Synthetic Diversity: How Researchers Perceive and Engage with LLM-Generated Diverse Research Feedback

ANONYMOUS AUTHOR(S)*

Obtaining diverse expert feedback on academic research is valuable yet challenging. Large Language Models (LLMs) show promise in simulating varied perspectives and generating paper reviews, but perceptions of synthetic diverse research feedback remain under-explored. This study investigates how researchers perceive LLM-generated reviews compared to human reviews. We generated synthetic diverse reviews for participants' papers and conducted a mixed-methods study with 18 experienced researchers. Participants recognized synthetic diversity in the reviews' expertise and attitudinal stances, along with benefits in uncovering blind spots, identifying critical issues, and enhancing willingness to improve. We found differences between perceptions of LLM-generated versus human reviews in degree of diversity, homogeneity, authenticity of expertise, and divergent opinions. Our findings offer insights into LLM's role in academic discourse and inform guidelines on generating meaningful LLM-augmented diverse feedback. We also contribute a dataset of over 800 sentence-level annotations from 54 synthetic and 62 human reviews to facilitate future research.

ACM Reference Format:

Anonymous Author(s). 2018. Synthetic Diversity: How Researchers Perceive and Engage with LLM-Generated Diverse Research Feedback. In . ACM, New York, NY, USA, 24 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Exposure to diverse perspectives and opinions has an important role in many contexts such as creative performance at work [19, 48, 52], online participatory culture [21, 32], and journalism [5]. Specifically, in organizational work teams, diversity can provide a greater variety of knowledge, perspectives and approaches that facilitate creativity and problem-solving [49]. When making decisions, combining even a small number of diverse opinions, especially more independent ones, can improve judgment accuracy compared to relying on a single individual [73]. In scientific research, having diverse research communities in terms of researchers' lived experiences, expertise, backgrounds, and views is important for generating new research questions, identifying limitations in existing models, accessing more complete data, and revealing biases [14, 31].

In the scholarly research process, prior work has emphasized the importance of gathering feedback from multiple sources for different stages of research such as prototyping [20], iterative design [76], academic writing [72]. Diverse feedback enables "the critique recipient to blend the feedback, discounting the outlying overly positive or negative comments," yielding a clearer sense of the overall reception [68]. Varied perspectives in feedback can also reduce blind spots and cover more problems, surfacing issues that may otherwise go unnoticed [12]. In this work, we focus on opinion diversity that has been explored in the context of design feedback from multiple providers [74] and online discussions and deliberation on social issues [21, 33, 60].

While researchers may benefit from diverse feedback, obtaining this diverse expert feedback can be challenging. Crowdsourcing systems [22, 25, 74] and social media platforms [27] support getting feedback from diverse sources.

However, these channels may not be suitable for providing expert feedback on academic research writing due to the demanding nature of the task [62] and researchers' concerns about premature criticism or scooping. Traditionally, researchers have relied on various channels to seek out diverse expert feedback, such as mentors, advisors, peers [18, 50], formal review processes, and workshops. However, these methods often come with high costs in terms of turn-around time, effort, and social capital, and may not always fully convey varied perspectives due to social dynamics [34, 35]. The reliance on social capital further raises concerns about inequity in academia [40, 51]. Researchers from smaller institutions or with limited professional networks may struggle to access the same quality and diversity of feedback as their counterparts at more prestigious or well-connected institutions, perpetuating existing inequalities in academic and career advancement.

Recent advancements in large language models (LLMs) have shown promise in assisting the review process [56, 67] and generating reviews [39]. However, it remains unclear if LLMs can generate *useful* diverse reviews and how researcher perceive this type of artificially generated diversity in the form of feedback. "Expert" is central to two aspects of this context: (1) this can refers to the task that requires *LLMs* to comprehend an academic research work before making judgment and recommendations that is used to be exclusively performed by experts; (2) the *recipients* of this feedback are also experts in their own fields, specifically academic researchers who authored the proposal. Hence, the evaluation of LLM-generated output demands domain expertise, prior experience, cognitive engagement, and reflective thinking. Investigating how expert users perceive and engage with artificially generated diverse feedback is crucial to revealing the fundamentals of LLMs' ability in expert tasks, which can guide the design of an LLM-powered system to support expert feedback tasks. Our stance is that high-quality human feedback may always benefit researchers, and synthetic review should *not* replace human review without proper evaluation; however, given the LLMs' disruptive impact on research [7], we discuss implications for the research community to design useful and responsible ways for researchers to interact with diverse synthetic feedback.

To summarize, the primary objective of this project is to explore how LLMs can be leveraged to generate diverse feedback on written academic work. Our work is one of the first empirical studies to understand the perceived benefits and barriers of LLM-generated diverse synthetic research feedback. Our work contribute to the following:

- Empirical insights from a user study using LLM-generated synthetic diverse feedback based on participants' actual research work, uncovering how academic researchers perceive LLM-generated synthetic diversity in feedback and how it compares with natural variations in human reviews.
- A set of design implications for future work in generating meaningfully diverse feedback that and important areas to scaffold, including
- A novel dataset comprising 858 sentence-level annotations of synthetic and human reviews with researchers' rationales, providing a rich foundation for future analysis and model development aimed at enhancing LLMs' capabilities for generating diverse research feedback.

2 RELATED WORK

2.1 Benefits and Challenges of Diverse Feedback

Our work builds on a rich body of research exploring the benefits and challenges of diverse feedback in various contexts, particularly in academic writing, peer review, and design critique. The importance of diverse perspectives has been widely recognized across domains such as organizational work, creative performance, and scientific research [19, 48, 49, 52]. In academic settings, diverse feedback has been shown to enhance creativity, problem-solving, and decision-making accuracy [73]. Xu and Zhang [72] found that multiple sources of feedback (automated, peer, and teacher) complement each other in different areas of academic writing, while Cho and MacArthur [11] demonstrated that students receiving feedback from multiple peers made more complex revisions and showed greater improvement in writing quality. These findings align with Intemann's [31] argument that diverse scientific communities are more likely to identify limitations in existing models, propose alternative hypotheses, and reveal biases in research. Clark and Jagsi [14] emphasized the importance of diverse peer reviewers in scientific journals to ensure that published science reflects the diversity of the communities it serves, a principle we extend to earlier stages of the research process.

