Supporting Reviewing Reviews: How HCI Authors Handle Peer Reviews of Manuscripts

Colin Au Yeung
colin.auyeung@ucalgary.ca
University of Calgary
Calgary, Alberta, Canada

Jessi Stark
jtstark@dgp.toronto.edu
University of Toronto
Toronto, Ontario, Canada

Jiannan Li
jiannanli@smu.edu.sg
Singapore Management University
Singapore, Singapore

Fanny Chevalier
fanny@cs.utoronto.edu
University of Toronto
Toronto, Ontario, Canada

Joonsuk Park
park@joonsuk.org
University of Richmond
Richmond, Virgina, USA

Young-Ho Kim
yghokim@younghokim.net
NAVER AI Lab
Seongnam, Gyeonggi, Republic of Korea

Anthony Tang
tonyt@smu.edu.sg
Singapore Management University
Singapore, Singapore

ABSTRACT

Responding to peer reviews is a critical but under-supported stage of academic writing. Authors must interpret reviewer comments, infer underlying concerns, and coordinate revisions across teams. We report findings from interviews with 14 HCI authors that reveal how they engage in this interpretive and collaborative process. Authors distinguish between surface-level content and subtextual meaning in reviews, and often rely on intermediary documents to track issues, assign tasks, and develop response strategies. These documents support sensemaking, communication, and planning, but must be built manually. Our findings suggest that while interpretation of subtext remains a human judgment task, there are clear opportunities for interactive tools to support coordination, document linking, and traceability. We offer design implications for next-generation writing tools, including those powered by language models, that align with authors' workflows and preserve their interpretive agency.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; **Empirical studies in interaction design**; **Interaction design**.

KEYWORDS

academic peer review, qualitative methods, interview study, writing support

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

ACM Reference Format:
Colin Au Yeung, Jessi Stark, Jiannan Li, Fanny Chevalier, Joonsuk Park, Young-Ho Kim, and Anthony Tang. 2018. Supporting Reviewing Reviews: How HCI Authors Handle Peer Reviews of Manuscripts. In *Conference acronym 'XX: Make sure to enter the correct conference title from your rights confirmation email, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 16 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Even in the earliest days of computing systems, researchers and organizations have sought to help people to author documents [15]. For instance, in the famous *"Mother of All Demos,"* Englebart demonstrated, among other things, synchronous collaborative editing software that allowed collaborators to work together on a shared document—an activity for which today's digital tools (Overleaf, Google Docs, Word) provide extensive support.

Yet while current systems support document drafting and collaboration well, far less attention has been paid to the editorial phase of writing, where authors must interpret and respond to critical feedback. In academic publishing, this occurs through peer review, where authors revise manuscripts in light of reviewer assessments [7, 44]. This process is time-consuming, cognitively demanding, and undersupported by current tools. Emerging generative AI tools powered by large language models (LLMs) [9, 30, 40] offer new possibilities for support during this phase. But to design these tools responsibly, we must first understand how authors actually engage with peer reviews: how they interpret reviewer remarks, coordinate with co-authors, and develop revision strategies.

In this study, we focus on academic authors in Human-Computer Interaction (HCI): a field with well-established peer review norms, multi-author collaborations, and high expectations for revision quality. HCI offers a rich starting point for investigating how authors navigate the review process, while providing a consistent disciplinary context. Although practices vary across fields, the challenges of interpreting critique, coordinating responses, and tracking revisions are common across academia.

While prior research has examined collaborative authoring and document preparation, the review response phase, where critique is translated into action, has received little empirical attention. We address this gap by investigating how authors interpret reviewer feedback, coordinate revisions, and organize their response strategies. Our goal is to understand these practices not only to characterize current behaviors, but also to inform the design of tools that support this collaborative, interpretive, and pedagogical stage of academic writing. We ask: *What are HCI authors' practices for working with peer review assessments?* By exploring this question, we identify opportunities for tools that better assist authors during a critical and often overlooked stage of scholarly work.

To investigate this question, we conducted an interview study with 14 academic authors (within HCI) at various career stages. In one-hour interviews conducted over video, we asked participants to walk us through how they had handled peer review comments for two of their previously published papers. In the second half of each interview, we presented a series of video sketches—design probes that illustrated speculative tool capabilities powered by large language models (LLMs). These probes helped elicit reactions to potential system designs and surfaced authors' preferences, concerns, and expectations.

Our findings highlight two central themes. First, authors spend considerable effort interpreting the *subtextual meaning* of reviewer comments—not just what was said, but why it was said and what reviewers may have been pointing to (e.g. unspoken concerns, tacit reviewer intentions, etc.). This interpretation work is essential to planning responses and revisions, yet highly contextual and cognitively demanding. Second, authors rely heavily on *intermediary documents*—custom spreadsheets, notes, and shared outlines—to support interpretation, coordinate across co-authors, and track changes. These documents act as scaffolds for sense-making, communication, and shared memory, but today must be built and maintained manually.

Based on these findings, we identify several design implications for authoring tools. Authors welcomed support for coordination, tracking, and document navigation, especially tools that could extract and link related ideas across artifacts. However, they were cautious about tools that attempted to interpret reviewer intent or provide revision suggestions based on subtext. These tasks were seen as deeply tied to authorship, disciplinary norms, and scholarly judgment.

This paper makes two primary contributions:

- We contribute the first empirical account (to our knowledge) of how HCI authors interpret and coordinate around peer review, and offer concrete design principles for tools that support coordination, contextual feedback, and author-driven interpretation.
- We derive design considerations for next-generation authoring tools, particularly those augmented with LLMs, highlighting where automation can meaningfully support review response and where human judgment must remain central.

Peer review plays a central role in shaping careers, publications, and research trajectories, yet the tools to support this phase of writing remain underdeveloped. Our findings point toward a shift in how this stage is conceptualized: not just as a document revision task, but as a collaborative, interpretive process that spans people, texts, and tools. These insights have implications for tool designers, mentors, early-career researchers, and anyone interested in improving how scholarly work is shaped through critique and revision.

2 RELATED WORK

Peer Review in Academic Writing. A common practice in scholarly manuscript preparation involves incorporating feedback received from peer review to improve and refine a manuscript [7]. Peer reviews are formal critiques of manuscripts. For the editor, they provide an assessment of quality, validity and significance of the work; for authors, they offer an in-depth analysis of the manuscript's strengths, weaknesses, and may contain suggestions for refining the research, the arguments, or the methodological/conceptual limitations [52]. Yet, these review documents are prepared by peers with varying levels of experience and domain expertise [28]. The result is that the documents themselves vary significantly in terms of tone, message, and may also be contradictory in judgement.

Reviewers engage in this service activity to give back and maintain the community [42–44]. Reviewers act as gate-keepers, assessing manuscripts' publication readiness, and so this part of the process is stressful for authors, since decisions can have dramatic career impacts [25, 52]. Furthermore, reviews can vary widely in quality and actionability [28]. Considerable recent effort has explored approaches for improving quality (e.g. [21, 22, 28]) and practices (e.g. [43]), as well as fairness in the peer-review process (e.g. [56]).

While other researchers have explored how the peer review process serves the community and individuals (e.g. [52]), our goal here is to explore how to design tools to support *authors* from the moment they receive these reviews. To our knowledge, there has not been explicit exploration to make authors' workflows with peer reviews smoother. For instance, how can we support authors in understanding the content of the peer reviews? How can we support authors in synthesizing this understanding to develop action plans with co-authors?

Supporting Authors with NLP-Based Tools. Recent advances in NLP have brought considerable attention to the potential of AI tools to support authors beyond simple routine tasks (e.g., spell check), proactive revision support (e.g., sentence rephrasing suggestions), and collaborative editing (e.g., shared change tracking), all of which are now standard features of modern document editing software.

Large language models (LLMs) take input text and repeatedly predict text that naturally continues the input. Because these models have now been trained on extensive troves of text, they have reached a high level of apparent capability. Breathless media reports, followed by some academic work (e.g. [9, 30, 40]), have demonstrated that general-purpose LLMs can perform remarkably well in summarizing large quantities of text, synthesizing ideas in text, and generating text with only limited prompting [35]. In academic terms, LLMs can be good test-takers. They have been demonstrated to effectively identify correct responses to multiple-choice questions [40], as well as prepare both short answer responses and longer essays [30].

Recent work has begun exploring the design of systems to support the writing process, including the the design of language models to supporting writing [27, 33] and the ways in which writers may interface with LLM-based support tools [12, 13, 48, 49, 62]. In addition, researcher have investigated how a range of different writing domains that may have different needs from LLM-based tools; such as script-writing [36, 57], story writing [3, 41], poetry [5], and journalism [46]. Researchers have also explored how LLM-based tools can support specific authoring tasks; such as inspiration generations [10, 14, 19], idea organization [60, 61], metaphor creation [18, 31], word selection [17], character creation [47], and background research [1].

Beyond the initial writing process, LLM-based tools' ability to synthesize and summarize text have also shown promising in support authors in reflecting and revising their work [11, 66, 67]. For example, Synthia [68] supports authors by breaking multiple sources of feedback and uses visual marks to situate that feedback within the text document. On the other hand, Benharrak et al. [2] explore the use of LLMs for generation of audience specific feedback.

Despite their promise, work such as Gero et al. [20] and Varanasi et al. [59] have raised questions about the authorship tension between tool and author. In response to these concerns, a number of works have explore how tools may support ownership and transparency in the use of LLMs. For example, Wr-AI-ter [65] which explored how LLM-based support tools may be designed to preserve author's sense of ownership over their work. Similarly, Hoque et al. [39] explored the use of interactive visualization to support transparency in the use of LLMs. Another challenge is the risk of LLMs introducing inaccuracy through hallucinations [26, 29]. Work such as Laban et al. [32] how LLM-based tools may be designed to allow authors to verify the accuracy of LLM suggestions.

