in the ReviewFlow and used the features actively. Participants took significantly more notes in the ReviewFlow system (M=12.9, STD = 3.7) than baseline environment (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. All participants used the notes-to-outline features which helped them summarize 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 participants to compare them overall directly. As shown in Figure 8, participants highly prefer ReviewFlow over the baseline environment. 92.9% of participants perceived

	# of participants	Avg frequency	useful	accuracy
Used the section-specific reflection question	14	3.8	5.4	4.9
Used the contextual reflection question on highlight	10	1.6	3.8	4.3
Checked the in-situ citation support	10	1.0	4.6	6.2
Checked the missing citation support	9	1.0	2.5	3.4
Used the summarization notes to high-level outline	16	1.12	6.3	5.2
Expanded the high-level outline to detailed outline	13	1	5.2	4.3

Table 3. Usage of scaffolding features and participants' rating on the usefulness and accuracy. Average frequency is the average number of times the feature was used. The notes-to-outline feature is used by all participants, while some participants did not expand the high-level outline to a detailed outline.

the interface as enjoyable to use. Participants mentioned that the scaffolding features make the review process more engaging and less boring (N=5).

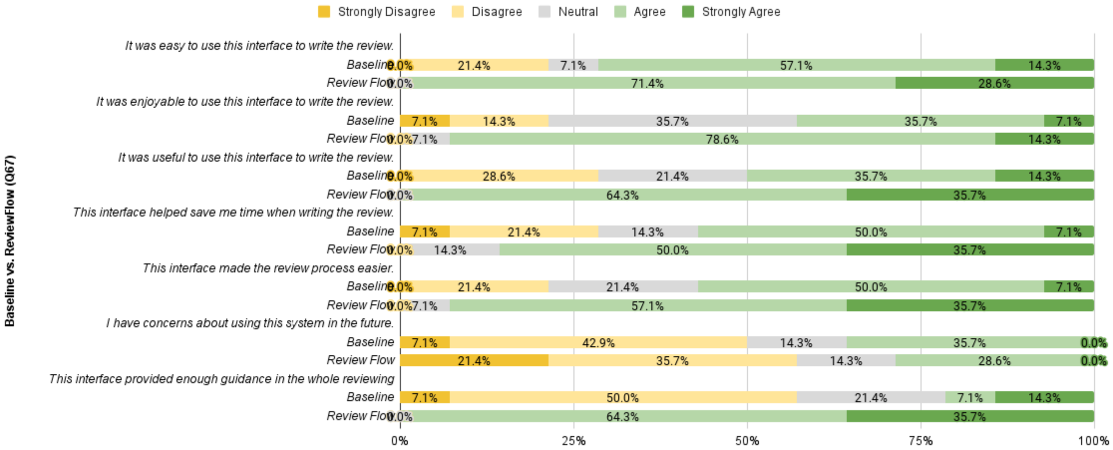


Fig. 8. Participants' perception on two systems

We further measured participants' perceptions of conducting the task together with the system. In the post-survey, we asked participants to rate cognitive workload, such as distraction and engagement [28], feeling of control and collaboration with the interface [70], and perceived learning outcomes with the interface, as shown in Figure 9. The add-on scaffolded features bring in some distraction since 57.1% of the participants believe that the interaction with the interface distracts them from the review. Interestingly, participants reflected that "It is actually not a bad distraction, since you don't want to keep reading the paper and get nothing out of it. This distraction is like a pop-up that keeps telling you about review criteria from different places"(P8). Even though the current system uses AI-generated features, such as notes-to-outline synthesis, all participants agree that they still have the control over writing process. One participant mentioned the reason is that "It just synthesizes what I have already written, and if I want, I can just delete them"(P9).

64.3% of participants have the feeling of collaborating with the interface. Participants identified the collaboration mainly happened in the process of answering reflections on each section by asking and answering questions and