Despite the potential benefits, getting useful diverse feedback presents several challenges for researchers. In the context of official peer review, researchers [34, 35] highlighted that traditional methods of seeking feedback may not always fully convey varied perspectives due to social dynamics within academic circles. Moreover, De Saá-Pérez et al. [17] found non-linear relationships between certain diversity attributes and research team performance, suggesting that diversity can be beneficial up to a point before communication and coordination problems arise.

In the context of design critique and classroom writing, researchers have implemented crowd-sourcing systems to elicit more feedback. Yen et al. [74] created Decipher, an interactive visualization tool to help designers interpret feedback from multiple providers, highlighting how diversity in feedback can surface contradictions and varying focuses. In academic writing, Greenberg et al. [25] developed Critiki, a system that improves critique quality by providing scaffolds such as question prompts and common errors. Tinapple et al.'s CritViz system [68] demonstrated how diverse feedback enables recipients to blend feedback, discounting outliers and gaining a clearer sense of overall reception.

However, these crowdsourcing approaches, while promising, face limitations when applied to academic research papers, which often require deeper domain expertise and engagement than typical crowdsourced tasks can provide. Moreover, researchers may hesitate to expose early-stage work to broad audiences due to concerns about idea appropriation or premature criticism [62]. Cho and Schunn [10], along with Van Zundert et al. [70], suggested that feedback from a large number of novices can be as effective as feedback from a small number of experts in certain contexts. This finding opens up possibilities for gathering diverse perspectives from a broader range of sources. Yet, their applicability to highly specialized research domains remains uncertain. This tension between the benefits of diverse perspectives and the need for domain expertise in research feedback presents a significant challenge, highlighting the need for innovative solutions that can provide diverse, expert-level feedback.

2.2 Large Language Models in Academic Review and Feedback Generation

The use of AI tools in the scientific publication process has gained significant attention. Early algorithms were developed to summarize papers [16], detect statistical errors [53], correct citations [75], and identify fairness issues [79]. Recent advances in Large Language Models (LLMs) like ChatGPT and GPT-4 have intensified interest in their potential for scientific feedback. Hosseini et al. conducted a small-scale qualitative investigation to gauge ChatGPT's effectiveness in the peer review process [30]. Similarly, Robertson et al. involved 10 participants to assess GPT-4's benefits in aiding peer review [59]. Liu et al. [43] demonstrated GPT-4's capability to identify errors and compare paper quality in computer science literature, while Verharen et al. [71] used ChatGPT to uncover gender disparities in neuroscience peer reviews.

In the context of peer review assistance, systems like CreBot [56] and ReviewFlow [67] leverage LLMs to generate relevant feedback and support novice reviewers, potentially enhancing the diversity and quality of reviews. Liang et al. [39] investigated LLMs' capacity to generate full academic reviews, finding promise in structure and content but limitations in domain-specific knowledge and critical analysis. Beyond traditional peer review, Liu and Sun [42] demonstrated high agreement between GPT-4 and human coders in qualitative research tasks, suggesting broader applications in academic analysis. These studies collectively highlight both the potential and current limitations of LLMs in augmenting human expertise across various aspects of the academic feedback process.

At the same time, NLP researchers have been exploring the ability for LLMs to embody the diversity of human opinions [65, 69]. Hayati et al. [29] investigated the extent to which LLMs can extract and generate diverse perspectives and rationales on subjective topics. Their findings indicate that while LLMs can generate semantically diverse rationales, human-written opinions still exhibit greater diversity in some aspects. This suggests that while LLMs have the potential to contribute to diverse feedback, they may not yet fully capture the breadth of human perspective. Efforts have also been made to enhance the diversity and controllability of LLM-generated perspectives. Li et al. [38] proposed a novel approach to fine-tune LLMs on debate-augmented data, significantly improving their capability to express diverse perspectives in a controllable manner. Their work demonstrates that with appropriate training, LLMs can generate high-quality statements representing various stances on controversial topics, outperforming existing models in both response quality and controversy controllability. However, the impact of LLMs on content diversity is not uniformly positive. Padmakumar and He [54] found that writing with a feedback-tuned LLM (InstructGPT) reduced lexical and content diversity compared to writing with a base LLM or without model assistance. This highlights the need for careful consideration of how LLMs are integrated into the writing and feedback process to avoid inadvertently reducing diversity. The relationship between LLM usage and opinion diversity appears to be complex. Research has found observed that partial reliance on LLMs can promote opinion diversity, while over-reliance may limit it [63]. Their study suggests that an optimal balance in LLMs usage could potentially enhance the diversity of perspectives in academic discourse.

While significant progress has been made in understanding LLM-generated reviews [39] and opinion diversity [29] separately, a gap exists at their intersection: the perceived usefulness of LLM-generated diverse feedback in academic contexts remain unexplored. Moreover, current research [56, 67] lacks empirically-grounded design principles for effective LLM-generated feedback systems. These gaps necessitate our work which examines both the perception of LLM-generated diverse feedback by academic experts and proposes the design strategies for its effective usage.

2.3 Perceptions and Engagement with AI-Generated Content by Experts

Recent HCI studies have explored the potential of Large Language Models (LLMs) in supporting domain experts across various fields. This body of research reveals both promising benefits and notable challenges in leveraging LLM-generated content for expert tasks. LLMs can inspire new ideas and angles that experts might not have considered. Liu et al. observed that their CoQuest tool helped researchers explore multiple angles for a given topic, particularly useful for interdisciplinary research [44]. In a different context, Gu et al. found that LLMs can broaden the analysis decision space for data analysts [26]. In the context of journalism, researchers also found that LLMs can significantly speed up expert workflows and reduce the cognitive load of brainstorming angles by providing specific angles that easily inspired next steps [57]. However, engaging with LLM-generated content also posed multiple challenges. A crucial challenge is verifying AI-generated outputs. LLM-generated content can sometimes be too general to inspire specific next steps, as observed in both journalistic [57] and research contexts [44].

These works collectively emphasize the significant potential and complex implications of LLMs in supporting domain experts' work, highlighting the increasingly influential role that AI-powered systems play in assisting expert tasks and the potential for these tools to not only facilitate work but to fundamentally alter expert processes and outputs. As LLM-based tools for expert support become more prevalent, it is crucial to examine the ways in which they impact individuals' cognitive processes and work practices. We build on this foundation and contribute to this space by investigating the range of characteristics of diverse synthetic feedback and how it is perceived by experts.