Academic writing has been no exception with regards to the impact of LLMs on writing. For example, in response to whether and the extent to which authors can use generative tools in manuscript preparation, publisher have developed new policies clarifying the acceptability of their use (see an analysis in [39]). On the hand, recent work has began exploring how LLMs can support different parts of the academic writing process. Rofferello et al. [37] examined how computer science researchers used LLM-based AI tools to support their writing. Similarly, research is increasingly investigating how LLMs can support writing peer reviews (e.g., [34, 51, 53, 54]). For instances, Metawriter [55] which explores the use of LLMs to support the writing of meta reviews. However, while prior work has explored supporting the writing of academic peer reviews, little focus has been on how LLMs can support authors' in responding to peer reviews. Our work explores how LLM-based tools may be designed to be incorporated into authors' workflows in dealing with peer review comments.

3 INTERVIEW STUDY

To understand how HCI authors handle and process manuscript reviews, we interviewed 14 authors who primarily submit to HCI-related venues. Our goal was to understand the steps that the authors took from the moment they received the reviews to the point of sending out their response: how they thought about the reviews, how they conceptualized ideas in the reviews, how they determined what actions to take, and how those actions were executed. We asked participants to show us "intermediary documents" that they may have used in preparing the response—for example, documents that had been prepared or highlighted (e.g. the reviews, a response letter, a spreadsheet, etc.), and asked them to guide us through how they developed these documents, why they developed them the way they did, and how they worked with co-authors through them.

We were also interested in participants' thoughts on next-generation tool support for this part of their authoring process. We used speculative video sketches that illustrated potential LLM-enhanced capabilities. To focus this thinking, we developed video sketches of design ideas for writing tools based on some prior work (e.g. [69]) and current/near-future capabilities of LLMs. We used these video sketches as probes to elicit feedback and ideas (see details in Appendix A).

In the recruitment for our study, we focused specifically on researchers engaged in HCI (and HCI-adjacent) research. We focus on HCI authors because the field has well-defined review norms, collaborative authorship practices, and a high rate of revise-and-resubmit decisions. These conditions create a rich context for studying review response. While some workflows may be specific to HCI, the interpretive and collaborative challenges we uncover should be common to scholarly writing more broadly.

Participants. We recruited 14 HCI authors (7 males, 7 females) with a range of prior publication history (3–182 peer-reviewed publications; median = 39, IQR = 98). We advertised our study on social media and our personal network, and used the snowball sampling approach. The primary inclusion criteria was that participants should have published at least two peer-reviewed manuscripts. Our participants published widely across a wide variety of venues within HCI and related disciplines. Further information about our participants is summarized in Table 1. Participants were compensated $20 CAD for their time. We discuss the limitations of our recruitment approach and the resulting sample in section 5.

Procedure. We invited participants to 1-hour Zoom interviews with two parts: a semi-structured phase (45 min), and a design probe phase (15 min). *Part 1: Semi-Structured Discussion.* This focused around two published manuscripts that participants selected. Here, we asked participants to share with us the practices that they used: (1) to understand the reviews—both individually and collectively with their co-authors; (2) to develop a plan of action (how they identified what would be done in response to the reviews), and (3) to develop the response. Over screen share, our participants shared and led discussions about "intermediary documents" that they had created to document ideas and discussions with their co-authors. Participants shared these documents over screen share, highlighting and discussing examples with us. *Part 2: Design Probe.* Participants responded to video sketches of fictional tools where the tools imagined different ways that academic authors could be supported in their process of responding to academic peer reviews. Each video sketch was played over the screen share (this time from the interviewer's computer), which illustrated how the tool worked, and how it might support academic authors' workflow. Participants then discussed aspects of the tool that were resonant and discordant with their goals and practices. Our study protocol was approved by our institutional research ethics board (Ref. REDACTED FOR REVIEW).

Table 1: Interview participant information

ID	Title	Years Active	Publications	HCI-Related Venues	Role on Papers Discussed
P1	PhD Candidate	4-10	3-10	CHI, ASSETS	First Author, First Author
P2	Full Professor	11-20	51-100	VIS, InfoVis	Supervisor, Collaborator
P3	PhD Student	4-10	3-10	ICT4D, HICCS	First Author, Second Author
P4	PhD Candidate	4-10	3-10	SIGSAC	First Author, First Author
P5	Full Professor	21+	101+	IMWUT, CHI	Supervisor, Supervisor
P6	Full Professor	21+	51-100	CHI, IUI	Supervisor, Supervisor
P7	PhD Candidate	4-10	3-10	VIS, CHI	First Author, First Author
P8	PhD Student	4-10	3-10	CHI	Second Author, First Author
P9	Professor Emeritus	21+	51-100	CHI, UIST	Supervisor, Supervisor
P10	Full Professor	11-20	101+	CHI, DIS, CSCW	Supervisor, Supervisor
P11	Assistant Professor	11-20	11-50	ACL, IUI	First Author, First Author
P12	Full Professor	11-20	51-100	CHI, ISS	Supervisor, Supervisor
P13	Associate Professor	11-20	11-50	CHI, DIS	Supervisor, Supervisor
P14	Associate Professor	11-20	51-100	VIS, CHI	Supervisor, Supervisor

Design Concepts and Video Sketches. Based on our experiences as authors, we ideated about potential tool support for this aspect of the academic authoring process. In particular, we focus on tools that may utilize recent advances in NLP given their potential ability to support the language interpretation and writing work involved in handling peer reviews. In our process, we developed several ideas and sketches with the aim of identifying tasks such tools could support, as well as dimensions along which these tool designs might sit. Based on this process, we developed a set of provisional organizing design dimensions for tools that could support the review handling process.

- *Tasks.* Based on our own experiences as authors, we identified five tasks such tools could support: (1) summarizing and rewording individual reviews; (2) categorizing issues and ideas across reviews—that is, identifying when several reviewers identified similar concerns; (3) proposing potential resolutions to these issues, and identifying where in the source manuscript these issues might occur; (4) describing differences between the original manuscript and revised manuscript; (5) summarizing these differences in relation to issues identified in the reviews.
- *Conversational vs. Document-Focused Interaction.* Chat-GPT and current LLM-driven tools allow users to use natural language prompts, whereas most authoring tools have a document-focused structure. We explored how a chat-based interface might operate and contrasted this with document-focused interaction, where proposed revisions and suggestions are situated within the documents themselves (e.g. as annotations to the reviews, the intermediary document, the response letter, or the manuscript itself).

For our study, we designed four video sketches that exemplified different aspects of the design concepts, illustrated in Figure 1 (a lengthier description of each video sketch is in Appendix A). Each sketch featured a workflow or task involving the review handling process, and how the tool would support that process. The narrated videos were 1:30-2:30 in length, and involved fictional reviews and *lorem ipsum* text.

Data Collection and Analysis. We collected interview data and recorded video of the participants' screens while they were sharing the intermediary documents—a total of 910 minutes of interview data (some interviews ran long), and about 25 distinct co-authoring teams. Interview audio was transcribed. We used Reflexive Thematic Analysis (RTA) as our analytic approach [6], where we iteratively coded and explored the interview data, relying on both the data and our own interpretations of the data—as academic researchers (and authors ourselves!), and as tool builders within the HCI/NLP research space. Based on these, we developed codes that reflected participants' responses, and iteratively synthesized these codes into themes. We then organized these around prominent responses (emotional labour, subtextual interpretation, intermediary

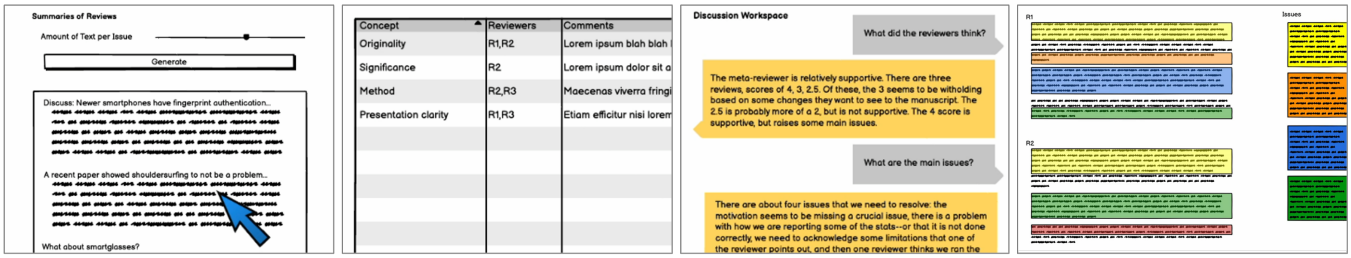

Figure 1: From left to right, Sketch 1 generated a summary of the reviews, along with a synthesis of which reviewer commented on the issue and how. This summary could be edited, such that the author could ask the tool to further "fill in" information about an issue. Sketch 2 explored presented summaries in a tabular rubric (e.g. clarity, organization, etc.), and the user could add/remove row items for the system to consider. Sketch 3 illustrated a dialogue-style interaction, where the author could ask questions about an interpretation or summary from the tool, for example, to ask for a clarification or an alternate interpretation of a review comment. Sketch 4 used a document-centric interaction, where similar issues in the review text were colour-coded with highlighting; this was colour coding was also carried through to the manuscript (i.e. where it would need to revised).

documents, collaboration, mentorship) that characterize our interpretation of the data. Our analysis approach asked us to consider our own practices (as academic authors), and to carefully reflect on practical differences that emerge out of habit, lessons learned, and "folk knowledge." The purpose of using the RTA approach was not to arrive at a comprehensive theory of how HCI academics handle manuscript reviews, but rather a more focused exploration of themes and patterns in our participants' practices.

4 FINDINGS

We organize our findings around three major themes: (1) how authors make sense of reviewer remarks, particularly the effort involved in interpreting subtext and managing emotional responses; (2) how intermediary documents are used to support understanding, communication, and coordination among co-authors; and (3) how authors envision and respond to future tool support, including where they see value and where they express skepticism.

Participant reflections span both phases of our interviews: discussions of past review experiences, and reactions to speculative tool designs. When relevant, we highlight which phase a particular insight came from. Collectively, these findings reveal opportunities for designing tools that support not just document revision, but the broader process of review response as collaborative and interpretive work.