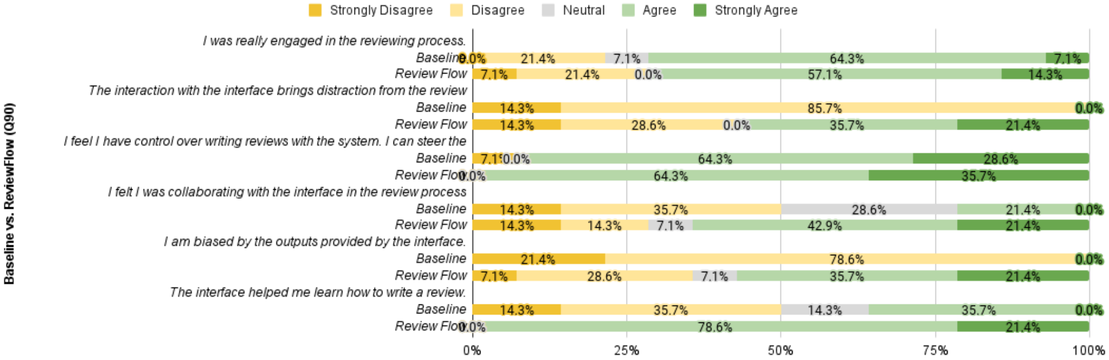


Fig. 9. Participants' perceptions on tasks together with each system

note-to-outline synthesis. The two-step process where participants can first summarize notes into a high-level outline and then edit and expand it provides participants the feeling of "iterating with an assistant"(P8). When we asked participants their reason for being biased by the outputs, most participants thought that the notes synthesis did not bias them in the current session, but they worried that "lazy people might randomly create notes and use the outline to make a decision "(P6). In addition, all participants believe that the ReviewFlow interface helped them learn the peer review process.

6.3.1 Participants have different preferences on the scaffolding strategies. As shown in Table 3, participants perceived that section-specific reflections were more useful and more accurate than the contextual question on the highlighted text. Participants reflected (N=2) that the uncertainty on the contextual questions is high since the questions' quality depends on how much text they highlight. P15 also reflected that "I kind of have a question in my mind when I highlight certain parts of the text, so if the question does not match with the question I was thinking, I feel this is not useful". 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 while reading. The most useful feature perceived by participants is the notes-to-outline synthesis feature where participants reflected that it is "very easy to get a sense on top of my notes"(P3). Participants reflected that the detailed outline bullets contain the same meaning as the high-level outline topic which makes it less useful. We further conducted an in-depth analysis of how participants used and perceived different types of scaffolding strategies:

Contextual reflections cues worked as an expert to guide reviews evaluated from different criteria. Participants described that their process of using the section-specific reflections is slightly different from the process of using context-specific reflections. P11 described that "I first check out the questions real quick before each section, and then I dive into the reading for this section. After I finish reading it, I come back and revisit these questions and reflect on how I feel". While, for contextual questions, "it is more like I feel confused or not sure what I should think about, I will highlight and use those questions to guide me"(P11). P14 described the section-specific question as "an expert who is sitting there that keeps giving you guided questions and prompts you to think"(P14).

Notes-to-outline synthesis helped participants structure their notes according to community practices. All participants used the notes-to-outline synthesis. Participants reflected that they prefer to have the notes in a "structured version" and that they don't need to conduct the mapping by hand (N=5). However, there are three participants did not use the expand button to obtain the detailed outline. Two participants intentionally avoided using it since they were worried the detailed bullets generated by AI may mislead them in their decision-making.

6.3.2 Participants' potential concerns of using ReviewFlow. When we asked participants about their main concerns about using ReviewFlow, they mostly concentrated on 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 is mainly reusing or summarizing 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 be misleading to the reviewer.

7 DISCUSSION

We conducted a within-subjects experiment to explore the potential of using AI scaffolding to support a workflow for novice peer reviewing. In the within-subject experiment, we found that novice reviewers not only preferred ReviewFlwow over baseline system, but they also wrote longer and more comprehensive reviews. Novice reviewers spent more time reading the paper and take more detailed notes. Novice reviewers improved their self-efficacy and perceived the system helped them learn the workflow of reviewing. By rethinking the design goals, we further discuss the potential open questions around creating AI scaffolding in the context of academic review:

7.1 AI-driven vs AI scaffolding vs. Human only

What role should machine intelligence play in this review process? In the current system, we intentionally define the system in the mode of human-driven AI scaffolding. One question here is what role the AI should play in complex cognitive tasks. Recent studies also explored the ways of using LLM to provide feedback to academic researchers and found that LLM feedback can benefit researchers in in earlier stages of manuscript preparation while researchers struggle with an in-depth critique of study methods [39]. AI can play the role of an assistant as well as a collaborator. This indicates that researchers may want to define the role at the beginning to specify what level of automation is beneficial for users.

7.2 Towards human-AI hybrid collaboration for academic review

In the recent discussion of human-AI collaboration, researchers proposed that AI and humans should maintain a "partnership relationship" where AI is designed to fit into the existing human task workflow and assist parts of tasks according to human needs [67]. Our study demonstrated one type of human AI collaboration where AI mainly play the role as an expert to scaffold novices. This could potentially be helpful in various applications, such as learning tasks or collaborative work.

Machine intelligence can potentially scaffold the entire peer review ecosystem and our findings can further guide the design of AI to support other tasks in the academic peer review cycle. For example, machine intelligence can potentially scaffold paper author to better analyze the reviews, write a persuasive rebuttal, and revise the submission more effectively [21]. In the future, we plan to extend our work to design more human-AI collaboration systems where AI scaffolding can support users in academic peer review.