3 FORMATIVE STUDY

To understand how to generate useful diverse synthetic feedback, we conducted a formative study with three experienced HCI researchers. Participants provided research work samples that had received human feedback. We conducted semi-structured interviews (60-90 minutes) to gather their reactions to both human and AI-generated feedback presented in a Google Doc. Our initial approach prompts language models to provide divergent decisions ranging from "Strong Accept" to "Strong Reject."

All researchers suggested that expertise-grounded perspectives were more beneficial than simply asking LLMs to generate varied feedback without a specific focus. P1 noted that while AI-generated feedback showed more individual variability than human feedback, their collective contribution was similar. P1 also observed, "There is value in receiving repeated feedback from different sources - it's an indication of a critical issue," reinforcing the importance of meaningful diversity. This prompted us to shift from simply prompting LLMs to generating diverse review points to incorporating profile-based diversity in our main study.

Concerns about the validity, reliability, and sufficiency of AI-generated feedback emerged. Issues with accuracy, specificity, and context-sensitivity were noted, with P2 and P3 highlighting the lack of in-depth theoretical and methodological advice. This revealed that participants valued quality over mere opinion variability in feedback. P2 emphasized, "Diversity has to be built on the basis of relevance and actionability." This led us to ensure the review quality in our approach to generating diverse feedback. Despite reservations, participants expressed curiosity about further engaging with the AI tool, suggesting that perceived usefulness would hinge on continued interaction. This indicated that limitations might stem from expectations based on traditional feedback, prompting us to consider ways to encourage more dynamic engagement in our main study.

Our initial method of using Google Docs for feedback collection revealed several limitations. Participants spent considerable time typing out reactions. Additionally, it was challenging for participants to quickly recall their annotations through the dense feedback when answering post-study interview questions. To address these issues, we designed a customized UI for the main study, allowing color coding and more structured annotation to reduce cognitive load, assist easier recall of their reactions, while maintaining flexibility.

Based on the findings, we adjusted our main study design: shifting to profile-based diversity in feedback generation, refining LLM prompts to improve relevance and actionability, designing protocols to encourage verbalization of thoughts and potential actions, implementing a customized UI for efficient annotation, and recruiting participants with diverse research experience levels.

4 PIPELINE AND IMPLEMENTATION

To generate diverse synthetic feedback for research proposals, we developed a pipeline leveraging LLMs. Our approach focused on quality and diversity in the generated feedback. We implemented two versions for each aspect of quality assurance and diversity promotion, resulting in four distinct combinations of feedback generation strategies. This approach was not aimed at comprehensiveness, but rather to cover a range of common techniques that would expose participants to varied characteristics of LLM-generated feedback. By doing so, we sought to elicit richer qualitative insights in our subsequent user study. Our goal was to ensure that the feedback examples would prompt diverse reactions and reflections from researchers, allowing us to gain a more nuanced understanding of how they perceive and interact with AI-generated research feedback across different generation approaches.

4.1 Generating Diversity

Building on our formative study (section 3), we conducted further prompting experiments to generate meaningful diversity in AI-generated feedback. We compared the performance of our approach against a baseline where we simply asked language models to generate diverse feedback, using our own research documents as test cases. The evaluation involved both qualitative examination of content and analysis of simple metrics. Eventually, we landed at two refined approaches:

4.1.1 LLM-defined Diverse Personas (D1). We prompted the LLM to generate three diverse reviewer profiles (Appendix A) qualified to review the work, focusing on diverse disciplinary backgrounds, research domain, methodological expertise, and related personal experience.

4.1.2 Viewpoint-based Diversity (D2). This approach involved a three-step process to generate diverse viewpoints (Appendix B):

1. Topic Extraction: We identified main topics in the proposal that lack academic consensus or rely on potentially biased assumptions.
2. Viewpoint Generation: For each topic, we generated a range of opinion statements reflecting different stances.
3. Constructive Viewpoint Profiles: We created profiles combining various viewpoints, such as optimistic vs. pessimistic views on technology feasibility, conservative vs. radical data interpretations, or practical vs. theoretical implications of findings.

4.2 Ensuring Quality

We implemented two methods to ensure the quality of the generated feedback:

One-shot Prompting (Q1). This method utilized a single comprehensive prompt that incorporated all the necessary components to generate a complete review based on a given opinion profile. We synthesized high-level general review guidelines from multiple sources (e.g. CHI unofficial review guidelines [1], review desiderata synthesized from prior work [77], and official review guidelines from conference websites) by iteratively adding unique points from different guidelines to ensure a balance of comprehensiveness and also conciseness. and instructed the model to generate reviews that adhere to best practices.

Iterative Refinement (Q2). We implemented an iterative refinement process, inspired by research on using multiple LLMs to critique and improve their own outputs [9, 37, 80]. This involved generating an initial review, critiquing it in a separate conversation, and then refining the original based on this feedback. We limited this to three iterations to prevent homogenization, which we observed would override the intentional diversity in our reviewer profiles. For example, the critiquing model would identify overlooked issues, potentially conflicting with our goal of having different profiles focus on distinct aspects. This approach balanced improving review quality with maintaining diverse perspectives, highlighting the tension between refinement and preserving unique viewpoints in our study design.

[1] "An Unofficial Guide to Reviewing for SIGCHI," Google Docs, URL: https://shorturl.at/ZzW1w accessed November 19, 2024.

4.3 Implementation Details

For end-to-end prompting (Q1), we used GPT-4 via Azure OpenAI Service. For iterative prompting (Q2), we used Claude 3.5 Sonnet via its chat interface. Both models are among the state-of-the-art best-performing LLMs available at the time of writing. For each participant, we randomly assigned them to one of the combinations of quality assurance (Q1 or Q2) and diversity promotion (D1 or D2) methods.