4.1 Sense-Making in Reviews

Participants described working through the reviews as a *"sense-making"* process (colloquially). While the ultimate goal was to determine how to respond to reviewer remarks, an important sub-goal was to interpret the remarks themselves, and participants explained that this was not always straightforward. Participants explained that authors needed to closely read reviewers' words, but that the words themselves might also be deceiving: a literal or surface interpretation of the remarks may not clarify the issue or identify the best resolution to the issue.

4.1.1 Emotional Responses to Reviews. One of the first challenges that authors faced when making sense of reviews is working through their initial emotional responses to reviews. Authors described that reviews can be emotionally charged documents to deal with. For some authors like P3, who described being excited to see others engage with their work and research interests, receiving reviews can be a positive experience. However, for other authors, the initial reaction to viewing reviews can be a difficult negative experience. For example, P1 describes *"... I was anxiously waiting on it. I was on the bus, constantly checking my phone, and bam there it was. And the scores and the reviews are not good, so I really... it was an emotional blow. That first time was really hard."* While these emotional responses were most pronounced among junior authors who were more likely to be the primary author on a manuscript and closest to the work, even senior authors who were more likely to serve as a supervisory role on a manuscript expressed that they had emotional responses to reviews. For instance, P12 describes that *"I, myself, read the reviews the first day they come out and I think I also sometimes need to give myself a bit of time to not be mad at the reviewers. I think I'm faster than the students at this now because when it's a student lead paper, it's very much I'm feeling protective of the student and that's why I'm upset, but at the same time, I'm thinking, 'Ah we missed this,' and I know that we missed this immediately, where as they are probably owning it a little bit more and take it a bit more personally."*

For the more senior authors who often served as supervisors to the primary author on the manuscript, one of the important initial processes that they engaged in was helping to manage the emotional response of the primary author. Senior authors had a variety of ways that they chose to handle this process. For P6 and P10, one of the ways that they managed this response was by being the first author to engage with the reviews, allowing them to manage that initial emotional response, as P10 explains, *"I'm setting it up. I'm trying to set an emotional response tone as to how things should go. I'm trying to let them know that, 'Hey you know, things aren't all totally bad,' or give them a bit of optimism... I think maybe the point is twofold. One, to set the direction of how we're going to work on the response, and two, to set the emotional mood of the rest of the process."* On the other hand, for P12, it was important for them to give time and space to the primary author to have the emotional

response to the reviews before coming back to the reviews with a clearer head, as they describe, *"But with people who have never seen reviews before, I actually do have a process where I say, 'Take a day or two and it's okay to be like mad at the reviewers,' and it's even useful sometimes to write a document that says like, 'This is why you're idiots.' And then I tell them it's okay then after those two days, go back and read the reviews again with a more critical eye and say, 'What are they really saying about your work?' and, 'Is there any validity to it, even if the way they said it wasn't the best way to say it?"* Ultimately, for the senior authors, the goal was not to prevent their junior authors from having emotional responses to reviews, but to help scaffold their initial interactions with reviews such that the junior author would be able to interact with the reviews in a constructive manner.

4.1.2 Surface vs. Subtextual Meaning in Reviews. One of the important challenges for authors in making sense of reviews was unraveling the distinction between the surface-level meaning that is conveyed via the words in the reviews and the deeper implied meaning (i.e. the subtextual meaning).

Participants described paying close attention to how reviewers' remarks were worded. P11 reports *"I rely a lot on how they phrase things, and other elements of nuance to interpret what they're actually telling me,"* since the importance of a particular comment or issue may be conveyed in a single word or phrase. Similarly P3 explains, *"The way in which they say it helps provoke a direction that I could explore or think about that I hadn't thought about."*

At the same time, participants indicated they would need to *"read between the lines,"* since how a reviewer worded a remark may obscure the meaning [P10, P12]. Sometimes, this would be out of kindness: *"The reviews are quite a diplomatic document. Some people sometimes don't say the literal meanings with the worst [words]. It could look positive, but the negativity was veiled. They all hide this through some wording"* [P2]. One participant likened this to other diplomatic writing: *"[reviews contain] very subtle wording, where it's a bit like reading a reference letter. They might be saying something kind of nice, but actually they're very critical"* [P1]. This means that while the specific choice of words that reviewers use to express their ideas is important, authors are aware that there is nuance and subtlety that may veil the meaning for an inexperienced author.

Authors relied on their experience or the experience of their co-authors to make sense of these nuances. For example, P10 refers to his experience and tacit knowledge when reading reviews, *"Because I've seen so many of these, and I've been on the other side of the table so much, I just have a sense for what the person is really getting at... a reviewer will sometimes be complaining about a detail when really there's a more fundamental problem that they're dancing around. And because I have a fair bit of experience reading these reviews and writing papers, too, I can sometimes read that."* This resonates with P7's commentary, who, as a less experienced author, noted *"Yeah, there's definitely instances where I read a reviewer comment, and I think, oh, they want this. But then my supervisor says, oh, they actually want this other thing"* [P7].

Participants viewed this communication and *"understanding what is meant"* as a two-way challenge: their hope was that the reviewers understood what they, as authors, were trying to express, and now it was their turn to try to understand what the reviewers were trying to express to them. P12: *"[Authors] can misinterpret the underlying meaning of what a [review] comment is actually about. I'm realizing that it's because a review, in and of itself, is also an imperfect articulation of what somebody is thinking. And it may be the case that they haven't really figured out what it is that they're thinking. They're just giving you the reaction."* This suggests that experienced authors view review remarks and their surface level meaning as indicators for the problematic aspects of a manuscript, rather than necessarily being a definitive accounting of the actual problem that the reviewer was identifying. That is, that the reviewer's intent (i.e. the subtextual meaning) may not be perfectly captured by the review remark itself.

4.1.3 Reviewer Suggestions. Even though reviewers had the best of intentions in providing suggestions to improve the manuscript, authors explained that these suggestions might not necessarily address the fundamental issues. For instance, P12 explains, *"Sometimes [reviewers] give us a fix that doesn't make any sense, because they didn't [understand our idea]. And so you have to slow the students down, who want to do what [the reviewer] said to do. Instead, you have to say, 'No, what they're asking us to do is not helpful for us.'"* The main issue seemed to be that sometimes, reviewers might not be expressing the issue clearly within their surface-level remarks, as expressed by P11: *"It's important to address the underlying issue. Someone could point something out as being problematic. But what they're pointing out, and the way in which they're pointing it out is sometimes really symptomatic of something else."*

P10 provides an example of this, where a reviewer might provide a paragraph of questions about the methods being described in the manuscript: *"[Quoting a hypothetical reviewer] 'Did you do a second round of coding, or have a second coder? Were you properly balancing different conditions in the study? Should have you done method X?' So maybe there's a paragraph, with lots of little questions, and it goes on about the methods. I would probably step back and say, from my experience, they're probably not so concerned about whether we did all these little things... [Instead] the bigger problem is that I've not really done a great job of explaining what we did for the methods. So I'm not going to worry so much about these specific questions; instead, we just need to rewrite this section better."* Here, we see that a surface-level interpretation of the reviewer remarks might have helped, but did not address the central issue; rather, the subtextual meaning of the reviewer remarks were that the section was difficult to interpret and understand.

Similarly, P11 provides an example where reviewers critiqued a theoretical lens the authors were using. *"The moment I added the picture [to explain the different theories], the complaints went away. The reviewers didn't say specifically to add the picture, but they're so used to that type of representation that without it, they were having trouble keeping track of all the theories and interactions."* The uncertainty expressed by the reviewers in their remarks did not address the central issue, which in this case was better addressed through a diagram rather than more explanatory text in the manuscript.

4.2 The Role of Intermediary Documents

How authors processed reviews varied significantly both between authors, and between different manuscripts of the same author. While authors had their own varied practices in terms of how they

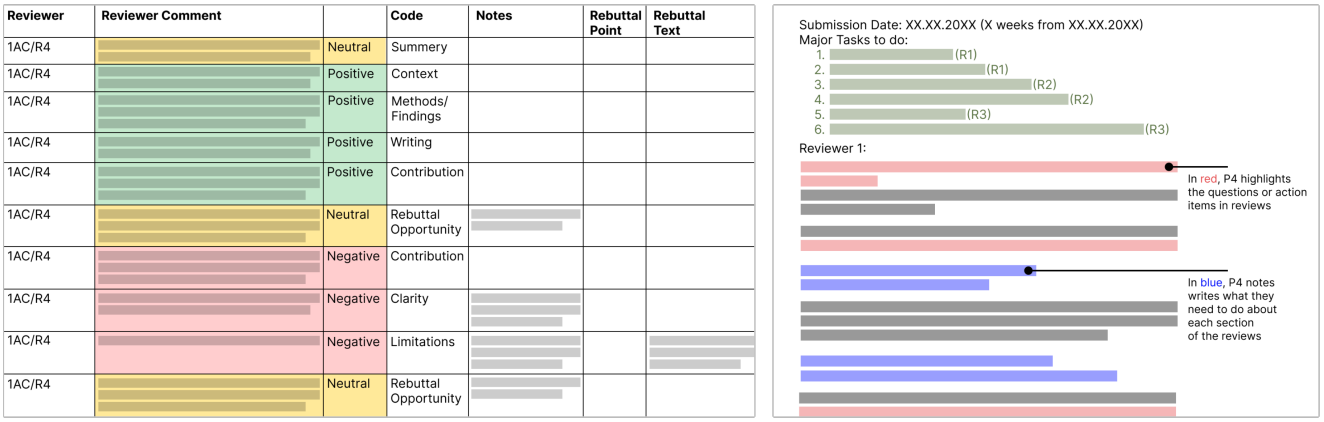

Figure 2: Left: P1's spreadsheet that she used to interpret review and plan their response. The columns included the original review text, the reviewer name (e.g. R1, R2), P1's short-form interpretation of the review text, a "category" (her own typology), a proposed fix field, and an additional comments field. Right: P4's working document where they had copied all of the text of their reviews into. In red text, they highlights the questions or issues raises in the review text, In blue text, they added into the reviews what they needed to do with regards to each issue raised in the reviews.

approached processing reviews, exactly how that process played out in handling particular reviews often varied depending on a range of factors. One key factor was the dynamics of the co-author team; for example, a senior author may have overarching practices in how they handle reviews; however, the particulars of how those practices occur may be shaped by the working styles of the other co-authors in the team. On the other hand, the reviews themselves could also shape the process: more favourable reviews may demand a less intensive process from the author, or the structure of the reviews might help inform the structure that authors used to interpret and respond to reviews. The documents that authors used in this process of responding to reviews were equally varied in style and number. While some documents became final output documents (e.g. response letter or revised manuscript), authors also created intermediary documents that might only ever be used "internally" by the author and co-authors. Intermediary documents served two important roles in how authors processed reviews: (1) as tools for helping authors to *make sense* of how to address the reviewer remarks; (2) as tools for helping authors *communicate and coordinate* collaborate with their co-author teams.