7.3 Ethical considerations of AI scaffolding for academic review

Generative AI and LLMs have been found useful to support writing tasks. This brings more ethical concerns of designing human-AI collaboration systems for writing. First, hallucinations in models may let the system generate factual incorrect information. Models can even induce bias, and misinformation [6]. In addition, directly using AI-generated content in writing artifacts may violate academic integrity and harm authorship. In ReviewFlow, we force the LLM to only use the user notes and highlighted content and synthesize them into an outline instead of paragraphs. We intentionally prevent the situation that reviewers may directly copy and paste the generated text as a review.

8 LIMITATIONS

Our study has several limitations. First, ReviewFlow combined potentially useful scaffolding strategies all in one place. Ideally, to test the effectiveness of each scaffolding strategy, ablation studies are needed. Second, we created a mock scenario and encouraged users to spend about 1 hour writing a review for a short paper. In a real review scenario, from our interview, reviewers can take hours to write a review. Third, we selected as core study materials using HCI short papers. Community review guidelines vary across different fields. The limit size of the scenario and the variance between research communities may bring the limitation of generalizing results to another research area. Last, the LLM we used here is GPT4. With the rapid development of ML and NLP, more advanced models can potentially make the system perform better.

9 CONCLUSION

We conducted formative study with 10 novices and experts to uncover the common practices of expert reviews and found that experts adopted a workflow of annotating, note-taking, and synthesizing notes into reviews. Modeling the expert workflow, we developed ReviewFlow – an AI-driven workflow that scaffolds novices using contextual reflections, in-situ knowledge support, and notes-to-outline synthesis. In within-subject experiments with 16 novice reviewers, we found that ReviewFlow can improve the comprehensiveness of the review and improve novice reviewers' self-efficacy.

REFERENCES

[1] 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.

[2] 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.

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

[4] 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.

[5] 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.

[6] Ali Borji. 2023. A categorical archive of chatgpt failures. *arXiv preprint arXiv:2302.03494* (2023).

[7] 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.

[8] Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel S Weld. 2020. TLDR: Extreme summarization of scientific documents. *arXiv preprint arXiv:2004.15011* (2020).

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

[10] Kathy Charmaz. 2014. *Constructing grounded theory*. sage.

[11] 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.

[12] 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).

[13] 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.

[14] 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.

[15] 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.

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

[17] 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.

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

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

[20] 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*.

[21] 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).

[22] 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.

[23] 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.

[24] 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.

[25] 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].

[26] Kathleen Gomoll. 1990. Some techniques for observing users. *The art of human-computer interface design* (1990), 85–90.

[27] 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.

[28] 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.

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

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

[31] 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.

[32] 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.

[33] 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).

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

[35] 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.

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

[37] 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

[38] Yoonjoo Lee, John Joon Young Chung, Tae Soo Kim, Jean Y Song, and Juho Kim. 2022. Promptiverse: Scalable generation of scaffolding prompts through human-AI hybrid knowledge graph annotation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–18.

[39] 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]

[40] 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.

[41] 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.

[42] Patrick E McKnight and Julius Najab. 2010. Mann-Whitney U Test. *The Corsini encyclopedia of psychology* (2010), 1–1.

[43] 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].

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

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

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

[47] Allen Newell, Herbert Alexander Simon, et al. 1972. *Human problem solving.* Vol. 104. Prentice-hall Englewood Cliffs, NJ.

[48] Milda Norkute, Nadja Herger, Leszek Michalak, Andrew Mulder, and Sally Gao. 2021. Towards explainable AI: Assessing the usefulness and impact of added explainability features in legal document summarization. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–7.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] Napol Rachatasumrit, Gonzalo Ramos, Jina Suh, Rachel Ng, and Christopher Meek. 2021. ForSense: Accelerating Online Research Through Sensemaking Integration and Machine Research Support. In *26th International Conference on Intelligent User Interfaces*. 608–618.

[55] 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

[56] 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.

[57] 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.

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

[59] 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).

[60] 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.

[61] 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.

[62] 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.

[63] 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.

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

[65] Thiemo Wambsganss, Tobias Kueng, Matthias Soellner, and Jan Marco Leimeister. 2021. ArgueTutor: An adaptive dialog-based learning system for argumentation skills. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–13.

[66] Thiemo Wambsganss, Christina Niklaus, Matthias Cetto, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2020. AL: An adaptive learning support system for argumentation skills. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14.

[67] Dakuo Wang, Justin D Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. 2019. Human-AI collaboration in data science: Exploring data scientists' perceptions of automated AI. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.

[68] 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].

[69] 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.

[70] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–22.

[71] 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

[72] 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).

[73] 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).

[74] 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.

[75] 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.

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

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

[78] 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).