Our prompt design underwent multiple iterations to ensure reasonable quality while serving as probes to showcase LLM's capabilities in generating research feedback. The generated reviews, potentially showing some level of usefulness, primarily aim to elicit researchers' perceptions and interactions with AI-generated feedback. This approach allows us to investigate both the strengths and limitations of current LLMs in producing expert-level feedback. We also opted against an agent-based approach due to the lack of a clearly superior model at the time of study design. Instead, this study serves to understand researchers' interactions with LLM-generated feedback and inform future design considerations for more advanced systems, including potential agent-based models. As discussed, we prompted the LLM to generate diverse reviewer profiles and diverse viewpoints. Examples of the personas and viewpoints generated are available in Appendix A and Appendix B.

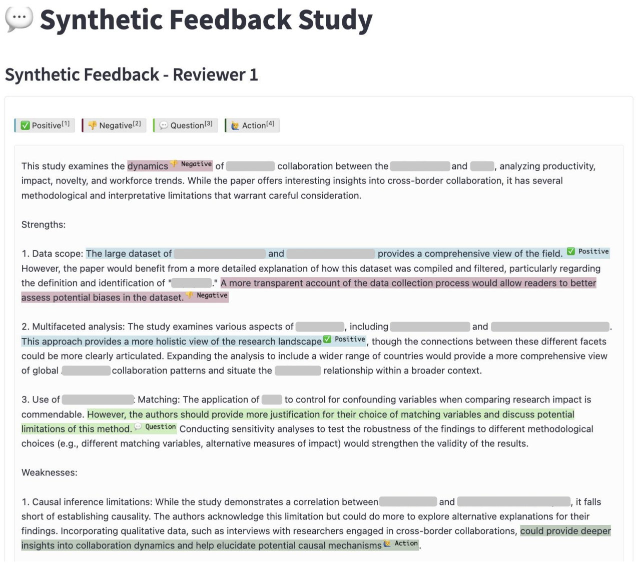
5 USER STUDY DESIGN

5.1 Participants

Participants were recruited through university Slack channels, social media platforms, and personal connections. The study was advertised as an opportunity to receive AI-generated reviews for participants' academic work. Based on the screening survey, we selected participants to ensure diverse backgrounds, considering review experience, research maturity, and work type (See Table 1). Participants received $30 compensation via Amazon gift card or Zelle. Informed consent was obtained, including permission to use research documents with AI models.

Table 1. Participant Information

PID	Current position	Experience	Research area	Current Status	Submission	Diversity	Method
p1	PhD Student	16-20	HCI	Accepted with minor revisions	AIES	Viewpoint	One-shot
p2	PhD Student	6-10	HCI	Rejected, planning resubmission	UIST	Viewpoint	One-shot
p3	PhD Student	11-15	HCI	Published and presented	ICCBR	Viewpoint	One-shot
p4	PhD Student	16-20	CS	Rejected after major revision	CSCW	Persona	One-shot
p5	PhD Student	11-15	HCI	Rejected, considering archiving	CSCW	Persona	One-shot
p6	PhD Student	1-5	HCI	Accepted after revisions	CSCW	Persona	One-shot
p7	PhD Student	6-10	HCI	Accepted and finalized	CHI PLAY	Persona	Iterative
p8	Research Scientist	20+	HCI	Inactive after rejection	NSF Grant	Persona	Iterative
p9	Industry Researcher	20+	HCI	Published	CSCW	Persona	Iterative
p10	Postdoctoral Researcher	6-10	HCI	Published after revisions	CHI	Viewpoints	Iterative
p11	PhD Student	20+	Info Sci	Under review at new journal	Nature	Persona	Iterative
p12	Associate Professor	20+	HCI	Accepted after revisions	mobileHCI	Persona	Iterative
p13	PhD Student	11-15	HCI	Archived after multiple rejections	UIST	Persona	Iterative
p14	Masters student	1-5	NLP	Under first review	EMNLP	Persona	Iterative
p15	PhD Student	6-10	CS	Rejected, under revision	CHI	Persona	Iterative
p16	PhD Student	11-15	HCI	Accepted after initial submission	CHI	Viewpoints	Iterative
p17	Assistant Professor	20+	Env Eng	Rejected, planning resubmission	NSF Grant	Viewpoints	Iterative
p18	Postdoctoral Researcher	20+	NLP	Rejected, revising for resubmission	Nature	Viewpoints	Iterative



💬 Synthetic Feedback Study

Synthetic Feedback - Reviewer 1

This study examines the dynamics 👎 Negative of [redacted] collaboration between the [redacted] and [redacted], analyzing productivity, impact, novelty, and workforce trends. While the paper offers interesting insights into cross-border collaboration, it has several methodological and interpretative limitations that warrant careful consideration.

Strengths:

1. Data scope: The large dataset of [redacted] and [redacted] provides a comprehensive view of the field. ✅ Positive However, the paper would benefit from a more detailed explanation of how this dataset was compiled and filtered, particularly regarding the definition and identification of "[redacted]." A more transparent account of the data collection process would allow readers to better assess potential biases in the dataset. 👎 Negative

2. Multifaceted analysis: The study examines various aspects of [redacted], including [redacted] and [redacted]. This approach provides a more holistic view of the research landscape ✅ Positive, though the connections between these different facets could be more clearly articulated. Expanding the analysis to include a wider range of countries would provide a more comprehensive view of global [redacted] collaboration patterns and situate the [redacted] relationship within a broader context.

3. Use of [redacted]: Matching: The application of [redacted] to control for confounding variables when comparing research impact is commendable. However, the authors should provide more justification for their choice of matching variables and discuss potential limitations of this method. 💬 Question Conducting sensitivity analyses to test the robustness of the findings to different methodological choices (e.g., different matching variables, alternative measures of impact) would strengthen the validity of the results.

Weaknesses:

1. Causal inference limitations: While the study demonstrates a correlation between [redacted] and [redacted], it falls short of establishing causality. The authors acknowledge this limitation but could do more to explore alternative explanations for their findings. Incorporating qualitative data, such as interviews with researchers engaged in cross-border collaborations, could provide deeper insights into collaboration dynamics and help elucidate potential causal mechanisms 🙋 Action.

Fig. 1. User interface used in the study to highlight review points that stand out to participants.