4.2.1 For Sense-Making. As we discussed previously, one of the major processes that authors undergo when handling reviews is the process of making sense of the review remarks, and this process can be challenging as authors identify layers of meaning within the reviews. Here we turn our attention to how participants created intermediary documents to support this process of making sense of reviews. We explore this as a possible site for technology augmentation. In our sample, participants had a number of strategies that they used to approach dissecting reviews into actionable items.

Spreadsheets. One of the strategies that participants took to understand reviews was to decompose the reviews into individual points using the structure of a spreadsheet. For instance, P1 showed us a spreadsheet where she had broken down each comment/observation from the reviews into their own row, ordered in the way they appeared in the review documents (shown in Figure 2 left). In P1's process, she used the "category" field to group related issues—both as an interpretive approach, as well as a pragmatic one for the purposes of developing a response to the reviews. P5 describes a similar process that their student went through where on their initial pass of the reviews: they broke down each of the review documents into smaller, individual review remarks into their own rows in a spreadsheet, categorizing each remark by what they thought their sentiment was, and then assigned each comment a similar "category" typology as P1. This strategy helped authors to work through the reviews on an issue-by-issue basis, understand what the reviewer meant in each comment, isolate which comments were actually actionable, and determine what should be done about them.

In-Situ Highlighting and Comments. Another strategy that authors used to approach making sense of reviews was to work with the reviews in-situ. Rather than breaking apart the reviews into individual comments, some participants worked directly within a copy of the review text. For example, on her initial pass of reviews, P6 used the features of a PDF reader to *"highlight what I [thought] we [needed] to address"*, and left comments that articulate their initial thoughts as well as open questions that they needed to discuss with their co-author. Similarly, P1 discussed how the primary author on one of their papers copied the review text into a Google Doc, and used the highlight feature to leave comments. For some (P1, P6), this was an unstructured approach; others would go about this in a much more structured way. P4, for instance, copied the review text into a separate document, and used red text to highlight the questions or suggestions in the reviews (illustrated in Figure 2, right). P4 then inserted blue text inside of the reviews to indicate things that they needed to do in response to each of those questions or suggestions. While this strategy required authors to work with the whole text of the reviews at once, it meant that the actionable items remained within the context of the review document.

Drafting the Response. Other authors would prepare a response letter as a way of organizing their thinking about how to address reviewer comments in revisions. P10 explained that they often first have their student draft an initial version of the summary of a changes document that would get submitted with the revisions, *"Basically you can see the student has summarized each of the main concerns with a heading, and then the concern, and then a description of what was done to address it. [This was written] in a sort of past tense, as though we had already done the work, but really it was a plan. It was a plan of what we would do to address the concerns."* Similarly, P7 worked within a Google Doc with all the reviews, where they *"[identified] each individual reviewer's concerns by reading their full reviews, parsing out what they're saying point-by-point"*, summarizing each of those points, and drafting their response to each of those points. He explained that this document that began as an intermediary working document would ultimately transform into the final document that would be submitted alongside the updated manuscript. For both P7 and P10, the point of drafting and iterating on these response letters prior to making changes to the manuscript itself was to synthesize an understanding of the reviews, and developing a plan of action that all of the co-authors agree upon. P7 describes that, *"I think in an ideal world, we would [first] make this document. And then we would say all the things that we're gonna do, and then just change the Overleaf document and then submit."*

4.2.2 For Communication And Coordination. While many of the papers that our participant described had a primary author who led the work, our participants noted that it was vanishingly rare that they would be the only author on a paper—no participant shared a single authored paper. Thus intermediary documents also served an important role in supporting communication and coordination between the co-authors. Every participant described these living documents as the centre of the discussion—be they synchronous or asynchronous—where they would track points of uncertainty or issues that would need to be discussed with the authoring team. These documents include records of suggestions, discussions and decisions about how issues were to be conceptualized, how they might be addressed, as well as who might be assigned to complete each task.

Supporting Meetings. Intermediary documents often served to support synchronous discussion within their co-author teams. Participants described that they would generate documents as a way of synthesizing what they needed to discuss with their collaborators. For instance, in P6's initial pass of the reviews, she focused on identifying challenging or contentious issues, so she knew what she needed to discuss with her co-authors during their synchronous meeting. Similarly, P4 would make note within the reviews about which points he needed to discuss with his supervisor. The documents themselves also often became things for the co-authors to talk over. P10 described that he would have his students draft an initial summary of the reviews prior to meeting with them because he felt that it was important, *"to enforce the fact that people need to do some homework first before we meet and talk about it."* These documents then became a way for authors to give and receive concrete feedback from their co-authors. For example, P8 explained that by going over a spreadsheet made by his student, he could identify where their student had misconceptions about the reviews and provide feedback on how they should interpret each review remark. Intermediary documents also served as a site for participants to take notes given the otherwise ephemeral discussion they were having. For instance, in the document where she was making sense of what reviewers had said, P1 would also annotate in red: *"[Those are] our conversation and my notes of our discussion."*

Supporting Asynchronous Discussions. The intermediary documents also served as site of asynchronous communication and coordination. Collaborators commonly used the commenting functionality in either Microsoft Word or Google Docs (6 participants showed or mentioned using comments). In some cases, these comments served as one-off pieces of feedback. For example, a comment in one of P6's documents simply noted to the primary author to, *"Keep the same active sentence structure as the other points."* In other cases, authors used comments to coordinate with the other co-authors, for example, to track which action items had been completed, and which items still needed to be handled. In these cases, the use of comments was sometimes accompanied by TODO lists. P2 and P6 showed how their co-author have written up TODO lists in the working document (shown in Figure 3). Finally, authors also used the comment feature as a way to have asynchronous discussions that were situated around the content that they were discussing. For example, in one of the documents that P2 shared, two of the co-authors engaged in a back and forth discussion about possible approaches to addressing a reviewer comment.

For some authors, the intermediate document and the asynchronous communication in it was, itself, the predominant way that a co-author team communicated. For example, P6 described a case of a manuscript having been submitted and rejected over several years. In the most recent iteration, the authors did not even meet synchronously to discuss how to handle the most recent set of reviews; instead, they opted to discuss strictly through the working draft document.

4.3 Implications for Tool Support

In the second part of our interviews, we explored with participants what kinds of tool support might help authors during the review handling process. We anchored this discussion in both their current workflows and reactions to four design probe sketches, each illustrating different interaction models and AI-driven capabilities. From these conversations, three clear implications emerged for the design of future tools: (1) authors need better support for tracking and coordinating information scattered across documents, (2) authors are skeptical of tools that attempt to interpret subtextual meaning in reviews, and (3) authors strongly prefer interactions that are situated within the review documents themselves, rather than abstracted or chat-based interfaces. We expand on each of these themes below.

4.3.1 Support Connecting Scattered Information. A common challenge was that while intermediary documents acted as hubs of communication and coordination, they were also not directly linked to the actual source documents (reviews, manuscript, revised manuscript, response letter). Thus, there was still often a need to go back to the original documents to locate things, or to ensure that tasks had been completed. P1 explains this difficulty in terms of finding information referred to from the reviews: *"Sometimes the*

Figure 3: P2's working document where their co-author created a TODO list for the team. The team used the comments features to either mark when items were completed or discuss the tasks.

meta reviewer has said something, and you go and search and search for where is it? Who said that? Where and how is it mentioned?" P10 describes this challenge in terms of ensuring that all reviewer comments have been adequately referenced and addressed in the final response document, *"Even after we have this document, and we're okay with it before the paper is accepted, I will go back through the reviews again to ensure for each item: Do we cover this? So it is almost like a checklist at the end. It's easy to miss something."*

Many participants felt that tool support with the intermediary documents would be valuable to help organization and coordination. Whereas the source documents (review and manuscript) were linear, Sketch 2 illustrated the idea of being able to pull out related ideas together. P6 *"This would be helpful to categorize things the reviewers said, and then to help us with [finding related ideas in] the paper."* Similarly P6 suggested, *"I hate always saying 'Who said that?' and having to go back to remember who said what in reviews, so that you could acknowledge them in the response letter—I can't stand doing that. It's super error prone too."*

This suggests that a key design area for future tools is supporting authors in consolidating the information that is scattered across documents. While authors currently use intermediary documents as hubs for organizing revision efforts, these documents still require continual manual effort from authors to link information between the reviews, the documents they use for communication, and the resulting manuscript. Future tools may be able to support authors in finding related ideas across documents, visualizing that corresponding information, and maintain continuity across documents.

4.3.2 Supporting, Not Performing, Interpretation. In response to the design probes, participants largely balked at the notion of a tool interpreting review remarks on their behalf. In particular, participants did not consider tools as being capable of ascertaining the subtextual meaning of reviewer remarks (or at the very least, did not trust them to do so). For instance, P10 explains, *"My worry about the tool automatically doing grouping [of review comments] for me is that I doubt the system has accumulated the same level of tacit knowledge and understanding that I have. And so I worry that it would be missing out on something. I worry it misses out on the tone, and the ordering of reviewer comments."* This resonates with P3's view: *"My apprehension would be whether [the organizational labels] are the same as what I would write, and whether this interpretation captured everything that reviewer actually had to say."* Similarly, P6 suggests, *"I don't know that I would use that to drive either my letter [to the editor] or the edits [to the manuscript], because it might not get it the way I would."* In general, in response to functionality in the design probes where the tool offered a subtextual interpretation of the original review text, participants explained they would not feel comfortable using it. Several participants were skeptical about a tool automatically interpreting and extracting reviewer intentions correctly, as described by P1: *"My guess is that it would be super hard to extract automatically."*

Participants pointed out that how they interpreted and responded to reviews was an important part of how they thought of themselves as researchers. Thus, giving this task up to a tool was difficult. P10 explains, *"I would just worry that if I use this, I will focus on*

what it's generating, and I will miss out on the other stuff that comes from the actual reviews themselves."