5.2 Materials and Procedure

Each participant submitted one research document (either a paper or a proposal) along with corresponding human reviews through a survey once we confirmed their participation. We asked for full paper and grant proposals only due to the relatively higher quality human reviews received compared to shorter papers or posters. We generated AI reviews for each submission using our developed pipeline (detailed in Section 4). The study was conducted remotely, with participants sharing their screens over a video conferencing tool and accessing our web-based design probe using their preferred browser. In the study session, we presented study materials via a custom-built Streamlit web interface (Figure 1), designed to facilitate smooth reading and annotation of reviews. This interface displayed 3 synthetic reviews alongside 2-4 human reviews for each participant, with the number of human reviews varying based on the original feedback received for their work.

The UI enabled annotation through a highlighting and tagging system employing four general labels:

- *Positive*: Captured favorable reactions, from slightly to highly positive, such as perceiving high accuracy.
- *Negative*: Encompassed unfavorable reactions, from slightly to highly negative, like identifying overlooked aspects.
- *Question*: Indicated follow-up queries, including requests for clarification.
- *Action*: Denoted actions prompted by the review, such as conducting additional experiments.

This set of labels was initially chosen based on insights from the formative study (Section 3), and refined through multiple rounds of pilot testing, was intentionally designed to be broad to capture participants' unique interpretations while minimizing cognitive load. It was designed to elicit richer insights and provide data suitable for post-hoc thematic analysis.

The study sessions, ranging from 60 to 90 minutes, were structured into three main parts: pre-study questions and tutorial phase (~10 min), annotation and assessment phase (~40 min), and semi-structured interview phase (~10 min).

During the annotation phase, participants engaged with two sets of reviews and were encouraged to think aloud, with researchers asking probing questions to elicit rationales and contextual information. To set appropriate expectations, we clarified that the synthetic reviews were not meant to be a gold standard, encouraging honest reactions. We also presented the following scenario to foster openness to feedback: "Imagine you are still in the revision phase of this work, and there is no immediate resubmission deadline. You should be open to a wider range of feedback than you typically would if you needed to submit soon." The semi-structured interview explored topics including signals of variation and diversity in reviews and their impact on perceptions, comparisons between all reviews seen, both within each set and between human and synthetic, and prior experiences with feedback and reviews, focusing on quality and diversity aspects.

5.3 Data Collection

We collected two types of data in our user study:

(1) **Sentence-level review annotation and explanations**: Participants' annotations on sentences in both synthetic and human reviews across the four labels, i.e., *Positive, Negative, Question, Action*. We also include the participants' transcribed verbal explanations for the annotations. With participants' consent, we will open-source this data to help future research establish robust benchmarks to test LLM's capabilities [45] on real-world use cases of synthetic reviews. It covers over 800 annotations with detailed rationales.

(2) **Interview data**: Participants' response to semi-structured interview questions. We divided our interview into three sections (1) their prior experiences with feedback *before reading the text*; (1) participants' perceptions of review diversity *as they read the text*, (2) comparisons between human and synthetic reviews *after reading each review*.

5.4 Data Analysis

We used reflexive thematic analysis (RTA) to guide our data analysis. Braun & Clarke describe RTA as a theoretically flexible method for analyzing and interpreting patterns across a qualitative dataset [15]. This approach acknowledges that the researcher's position and contribution is a necessary and important part of the process, emphasizing the term "reflexive": as researchers, we draw from our own experiences, pre-existing knowledge, and social position to critically interrogate how these aspects influence and contribute to the research process and potential insights into qualitative data [15].

Our research team comprises three co-authors with extensive experience in both receiving and providing academic reviews in the broad field of computing research, as well as working with LLMs. This interdisciplinary expertise informs our approach to analyzing participants' interactions with human and AI-generated feedback. Our interdisciplinary backgrounds shape our user-centered approach and provide insights into the technical aspects of generating and analyzing text. Specifically, we are informed by our working understanding of LLMs (both in terms of practical know-how, as well as near future capabilities of these foundational models). These experiences inform and shape how we conceptualized this work, and therefore how we analyzed our data.

Our analysis process encompassed three main components, corresponding to the types of data collected (as described in Section 5.4). Initially, we transcribed the interviews using Otter.ai [2], with manual corrections for any system misunderstandings. We then enriched the data by integrating verbal rationales for each annotated review portion, using video transcripts of the study sessions. For the interview data, we employed an iterative thematic analysis approach.

We began with open coding of interview transcripts using Google Sheets [24]. These codes were then clustered and grouped on Mural [1] to develop potential themes. Through ongoing discussions of codes, participant quotes, and emerging themes, we refined our analysis to a set of candidate themes. As we drafted the paper, these themes were further developed, culminating in the final themes reported here, which reflect our perspective as HCI researchers. To gain a nuanced understanding of participants' perceptions of synthetic and human reviews, we conducted a separate analysis of the four annotation labels (positive, negative, action, question). This process involved identifying common features, reactions, and characteristics of feedback that participants associated with each label.

6 PERCEIVED VARIATIONS OF SYNTHETIC FEEDBACK

In this section, we demonstrate that the diversity in our generated synthetic feedback was indeed reflected and perceived by participants. We present several ways in which participants recognized these variations across the synthetic reviews.

6.1 Backgrounds and Expertise

To generate the diverse synthetic reviews, we created diverse personas and opinion profiles to generate varied synthetic reviews. These personas incorporated different domain expertise, methodological preferences, and research experience relevant to the paper, and the opinion profiles reflected varying levels of agreement on different topics in the paper. Our findings indicate that variation was indeed reflected in the synthetic feedback, and participants were able to infer the different academic backgrounds of the synthetic reviewers. For some participants, the way they perceive this type of diversity stems from the habit of constructing reviewer personas to help with processing feedback. P3 illustrated that they would "*create different kind of personas based on these reviews. [...] I'll add that kind of to like my corpus of reviewers who are possible for my argumentative writing, who am I applying this to? Or who am I trying to like write this for?*"

More specifically, to understand the unique expertise of synthetic reviewer, participants often drew on the review's use of terminologies and theories, as well as their focuses in review topics. Participants discussed noticeable differences in review's use of terminologies and theories. For instance, P5 noted a distinct focus on machine learning and engineering in one review because of the frequent terms used: "*Yeah, I think this reviewer seems to be focusing more on the ML stuff, like, the vocabulary that they're using, right? Like, edge cases, etc. They seem to have a bit more of, like, an engineering or data science kind of background compared to the other two reviewers.*"

P16 discussed how the synthetic feedback's use of different theories indicates their domain expertise: "*Reviewer three is definitely a development psychologist or community psychologist, probably teaches a class in child development, it's very theory heavy on that field. Reviewer two is probably coming trained in computer science, it's all about developing an intervention, interventionist type of perspective.*"For example, synthetic review three said "*The authors' attempt to examine [...] across developmental stages (infancy to pre-teen) is commendable. However, the paper lacks [...] in developmental psychology, which significantly weakens its analysis and conclusions.*" Synthetic review two instead emphasized on the technological implications of the work: "*As a technologist, I'm particularly interested in how these findings could inform future technology design: [...] Discussing how emerging technologies (e.g., AI, IoT) could potentially be applied to address some of the unique challenges of CF management identified in the study.*"

Participants also discerned how the reviews' particular emphasis on a specific point in the paper could indicate their different backgrounds. For instance, P7 observed that one review is particularly technology-focused because it was "very focused on the technical, the system implications of the paper," whereas their paper was mostly qualitative focused. Similarly, P18 shared her perceptions of two reviews: "*I also feel like Reviewer two provides more like diverse perspectives feedback of this work. So for example, you mentioned like arrow analysis, you mentioned ethical considerations* *but I feel like Reviewer one is, have like a more narrow perspective compared to Reviewer two.*" In this case, Reviewer 2 was perceived as addressing a different set of issues, including methodological considerations (error analysis) and broader implications (ethical considerations) than Reviewer 1, which was seen as having a narrower focus, potentially delving deeper into fewer areas.

6.2 Attitudes

When generating the sets of diverse reviews, while we did not directly prompt the variation in the overall attitudinal stance of the synthetic reviews on the paper, in some cases, the variations in viewpoints and personas indirectly lead to the variations in attitude. This aspect was surfaced in participants' discussions of the review, as participants mentioned notable differences in tone at the sentence level and in the balanced discussion of pros and cons at the high level.

First, we found that participants noticed diversity in the tone of the reviews. P15 reflected on this aspect: "*I will say there was a diversity of tone. And I do think like getting reviews is scary. And sometimes having like having the um, empathetic statements sometimes do find and having the mix.*" This observation highlights how the synthetic reviews successfully replicated the range of tones typically encountered in academic feedback, from critical to empathetic. Importantly, P15 found value in this tonal diversity, noting that it can help manage the emotional aspects of receiving feedback. This suggests that variability in tone is not just about realism but also about providing a more supportive and psychologically considerate feedback experience.

Additionally, participants noted that there were attitudinal differences in the reviews. Specifically, P14 noted that the balance in the discussions of strengths and weaknesses led to perceived differences in attitudes: "*Compared to the other one, this one has a much better balancing of strengths and weaknesses. I haven't read it in depth yet, but it has sections that aren't just two lines on what's good and then ten points on what's bad.*" Additionally, P11 highlighted how the reviews' emphasis on positives and negatives indicated their recommendation to the paper (e.g., accept, reject, etc.): "*I think the last reviewer ended by saying this is a good paper and should be accepted. Reviewer two, on the other hand, said it's a good paper but there are some limitations to it. So the endings seemed as if there were two different people, one recommending publication and one highlighting more limitations*". This highlights that the synthetic reviews could replicate the range of judgments typically encountered in academic peer review. Some reviews provided more positive assessments and recommendations for acceptance, while others emphasized limitations and areas for improvement.

7 PERCEIVED BENEFITS OF SYNTHETIC DIVERSE FEEDBACK

Participants can not only detect the feedback variation through backgrounds and attitudes in the probe, they also **recognized and valued diversity across synthetic reviews and its potential for comprehensive evaluation of academic work**. This appreciation stemmed from the understanding that diverse perspectives offer a more holistic assessment. In participants' prior experience getting diverse feedback, they benefit from the varied backgrounds and interpretations for more holistic evaluations. For instance, P14 highlighted the value of expertise diversity "*It's sometimes nice to have different people take away different things from the paper or just focus on different things.*" In this section, we dive into the specific key benefits of synthetic variability in feedback.

7.1 Suggestions of Novel Ideas and Blindspots

The synthetic diverse feedback shows potential in generating different perspectives that lead to useful revision ideas, some of which are absent in the human reviews. As encapsulated by P5, "*I was actually surprised at how many different angles for future work or discussion that the synthetic feedback brought to mind. So I think that almost felt equally if* *not more helpful than the human review.*" P18 further noted that the synthetic feedback suggested "*testing the impact of the generated response lens,*" which wasn't mentioned in human reviews, remarking, "*Only AI mentioned this. I'm surprised too.*" By generating diverse viewpoints and using these profiles to create feedback, our approach produced suggestions that combined different areas of expertise – in this case, misinformation research with AI-generated content analysis. P18 further articulated their surprise and appreciation for this identification of a useful research angle: "*I feel like it inspires me to test [ModelName]'s ability across, for example, we can divide all of the misinformation posts into different groups based on their manipulation strategies and test whether our misinformation detection model performs consistently better than the baselines.*" This demonstrates how the synthetic feedback's suggestion led to an actionable, novel experimental design idea.

Our diversification of synthetic reviewer profiles also introduced perspectives from related but unfamiliar fields to the researchers. For example, P8 expressed surprise when encountering considerations of quantum communication in their proposal about internet infrastructure: "*It gave me stuff I legitimately hadn't thought about before, like quantum communication stuff. Yeah, that is potentially a game-changer.*" The synthetic feedback not only offered conceptual ideas but also pointed to real-world resources previously unknown to the researcher (P8, P15). As P8 encapsulated, "*it told me about this lab that I didn't know about, it told me about this thing... this looks like a super relevant piece of work that I had never heard of.*"

The diversity in synthetic reviews revealed subtle but important communication gaps that might not be apparent from a single perspective. When different reviews interpreted or emphasized aspects of the work differently, it highlighted areas where the authors' intended message wasn't consistently coming across. P17 noted: "*It makes me know which of the scientific elements for the experimental design to keep because some of the feedback [...] wasn't necessarily accurate because they might not have a deep understanding of the technology itself or of the experiments.*" This variability in interpretations across reviews prompted reflection on how to communicate more effectively to diverse audiences. P17 continued: "*So I think we need to hone in on that a little bit more. That's really helpful.*" The reflection encourages the researcher to refine the work.