On the other hand, several participants proposed the idea of using tools as a "checking" device within the context of their authoring workflow. Here, the tool would provide an additional assessment (e.g. like a spell checker) that an author could use to see whether they had interpreted everything well. *"I think that having a program go through and try to do this annotation of reviews makes some sense, but I think that the manual aspect of it is what's going to be most useful to most authors. I would wonder if prudent people would do it manually first, and then maybe do the automated one to see if they missed anything they've got all these things"* [P9]. Similarly, P10 described this within the context of training students, *"If I want to train a student, I would want them to go through the process of handling reviews on their own first, and then use this as a tool to go back and check to see that they're doing a good job, so kind of like a validation tool."* Finally, P4 conceived of the tool as a way to provide another perspective on the reviews, *"One thing also I take from this idea is to use the technology to see if we can combine it with what I do, and improve my understanding of the reviews."* In these cases, the clear intention is to supplement one's own process of determining subtextual meaning, rather than relying on the tool to do this.

This suggests that future tools should explore not how they might automate interpretation of reviews, but rather how they might support authors in the task of interpreting reviews. One direction may be exploring how tools might support the micro-tasks involved in interpreting reviews rather than the process as a whole. For example, future tools may explore highlighting ideas repeated through reviews or supporting authors in comparing their interpretation to the original reviews.

4.3.3 Support Contextually Situated Interactions. Participants all preferred document-focused interaction as expressed in Sketch 4 over the chat-based and prompt-based interfaces of Sketches 1-3. In part, this had to do with trusting the algorithm underlying the tool, but even more so, participants wanted to see the provenance of ideas within the context of the source documents themselves. For instance, in response to Sketch 2, P4 indicated, *"I want the annotations to appear in the Review itself, like in the margins. This would be useful because I want to see where it is actually occurring."* Similarly, P10 explained that this was important to build trust in the system: *"I would prefer to be able to review concepts merged together—that can be really valuable when sections are collapsed together. But I still want to see the linkages and the concepts—maybe also within the actual reviews."*

When participants saw Sketch 4, which presented the tool's suggestions within the context of the documents, this was a revelation. The tool highlighted and annotated the raw review text (adding ideas as opposed to removing text as in Sketches 1-3), as it provided context to interpret and build trust in the tool. P10 enthusiastically explained, *"I really like this in particular, because it's in the context of the actual views. I love the ability to visually see the different colors of the areas, the problems, and then the linkage to the actual document itself is quite nice as well."* In general, participants felt that Sketch 4, which presented the tool as augmentations to the documents, more closely aligned with their existing practices (i.e. highlighting review remarks for discussion).

Participants wanted the interaction with the tool to be bidirectional in that they would be able to provide the system with feedback about its performance, with the idea that it could, over time, improve its assessment of reviewer intention (i.e. the subtextual meaning). Sketch 4 suggests that users can re-classify or rename issues—a way of correcting the system, or providing it with this feedback. This approach resonated with participants who had been skeptical of the "snippet" approaches illustrated in Sketches 1-3. For instance, P1 suggests, *"If you let me add or remove text, or change its assessments: this would give me more confidence."* P2 agrees with this approach, *"I really like the way it is organized where the interpretation was backed by the reviewer number. So, I can always double check and correct it if anything needs to be adjusted."*

This suggests that tools may be best utilized by authors when they are contextually situated within the working documents of authors rather the separate chat-based interfaces common within current LLM tools such as ChatGPT.

4.4 Summary of Findings

Our findings reveal that responding to peer review is more than an act of document revision, rather, it is an interpretive, emotional, and collaborative process.

Authors emphasized the effort involved in making sense of reviews, particularly when inferring subtextual meaning or managing initial emotional responses. These interpretive judgments were seen as essential to developing appropriate revision strategies. To support this process, authors routinely created intermediary documents (such as annotated drafts, spreadsheets, and draft response letters) to aid in understanding reviewer concerns, track changes, and coordinate with co-authors. These documents also acted as shared memory and planning scaffolds.

When reacting to speculative tool designs, participants welcomed features that supported organization, traceability, and contextual integration into their workflows. However, they were skeptical of tools that attempted to interpret reviewer intent, favoring designs that preserved author control and judgment.

Together, these findings highlight opportunities for tool support that respects authors' interpretive expertise while easing the logistical and collaborative burdens of handling peer review.

5 DISCUSSION AND FUTURE WORK

Our findings surface key design tensions and overlooked practices in how authors interpret and respond to peer reviews. Rather than offering a comprehensive theory of authoring, we identify actionable opportunities where current workflows and available tools are misaligned. In particular, we point to two design directions: supporting routine coordination and tracking tasks that current editors neglect, and addressing the collaborative and pedagogical dimensions of review handling. Below, we outline these implications and suggest directions for future work.

Delegating Routine Tasks, Preserving Human Interpretation. Participants were generally comfortable with delegating more "routine" tasks to the tool. For instance, highlighting and labeling text and ideas were considered acceptable, as was reordering or reorganizing concepts or ideas. Similarly, relying on the tool to track and connect changes across documents—this was a routine,

mundane task that was acceptable to delegate to the tool. The results of each of these micro-tasks are reviewable, and therefore easily corrected.

Much like in supporting the initial writing process [31, 61], this suggests a design opportunity for tools that perform routine micro-tasks, such as labeling passages of text based on criteria (e.g. allowing authors to attach "interpretation notes"), re-ordering or re-organizing concepts or passages in the manuscript, or tracking changes across multiple documents. In principle, these tools might take the place of conventional Find-and-Replace or Find-and-Highlight tools, though with slightly more flexible, higher-level inputs (e.g., "*Identify every time we make claim X*" or "*Re-order how the ideas are presented so there is parallel structure across these sections*"). These are tasks that are achievable with today's LLMs, and are beyond the capacity of existing word processing tools. In addition, our work also affirms the importance of the development of tools that are reviewable and verifiable [24, 32].

On the other hand, similarly to prior work on professional writer's use of LLMs [37, 59], our participants were resistant about involving LLMs in the all the tasks of responding to reviews. Some of our design concepts considered the possibility of tools having a deeper understanding of the review text: summarizing and interpreting review comments, grouping review issues together, and even proposing potential fixes. However, participants were uneasy with these ideas: they were not confident a tool would catch every issue in the reviews, or that it might misinterpret a review, or perhaps even propose poor solutions to the issues in the reviews. The challenge seems to be the distinction between surface-level meaning and subtextual-level meaning of review commentary. Our participants generally were not confident that the tools would be able to discern (correctly) the subtextual meaning implied by reviewers' remarks on their own. This is perhaps unsurprising: at the best of times, participants described trying to make "educated guesses" about what reviewers were hinting at, based on their understanding of the review process, and their own deeper understanding/knowledge of the manuscript. This mirrors the concerns of writers in Gero et al.'s work [20] who felt that were deeper notions of creativity that LLMs were not capable of. Most circumspect were our participants who had a working knowledge of how generative models operated—these participants pointed to towards notions of hallucinations in language models [26, 29] and felt firmly that language models did not fundamentally "understand" the text they were generating, let alone the text they were apparently interpreting. On the other hand, perhaps here there is an opportunity for authors to provide annotations or labels to tools the explain their categorization/rationale (much as they do in intermediary documents). A tool could perhaps make use of these to coordinate resolution finding for authors, as well as be there to help track tasks on a per issue basis.

Review Text as Imperfect Dialogue. While peer-review is nominally a dialogue between reviewers and authors, many researchers have previously commented on the paucity of the communication medium (i.e., a manuscript, a review letter, a response letter). Reviewers can only communicate through their review text. The review text is drafted and written with an idea, but even this is an imperfect reflection of the reviewer's ideas and intentions. It may be worded stronger or weaker than it should be interpreted, or worded awkwardly, or in a way that may be misinterpreted. It may be built on an incorrect assumption or misreading of the original manuscript itself. Or, one issue being identified may just be a single instance of a whole class of problems in a manuscript; thus, relying on a naïve, surface-level reading of the review text itself to create a task list to improve a manuscript may be misleading. Instead, our participants were clear that the role of the experienced co-authors was to provide an interpretation of the reviews, as well as to decide on the action plan. As we saw earlier, there are instances where the review text suggests a certain type of change, but that a different type of modification may ameliorate a whole class of problems. This hard-won awareness and understanding of a discipline's or a publication venue's norms and its application to manuscript review handling was difficult to articulate in a way that would be easily "codeable."

More importantly though, this means that we should view all text involved in this process (manuscript, review text, response text) as provisional: a best attempt at articulating an idea, but only that—a best attempt. This core idea underpinned many of our respondents' ideas about how AI tools that support the review process ought to be considered: rather than considering any of the text (e.g., the reviews) as rote instructions, they ought to be considered as guides and ideas—some of which might be wrong.

Collaboration and Mentorship as Intertwined Practices. Academic manuscripts are not only vehicles for reporting research findings, they are also key sites for collaboration and mentorship. Much like prior work [8, 50], many of the senior authors in our study highlighted the importance of their role as mentors for their junior collaborators. For these authors, the process of handling peer reviews was not simply about preparing revisions or composing a response letter. Rather, it was an opportunity to help junior collaborators develop essential academic skills: interpreting critique, forming a response strategy, coordinating with co-authors, and managing emotional reactions to criticism. As such, the task of review handling inherently blends the logistical demands of coordination with the pedagogical goals of mentorship.