Furthermore, the complementary nature of different reviews was also found beneficial. P17 observed that "*out of the three, probably a mixture of one and two had the most helpful feedback to act on.*" This participant noted that Reviewer 1 focused more on technical aspects and project feasibility, while Reviewer 2 emphasized regulatory compliance and legal considerations. These complementary emphases enable the researcher to consider a broader range of factors that could impact their work.

7.2 Convergence of Critical Issues

Since participants can perceive diversity across the different reviews, when they find that multiple reviewers converge on the same point, they find that point to be more convincing and of higher significance. The convergence of opinions from diverse sources lends more weight and credibility to the identified issues, enabling researchers to prioritize and address critical concerns with increased confidence. Participants in our study consistently emphasized the importance of this convergence. P2 highlighted the value of multiple perspectives, stating "*[If] multiple people [...] [have] a problem with something [...] then this is something I need to address.*" The benefits are further elaborated by P15, "*I think having both, well, having multiple reviews is always helpful because you can see what are the consistent points, I guess... and that's valuable. So it's valuable seeing the intersection and then also getting that variety.*"

We found that the diverse backgrounds of the synthetic review when they converge on the same point, the perceived diversity in expertise also allows for a more holistic critique of academic work. For instance, P17 observed, "*One is* *saying the tech part of this proposal is too ambitious. The other one is saying that the community engagement interview social factors part is too ambitious. And they're both like, this is too ambitious for this grant because it's not enough time, not enough money.*" This shows different reviewers approaching the same issue from distinct angles can help the researcher form a richer understanding of the problem. Moreover, the variety in reviewer focus and questioning styles contributes to a more comprehensive evaluation of academic work. As P15 noted, "*R5 is mostly like, why is it the way that it is? And then R7 was like, but what does it mean? And I think both of those were important questions and needed to be addressed in different ways.*"

7.3 Encouragement and willingness to improve the work

The diversity in attitudes and tones across synthetic reviews provided participants with a balanced feedback experience, offering both validation and constructive criticism. This variation in attitudinal stances contributed to a more encouraging overall experience, as participants could find affirmation in positive comments while also receiving actionable suggestions for improvement.

The presence of varied attitudes across reviews allowed participants to validate their own expectations while also considering new viewpoints. As P13 observed, "*It's interesting, the previous reviewer said we did a thorough evaluation. And also, the dataset limitation, especially the size, the diversity, I think we kind of claim that the main contribution is the approach*". This contrast in opinions prompted reflection on different aspects of their work. The mixture of tones across reviews allowed participants to critically evaluate the feedback received. As one P13 stated, "*I think, either he or she does not understand the concept or maybe they made a mistake [...] I think this is just bullshit, but this review [...] I think I discussed how many poses were used for calibration. If not, it's not that important*". This demonstrates that participants felt empowered to dismiss feedback they disagreed with while still considering other points.

Similarly, P6 found that synthetic reviews offered a mix of validation and critique, noting that one reviewer provided "*more validating, on-topic feedback*" while another offered "*more critical, in-depth comments*". This combination of on-topic validation and comprehensive critique allowed the participant to feel reassured about their paper's focus while also receiving detailed suggestions for improvement.

8 COMPARING HUMAN AND SYNTHETIC VARIATIONS IN FEEDBACK

While participants were able to perceive useful diversity across the synthetic feedback, we also found some key differences in how this synthetic diversity compares to natural variations among human reviewers. We identified three sources of key differences: the perceived degree and nature of diversity, the perceived homogeneity and repetition among the diverse reviews, and the perceived depth and authenticity of diversity in expertise.

8.1 Perception of Degree and Nature of Diversity

Participants perceived synthetic reviews as offering a higher degree of diversity compared to human reviews. As P3 noted: "*The human reviews were less diverse. [...] They both have similar expertise within this domain. [...] Synthetic reviewer three had a really good, like different perspective*". This diversity was often viewed positively, but could also be seen as inconsistent or scattered. P10 observed: "*[The] synthetic feedbacks are a lot more diverse, but human feedback converges on similar topics and issues. [...] synthetic reviews are [...] all over the place*".

The perception of diversity stemmed from how participants constructed reviewer personas based on cues in the reviews (Section 6.1). However, synthetic reviews were sometimes seen as having greater internal variability, leading to more piecemeal interpretation. P8 remarked: "*With the AI reviews, I find myself having to evaluate each point individually.*

It's like each paragraph could have been written by a different person". In contrast, human reviews were perceived as more coherent in expressing a consistent perspective. P12 explained: "*Human reviews are usually more consistent. You can tell they're coming from a specific person with a particular background and set of concerns*". This internal variability within synthetic reviews impacted how participants engaged with the feedback. While diversity of perspective was valued, the lack of a consistent voice or stance could make it challenging to contextualize the feedback. As P12 noted: "*The AI [...] jumps around more. One paragraph might be spot-on, the next totally off base. It's harder to get a sense of where it's coming from*". This highlights a tension between the desired diversity of viewpoints and the need for perceived consistency within a single review.

8.2 Perceived Homogeneity and Repetition

Participants noted that sometimes synthetic reviews would use generic terms or big words that are not specific to the paper. P14 described some reviews points to have "*very flowery language that is kind of over the top*". This observation was echoed by P11, who expressed a strong dislike for certain generic terms: "*I don't like the very generic words like 'dynamics' or 'multifaceted analysis'. I know this is from an AI. [...] It's a big word which doesn't mean anything. I feel like it's just random.*" These terms were not specific to paper but were repeatedly used, leading to an impression of homogeneity and potentially undermining the perceived diversity of the feedback, as P11 articulated, "*[...] some of them were similar. Some of the keywords were like [...] using 'monolithic' or 'dynamics.' It seems like this is coming from the same person because these are very specific.*"

The issue of repetition was not limited to the use of words but extended to broader concepts as well. The repeated mentioned focus in the reviews would also lead to a perceived similarity across feedback even if the exact review points are different. P13 noted that "*I feel that R1, R2, and R3 look very similar because they actually mention many of the same things, like context-aware, effect size, and then generalizability.*" Similarly, P17 observed that "*The synthetic feedback mentioned 'data security and data privacy' repeatedly.*" The same participant also noted "*The synthetic reviewer's repetitive emphasis on 'gold standard technologies and citing protocols' that were somewhat generic.*" This repeated mentioning of broad concepts across reviews further contributed to the perception of homogeneity and lack of tailored feedback.