This dual function has several implications for the design of tools in this space. First, it challenges designers to consider review handling as more than just an exercise in document management (e.g., editing reviews, crafting response letters, or updating manuscripts). As foreshadowed by prior work on the collaborative nature of academic authoring [69], and reinforced by our findings, the review phase involves a complex mesh of conversations, decisions, and tracking tasks, and only some of these are captured in formal documents.

With recent advances in LLMs and natural language processing, we see an opportunity to move beyond static documents. Much like work explore systems for supporting literature reviews [16, 45], our work highlights the potential for tools to assist in the maintenance of information across documents. For instance, tools might analyze emails or collaborative conversations to detect important decisions or unresolved questions, then annotate the corresponding locations in the reviews, manuscripts, or response drafts. Alternatively, systems might help co-authors maintain shared context by tracking action items, changes, or rationale for revisions over time—support that is currently missing in most academic authoring platforms. At present, authors must manually create and maintain these systems using ad hoc combinations of spreadsheets, comments, and memory.

While collaborative writing tools have been explored in prior work [4, 58, 64], similarly to Wang et al.'s work [63], this dual function highlights the importance of the different roles that members of a research team play in the writing process. Tools that aim to support this space must not only attend to the surrounding communication and coordination practices, but consider the variable roles that research members play in these practices.

Finally, mentorship-specific considerations further shape what authors find acceptable in tools. Similarly to concerns in lower-levels of education [23, 38], several senior participants expressed concern that if a tool were to pre-interpret reviewer remarks or automate too much of the response planning, it could deprive junior authors of critical learning opportunities. For them, interpreting review comments was not only a necessary task but also a developmental one. As such, tools in this space should augment, and not replace, early-career researchers' engagement with review materials, and provide opportunities for learning through reflection, not just automation.

These themes were not only described in participants' reflections but were reinforced during their responses to our design probes. Participants expressed clear preferences for tools that support surface-level coordination and contextually embedded interactions (as in Sketch 4), while showing resistance to systems that abstract away too much interpretive responsibility.

Limitations and Future Directions. Our sample engaged in HCI-specific norms (e.g., brief rebuttals, tone calibration). We acknowledge that other fields (e.g., qualitative social sciences, engineering) may have different interpretive cultures. Future work should explore how our findings translate across disciplinary contexts.

Nevertheless, our findings bring attention to tool support for academic authors. We highlight that review response is not just a document editing task; instead, it is a socially embedded, collaborative, and pedagogical process. Authors navigate not only revision logistics but also team coordination, emotional labor, and mentorship. Current tools offer little support for this broader context.

We surface a key distinction between surface-level and subtextual meaning in reviews, and show why authors resist automation in interpretive tasks. Yet we also identify where support is both welcome and feasible: tracking tasks, organizing reviewer input, and coordinating across documents and co-authors. Tools that foreground these practices, rather than aiming to replace author judgment by interpreting the reviews, better align with how review handling actually unfolds.

Finally, our study offers design directions for leveraging LLMs to support this phase in ways that authors find acceptable—not by interpreting reviews, but by helping authors synthesize, annotate, and coordinate responses across the documents and people involved. We see this not as the end of the authoring pipeline, but as a key stage in scholarly practice deserving of tailored, thoughtful support. Future work can build on this foundation through deployments, cross-disciplinary comparisons, and deeper engagement with the lived experience of co-authoring.

ACKNOWLEDGMENTS

Thank you to our participants for assisting with the study, and to prior reviewers for their ideas on how to improve the clarity of our manuscript. We thank NAVER AI, MITACS Globalink Program, NSERC Canada Graduate Scholarships and Foreign Study Supplement for partially supporting this research. This work was also partially funded by the School of Computing and Information Systems at the Singapore Management University.

REFERENCES

[1] Paulo Bala, Stuart James, Alessio Del Bue, and Valentina Nisi. 2022. Writing with (Digital) Scissors: Designing a Text Editing Tool for Assisted Storytelling Using Crowd-Generated Content. In *Interactive Storytelling*, Mirjam Vosmeer and Lissa Holloway-Attaway (Eds.). Springer International Publishing, Cham, 139–158.

[2] Karim Benharrak, Tim Zindulka, Florian Lehmann, Hendrik Heuer, and Daniel Buschek. 2024. Writer-Defined AI Personas for On-Demand Feedback Generation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1049, 18 pages. https://doi.org/10.1145/3613904.3642406

[3] Oloff C. Biermann, Ning F. Ma, and Dongwook Yoon. 2022. From Tool to Companion: Storywriters Want AI Writers to Respect Their Personal Values and Writing Strategies. In *Designing Interactive Systems Conference*. ACM, Virtual Event Australia, 1209–1227. https://doi.org/10.1145/3532106.3533506

[4] Robert P. Biuk-Aghai, Christopher Kelen, and Hari Venkatesan. 2008. Visualization of interactions in collaborative writing. In *2008 2nd IEEE International Conference on Digital Ecosystems and Technologies*. 97–102. https://doi.org/10.1109/DEST.2008.4635141

[5] Kyle Booten and Katy Ilonka Gero. 2021. Poetry Machines: Eliciting Designs for Interactive Writing Tools from Poets. In *Proceedings of the 13th Conference on Creativity and Cognition* (Virtual Event, Italy) *(CC '21)*. Association for Computing Machinery, New York, NY, USA, Article 51, 5 pages. https://doi.org/10.1145/3450741.3466813

[6] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health* 11, 4 (Aug. 2019), 589–597. https://doi.org/10.1080/2159676X.2019.1628806

[7] John C. Burnham. 1990. The Evolution of Editorial Peer Review. *JAMA: The Journal of the American Medical Association* 263, 10 (March 1990), 1323. https://doi.org/10.1001/jama.1990.03440100023003

[8] Susan T. Charles, Melissa M. Karnaze, and Frances M. Leslie. 2022. Positive factors related to graduate student mental health. *Journal of American College Health* 70, 6 (2022), 1858–1866. https://doi.org/10.1080/07448481.2020.1841207 arXiv:https://doi.org/10.1080/07448481.2020.1841207 PMID: 33522446.

[9] Jonathan H. Choi, Kristin E. Hickman, Amy Monahan, and Daniel B. Schwarcz. 2023. ChatGPT Goes to Law School. *SSRN Electronic Journal* (2023). https://doi.org/10.2139/ssrn.4335905

[10] 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 *Proceedings of the 23rd International Conference on Intelligent User Interfaces* (Tokyo, Japan) *(IUI '18)*. Association for Computing Machinery, New York, NY, USA, 329–340. https://doi.org/10.1145/3172944.3172983

[11] Hai Dang, Karim Benharrak, Florian Lehmann, and Daniel Buschek. 2022. Beyond Text Generation: Supporting Writers with Continuous Automatic Text Summaries. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 98, 13 pages. https://doi.org/10.1145/3526113.3545672

[12] Hai Dang, Sven Goller, Florian Lehmann, and Daniel Buschek. 2023. Choice Over Control: How Users Write with Large Language Models using Diegetic and Non-Diegetic Prompting. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–17. https://doi.org/10.1145/3544548.3580969

[13] Paramveer S. Dhillon, Somayeh Molaei, Jiaqi Li, Maximilian Golub, Shaochun Zheng, and Lionel Peter Robert. 2024. Shaping Human-AI Collaboration: Varied Scaffolding Levels in Co-writing with Language Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1044, 18 pages. https://doi.org/10.1145/3613904.3642134

[14] Giulia Di Fede, Davide Rocchesso, Steven P. Dow, and Salvatore Andolina. 2022. The Idea Machine: LLM-based Expansion, Rewriting, Combination, and Suggestion of Ideas. In *Proceedings of the 14th Conference on Creativity and Cognition* (Venice, Italy) *(CC '22)*. Association for Computing Machinery, New York, NY, USA, 623–627. https://doi.org/10.1145/3527927.3535197

[15] Douglas C Engelbart and William K English. 1968. A research center for augmenting human intellect. In *Proceedings of the December 9-11, 1968, fall joint computer*

conference, part I. 395–410.

[16] Raymond Fok, Joseph Chee Chang, Tal August, Amy X. Zhang, and Daniel S. Weld. 2024. Qlarify: Recursively Expandable Abstracts for Dynamic Information Retrieval over Scientific Papers. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 145, 21 pages. https://doi.org/10.1145/3654777.3676397

[17] Katy Ilonka Gero and Lydia B. Chilton. 2019. How a Stylistic, Machine-Generated Thesaurus Impacts a Writer's Process. In *Proceedings of the 2019 Conference on Creativity and Cognition* (San Diego, CA, USA) *(CC '19)*. Association for Computing Machinery, New York, NY, USA, 597–603. https://doi.org/10.1145/3325480.3326573

[18] 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* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300526

[19] Katy Ilonka Gero, Vivian Liu, and Lydia Chilton. 2022. Sparks: Inspiration for Science Writing using Language Models. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference* (Virtual Event, Australia) *(DIS '22)*. Association for Computing Machinery, New York, NY, USA, 1002–1019. https://doi.org/10.1145/3532106.3533533

[20] Katy Ilonka Gero, Tao Long, and Lydia B Chilton. 2023. Social Dynamics of AI Support in Creative Writing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–15. https://doi.org/10.1145/3544548.3580782

[21] Tirthankar Ghosal, Sandeep Kumar, Prabhat Kumar Bharti, and Asif Ekbal. 2022. Peer review analyze: A novel benchmark resource for computational analysis of peer reviews. *PLOS ONE* 17, 1 (Jan. 2022), e0259238. https://doi.org/10.1371/journal.pone.0259238 Publisher: Public Library of Science.

[22] Jonathan Grudin. 2013. Varieties of Conference Experience. *The Information Society* 29, 2 (March 2013), 71–77. https://doi.org/10.1080/01972243.2012.757263 Publisher: Routledge _eprint: https://doi.org/10.1080/01972243.2012.757263.

[23] Emma Harvey, Allison Koenecke, and Rene F. Kizilcec. 2025. "Don't Forget the Teachers": Towards an Educator-Centered Understanding of Harms from Large Language Models in Education. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 1064, 19 pages. https://doi.org/10.1145/3706598.3713210

[24] Md Naimul Hoque, Md Ehtesham-Ul-Haque, Niklas Elmqvist, and Syed Masum Billah. 2023. Accessible Data Representation with Natural Sound. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–19. https://doi.org/10.1145/3544548.3581087

[25] Sierk A. Horn. 2016. The Social and Psychological Costs of Peer Review: Stress and Coping With Manuscript Rejection. *Journal of Management Inquiry* 25, 1 (Jan. 2016), 11–26. https://doi.org/10.1177/1056492615586597

[26] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.* 43, 2, Article 42 (Jan. 2025), 55 pages. https://doi.org/10.1145/3703155

[27] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-Writing with Opinionated Language Models Affects Users' Views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 111, 15 pages. https://doi.org/10.1145/3544548.3581196

[28] Yvonne Jansen, Kasper Hornbæk, and Pierre Dragicevic. 2016. What Did Authors Value in the CHI'16 Reviews They Received?. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '16)*. Association for Computing Machinery, New York, NY, USA, 596–608. https://doi.org/10.1145/2851581.2892576

[29] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* 55, 12, Article 248 (March 2023), 38 pages. https://doi.org/10.1145/3571730

[30] Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2023. GPT-4 Passes the Bar Exam. *SSRN Electronic Journal* (2023). https://doi.org/10.2139/ssrn.4389233

[31] Jeongyeon Kim, Sangho Suh, Lydia B Chilton, and Haijun Xia. 2023. Metaphorian: Leveraging Large Language Models to Support Extended Metaphor Creation for Science Writing. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (Pittsburgh, PA, USA) *(DIS '23)*. Association for Computing Machinery, New York, NY, USA, 115–135. https://doi.org/10.1145/3563657.3595996

[32] Philippe Laban, Jesse Vig, Marti Hearst, Caiming Xiong, and Chien-Sheng Wu. 2024. Beyond the Chat: Executable and Verifiable Text-Editing with LLMs. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 20, 23 pages. https://doi.org/10.1145/3654777.3676419

[33] Mina Lee, Percy Liang, and Qian Yang. 2022. CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 388, 19 pages. https://doi.org/10.1145/3491102.3502030

[34] Ryan Liu and Nihar B. Shah. 2023. ReviewerGPT? An Exploratory Study on Using Large Language Models for Paper Reviewing. arXiv:2306.00622 [cs.CL]

[35] Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey. *Comput. Surveys* (June 2023). https://doi.org/10.1145/3605943 Just Accepted.

[36] Piotr Mirowski, Kory W. Mathewson, Jaylen Pittman, and Richard Evans. 2023. Co-Writing Screenplays and Theatre Scripts with Language Models: Evaluation by Industry Professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–34. https://doi.org/10.1145/3544548.3581225

[37] Alberto Monge Roffarello, Tommaso Calò, Luca Scibetta, and Luigi De Russis. 2025. Investigating How Computer Science Researchers Design Their Co-Writing Experiences With AI. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 1215, 17 pages. https://doi.org/10.1145/3706598.3713205

[38] Fatemeh Mosaiyebzadeh, Seyedamin Pouriyeh, Reza Parizi, Nasrin Dehbozorgi, Mohsen Dorodchi, and Daniel Macêdo Batista. 2023. Exploring the Role of ChatGPT in Education: Applications and Challenges. In *Proceedings of the 24th Annual Conference on Information Technology Education* (Marietta, GA, USA) *(SIGITE '23)*. Association for Computing Machinery, New York, NY, USA, 84–89. https://doi.org/10.1145/3585059.3611445

[39] Md Naimul Hoque, Tasfia Mashiat, Bhavya Ghai, Cecilia Shelton, Fanny Chevalier, Kari Kraus, and Niklas Elmqvist. 2023. The HaLLMark Effect: Supporting Provenance and Transparent Use of Large Language Models in Writing through Interactive Visualization. *arXiv e-prints* (2023), arXiv–2311.

[40] Philip Mark Newton. 2023. *ChatGPT performance on MCQ-based exams.* preprint. EdArXiv. https://doi.org/10.35542/osf.io/sytu3

[41] Eric Nichols, Leo Gao, and Randy Gomez. 2020. Collaborative Storytelling with Large-scale Neural Language Models. In *Proceedings of the 13th ACM SIGGRAPH Conference on Motion, Interaction and Games* (Virtual Event, SC, USA) *(MIG '20)*. Association for Computing Machinery, New York, NY, USA, Article 17, 10 pages. https://doi.org/10.1145/3424636.3426903

[42] Syavash Nobarany and Kellogg S. Booth. 2015. Use of politeness strategies in signed open peer review: Use of Politeness Strategies in Signed Open Peer Review. *Journal of the Association for Information Science and Technology* 66, 5 (May 2015), 1048–1064. https://doi.org/10.1002/asi.23229

[43] Syavash Nobarany and Kellogg S. Booth. 2017. Understanding and supporting anonymity policies in peer review. *Journal of the Association for Information Science and Technology* 68, 4 (April 2017), 957–971. https://doi.org/10.1002/asi.23711

[44] Syavash Nobarany, Kellogg S. Booth, and Gary Hsieh. 2016. What motivates people to review articles? The case of the human-computer interaction community. *Journal of the Association for Information Science and Technology* 67, 6 (June 2016), 1358–1371. https://doi.org/10.1002/asi.23469

[45] Srishti Palani, Aakanksha Naik, Doug Downey, Amy X. Zhang, Jonathan Bragg, and Joseph Chee Chang. 2023. Relatedly: Scaffolding Literature Reviews with Existing Related Work Sections. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 742, 20 pages. https://doi.org/10.1145/3544548.3580841

[46] Savvas Petridis, Nicholas Diakopoulos, Kevin Crowston, Mark Hansen, Keren Henderson, Stan Jastrzebski, Jeffrey V Nickerson, and Lydia B Chilton. 2023. AngleKindling: Supporting Journalistic Angle Ideation with Large Language Models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 225, 16 pages. https://doi.org/10.1145/3544548.3580907

[47] Hua Xuan Qin, Shan Jin, Ze Gao, Mingming Fan, and Pan Hui. 2024. CharacterMeet: Supporting Creative Writers' Entire Story Character Construction Processes Through Conversation with LLM-Powered Chatbot Avatars. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1051, 19 pages. https://doi.org/10.1145/3613904.3642105

[48] Mohi Reza, Jeb Thomas-Mitchell, Peter Dushniku, Nathan Laundry, Joseph Jay Williams, and Anastasia Kuzminykh. 2025. Co-Writing with AI, on Human Terms: Aligning Research with User Demands Across the Writing Process. *Proc. ACM Hum.-Comput. Interact.* 9, 7, Article CSCW385 (Oct. 2025), 37 pages. https://doi.org/10.1145/3757566

[49] Shalaleh Rismani, Su Lin Blodgett, Alexandra Olteanu, Q. Vera Liao, and AJung Moon. 2024. How different mental models of AI-based writing assistants impact writers' interactions with them. In *Proceedings of the Third Workshop on*

Intelligent and Interactive Writing Assistants (Honolulu, HI, USA) *(In2Writing '24)*. Association for Computing Machinery, New York, NY, USA, 34–37. https://doi.org/10.1145/3690712.3690722

[50] S. Rönkkönen, L. Tikkanen, V. Virtanen, and K. Pyhältö. 2024. The impact of supervisor and research community support on PhD candidates' research engagement. *European Journal of Higher Education* 14, 4 (2024), 536–553. https://doi.org/10.1080/21568235.2023.2229565 arXiv:https://doi.org/10.1080/21568235.2023.2229565

[51] Shubhra Kanti Karmaker Santu, Sanjeev Kumar Sinha, Naman Bansal, Alex Knipper, Souvika Sarkar, John Salvador, Yash Mahajan, Sri Guttikonda, Mousumi Akter, Matthew Freestone, and Matthew C. Williams Jr au2. 2024. Prompting LLMs to Compose Meta-Review Drafts from Peer-Review Narratives of Scholarly Manuscripts. arXiv:2402.15589 [cs.CL]

[52] Anna Severin and Joanna Chataway. 2021. Purposes of peer review: A qualitative study of stakeholder expectations and perceptions. *Learned Publishing* 34, 2 (2021), 144–155.

[53] Xiaotian Su, Thiemo Wambsganss, Roman Rietsche, Seyed Parsa Neshaei, and Tanja Käser. 2023. Reviewriter: AI-Generated Instructions For Peer Review Writing. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, Ekaterina Kochmar, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Nitin Madnani, Anaïs Tack, Victoria Yaneva, Zheng Yuan, and Torsten Zesch (Eds.). Association for Computational Linguistics, Toronto, Canada, 57–71. https://doi.org/10.18653/v1/2023.bea-1.5

[54] Lu Sun, Aaron Chan, Yun Seo Chang, and Steven P. Dow. 2024. ReviewFlow: Intelligent Scaffolding to Support Academic Peer Reviewing. In *Proceedings of the 29th International Conference on Intelligent User Interfaces* (Greenville, SC, USA) *(IUI '24)*. Association for Computing Machinery, New York, NY, USA, 120–137. https://doi.org/10.1145/3640543.3645159

[55] Lu Sun, Stone Tao, Junjie Hu, and Steven P. Dow. 2024. MetaWriter: Exploring the Potential and Perils of AI Writing Support in Scientific Peer Review. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 94 (April 2024), 32 pages. https://doi.org/10.1145/3637371

[56] Mengyi Sun, Jainabou Barry Danfa, and Misha Teplitskiy. 2022. Does double-blind peer review reduce bias? Evidence from a top computer science conference. *Journal of the Association for Information Science and Technology* 73, 6 (2022), 811–819. https://doi.org/10.1002/asi.24582 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.24582.

[57] Yuying Tang, Haotian Li, Minghe Lan, Xiaojuan Ma, and Huamin Qu. 2025. Understanding Screenwriters' Practices, Attitudes, and Future Expectations in Human-AI Co-Creation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 26, 18 pages. https://doi.org/10.1145/3706598.3714120

[58] Sunny Tian, Amy X. Zhang, and David Karger. 2021. A System for Interleaving Discussion and Summarization in Online Collaboration. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW3, Article 241 (Jan. 2021), 27 pages. https://doi.org/10.1145/3432940

[59] Rama Adithya Varanasi, Batia Mishan Wiesenfeld, and Oded Nov. 2025. AI Rivalry as a Craft: How Resisting and Embracing Generative AI Are Reshaping the Writing Profession. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 1198, 19 pages. https://doi.org/10.1145/3706598.3714035

[60] Qian Wan, Siying Hu, Yu Zhang, Piaohong Wang, Bo Wen, and Zhicong Lu. 2024. "It Felt Like Having a Second Mind": Investigating Human-AI Co-creativity in Prewriting with Large Language Models. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 84 (April 2024), 26 pages. https://doi.org/10.1145/3637361

[61] Qian Wan, Jiannan Li, Huanchen Wang, and Zhicong Lu. 2025. Polymind: Parallel Visual Diagramming with Large Language Models to Support Prewriting Through Microtasks. *Proc. ACM Hum.-Comput. Interact.* 9, 7, Article CSCW316 (Oct. 2025), 29 pages. https://doi.org/10.1145/3757497

[62] Ruyuan Wan, Simret Araya Gebreegziabher, Toby Jia-Jun Li, and Karla Badillo-Urquiola. 2024. CoCo Matrix: Taxonomy of Cognitive Contributions in Co-writing with Intelligent Agents. In *Proceedings of the 16th Conference on Creativity & Cognition* (Chicago, IL, USA) *(CC '24)*. Association for Computing Machinery, New York, NY, USA, 504–511. https://doi.org/10.1145/3635636.3664260

[63] Dakuo Wang, Michael Muller, Qian Yang, Zijun Wang, Ming Tan, and Stacy Hobson. 2022. Organizational Distance Also Matters: How Organizational Distance Among Industrial Research Teams Affect Their Research Productivity. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 453 (Nov. 2022), 18 pages. https://doi.org/10.1145/3555554

[64] Dakuo Wang, Judith S. Olson, Jingwen Zhang, Trung Nguyen, and Gary M. Olson. 2015. DocuViz: Visualizing Collaborative Writing. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 1865–1874. https://doi.org/10.1145/2702123.2702517

[65] Christoph Johannes Weber, Sebastian Burgkart, and Sylvia Rothe. 2024. wr-AI-ter: Enhancing Ownership Perception in AI-Driven Script Writing. In *Proceedings of the 2024 ACM International Conference on Interactive Media Experiences* (Stockholm, Sweden) *(IMX '24)*. Association for Computing Machinery, New York, NY, USA, 145–156. https://doi.org/10.1145/3639701.3656325

[66] Chi-Lan Yang, Alarith Uhde, Naomi Yamashita, and Hideaki Kuzuoka. 2025. Understanding and Supporting Peer Review Using AI-reframed Positive Summary. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 171, 16 pages. https://doi.org/10.1145/3706598.3713219

[67] Chao Zhang, Kexin Ju, Peter Bidoshi, Yu-Chun Grace Yen, and Jeffrey M. Rzeszotarski. 2025. Friction: Deciphering Writing Feedback into Writing Revisions through LLM-Assisted Reflection. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 935, 27 pages. https://doi.org/10.1145/3706598.3714316

[68] Chao Zhang, Kexin Ju, Zhuolun Han, Yu-Chun Grace Yen, and Jeffrey M. Rzeszotarski. 2025. Synthia: Visually Interpreting and Synthesizing Feedback for Writing Revision. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25)*. Association for Computing Machinery, New York, NY, USA, Article 88, 16 pages. https://doi.org/10.1145/3746059.3747703

[69] Qixing Zheng, Kellogg Booth, and Joanna McGrenere. 2006. Co-authoring with structured annotations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Montréal Québec Canada, 131–140. https://doi.org/10.1145/1124772.1124794

A DESCRIPTION OF VIDEO SKETCHES

The video sketches that we used in the study are included as a video figure. We include here a lengthier description of each sketch as a short comic strip. The images are clipped from each video that was shown to participants. As such, there are extraneous artefacts (e.g. mouse cursors and annotations) that were intended for participants to understand the flow of interaction.

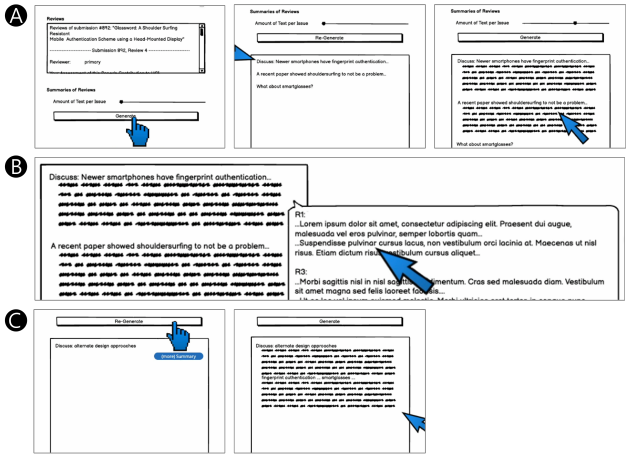

Figure 4: Sketch 1: Summaries. This sketch focused on the idea of varying length summaries generated by the system. In (A): After the user pastes their review into the system, they select a short amount of summary text to be generated (using the slider), and press the Generate button. This gives them an overview of the major comments. If they move the slider to generate more text, they can regenerate the summaries, which gives lengthier summaries. (B) shows that hovering over the summary text provides a call-out that explains which reviewer made a comment pertinent to that summary. Finally, (C) shows how the user can ask for a particular topic (by typing into the window, and then pressing generate), where the system then checks the reviews to see if there were comments pertinent to that topic.

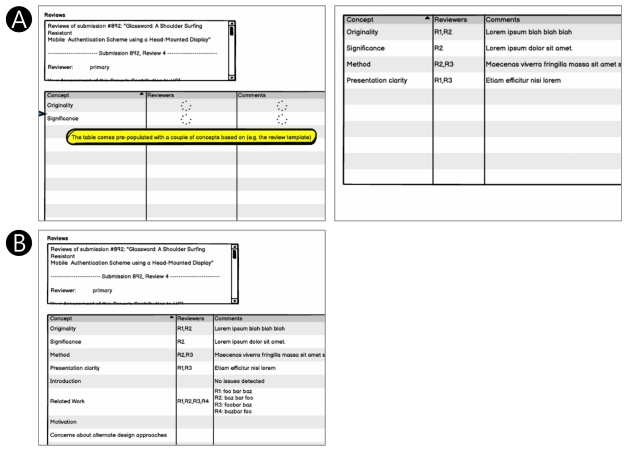

Figure 5: Sketch 2: Tabular. This sketch provides commentary in a tabular format. (A) shows that the initial concepts that are being explored are based on the the review template, which asks reviewers to comment on Originality and Significance. The user can, however, add additional concepts to consider: here, they have filled in Method and Presentation Clarity. In response, the system studies the reviews, and provides the reviewer IDs that commented on these concepts, and then provides a commentary based on the reviewer's comments. (B) shows that the user can also provide "spatial" concepts (e.g. section headers from their manuscript) to see if reviewers have commented on those sections specifically.

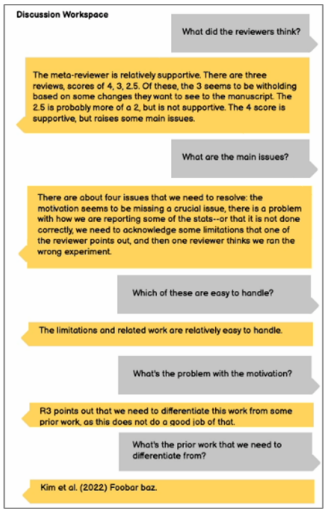

Figure 6: Sketch 3: Chat Partner. This sketch explores a dialogue-based interaction with a chatbot that understands the reviews. The user can ask questions of the chatbot, and the chatbot provides commentary based on its understanding of the reviews. It can provide direct summaries as well as direct quotes (and information from the reviews, such as references).

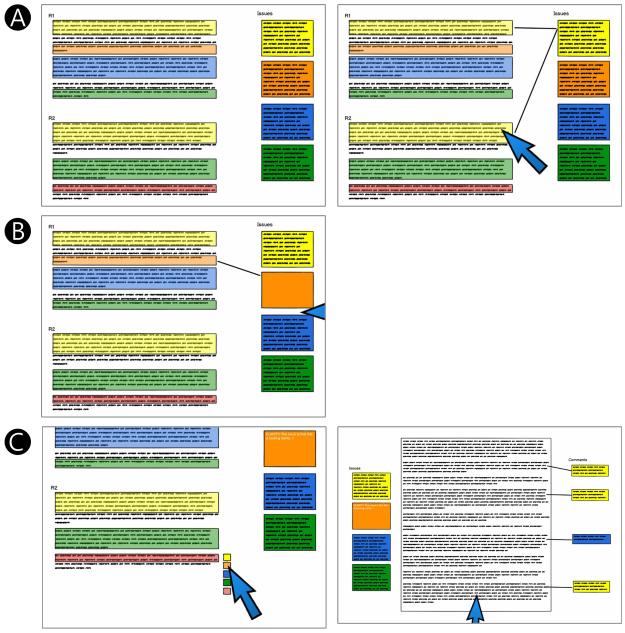

Figure 7: Sketch 4: Document-Centric. This sketch explores a document-centric interaction. In (A), the system begins by highlighting the original reviews. The colour coding represents similar issues across reviews from different reviews. On the left, the issues are summarised. Clicking on an issue provides links that show the user where the issue appears in different reviews. In (B), the user is able to re-label or re-characterize the nature of an issue (e.g. if they disagree with the system). In (C), the user can then re-classify a labeling in the original reviews, changing how each highlighting is classified in the system. Finally, these classifications and ideas are carried through to a view of the manuscript. Here, the manuscript is annotated with parts of the original manuscript that ought to be revised based on the review comments—again, these are colour coded.