Furthermore, repeated wordings across synthetic reviews were viewed less favorably than in human reviews, especially when participants initially disagreed. P15 explained: "*[If] I had received all of these reviews [...] I would have been like, ah, crap. We have to do a user study. Because they're all saying the same thing.*" This suggests pressure to act on repeated synthetic feedback, despite initial disagreement. However, when human validation aligned with synthetic feedback, researchers became more confident in accepting suggestions. P12's experience illustrates a shift in perception after encountering a human review that echoed a point repeatedly mentioned in the synthetic reviews. P12 noted, "*When I first saw the machine review, I wondered how it could spot this issue. [...] I had overestimated the depth of thought behind this comment. [...] I think it means our paper must be missing this part, for both humans and machines to say so.*"

8.3 Perceived Depth and Authenticity of Expertise Diversity

Participants perceived differences in the depth and authenticity of expertise diversity between human and synthetic reviews. Human reviews often demonstrated deeper domain knowledge and more nuanced perspectives, which participants associated with genuine expert diversity. P15 noted: "*The human [reviews] [...] were much more nuanced, and [...] comprehensive... something that I didn't see as much in the synthetic ones.*" While synthetic reviews conveyed a sense of diversity in domain knowledge and content focus, they sometimes lacked the in-depth knowledge. This absence of subjective elements paradoxically undermined the perception of authentic, deep expertise. P11 observed:

“*Obviously the attitude towards the paper and towards other issues of the humans have their own biases. So those are very prevalent, but those I don’t see in the AI reviews. AI reviews may be biased implicitly, but explicitly they are very agreeable because of the whole alignment [...] done on these models.*”

However, human reviewers can also anchor too much on their biases and unsubstantiated beliefs. P11 observed a difference in how synthetic reviews handled biases compared to human reviewers: “*Obviously the attitude towards the paper and towards other issues of the humans have their own biases. So those are very prevalent, but those I don’t see in the AI reviews. AI reviews may be biased implicitly, but explicitly they are very agreeable because of the whole alignment and that’s done on these models.*” This highlights a tension between authenticity and fairness in diverse reviews. On one hand, the depth and nuance of human expertise provide valuable insights that synthetic reviews may not replicate. On the other hand, the potential for human reviewers to be overly influenced by their biases can lead to unfair or skewed evaluations.

8.4 Perceptions of Divergent Opinions

Participants perceived divergent opinions more negatively when it comes to synthetic feedback. For example, when presented with overly positive feedback, P3 still reacted negatively, stating: “*I know, it’s a strength, but already, I’m getting like a very negative approach, where I don’t think the reviewer knows what I’m talking about... it’s way too optimistic for what was in the paper.*” Divergent opinions, even when positive, can be met with skepticism if perceived as exaggerated and not aligned with the researchers’ own opinion. P14 articulated their expectation for synthetic feedback: “*With LLM... I kind of have this understanding that it had been given all of this*”. This expectation potentially led to lower tolerance for divergent viewpoints in AI-generated reviews.

In contrast, when faced with differing opinions in human reviews, participants often rationalized or contextualized these differences rather than dismissing them outright. P4 noted: “*Usually where I feel like the human reviewers didn’t get [...] the intention of the study, I would just skip that part*”. Even when acknowledging significant deviations, participants displayed a more forgiving attitude towards human reviewers. P12 demonstrated this tendency: “*I think the 2AC, that R5, is very low quality [...] I feel like he didn’t even read the paper [...] But I think he could be an expert, he’s just writing carelessly*”. This suggests that participants were more willing to attribute divergent human opinions to factors like individual quirks or carelessness, rather than fundamental lack of expertise.

9 DISCUSSION

While recent work has explored LLMs’ capabilities in generating paper reviews [39], supporting academic research [46], and simulating different opinions [13], whether language models can generate diverse research feedback and whether researchers would perceive such diversity as useful remains unclear. Through our study, we generated a set of diverse synthetic feedback as a probe to explore researchers’ perceptions of diversity in academic reviews. We uncovered the types of synthetic variability that are perceivable and valuable, as well as where synthetic diverse feedback still falls short. Participants identified variability in reviewer backgrounds and attitudes (Section 6), appreciating how this diversity led to more comprehensive evaluations and novel insights, convergence of important issues, and validation that enhances the willingness to improve (Section 7). We also identified ways that AI-generated synthetic variations differ from naturally emerged variations among human reviewers (Section 8), informing future work designing more useful synthetic diverse feedback. These findings serve as a crucial first step in understanding how to generate meaningful diversity in academic feedback, pointing towards a future where AI-generated diverse perspectives could complement human expertise throughout the research process.

9.1 Implications for Engaging with and Designing for Synthetic Diverse Feedback

9.1.1 Intentional perspective seeking. Our study reveals that LLMs can simulate perceived and useful diverse perspectives in academic feedback. Participants recognized variability across reviews in reviewer backgrounds and expertise and attitudinal stances across synthetic reviews. Researchers could strategically use synthetic diverse feedback to complement their existing feedback-seeking processes. Here we outline a few scenarios. For experienced researchers, who clearly understand the types of feedback and specific domain expertise they need, can leverage this capability to their advantage. They can prompt LLMs to generate targeted reviews from specific viewpoints or areas of expertise that might not be readily available through traditional peer review processes. Also, researchers working on an interdisciplinary project could request feedback from the perspective of multiple relevant disciplines, gaining insights that might be challenging to obtain from a limited pool of human reviewers. Moreover, researchers can request feedback with different focus and levels of specificity based on their current stage in the research process. For instance, in the early stages of a project, researchers might benefit from feedback that suggests larger-scale revisions and broader conceptual shifts.

While our findings demonstrate the potential of synthetic feedback to simulate diverse domain expertise that led to novel insights often missed by human reviewers, we found limitations in replicating authentic, nuanced expertise. One fundamental challenge is that LLMs may not possess or put enough attention on domain experts' different sets of in-depth procedural or tacit knowledge [55, 61] - the "know-how" of conducting research, applying methodologies, or interpreting results - was often misrepresented or missing in AI-generated feedback. The procedural knowledge that shapes how experts approach problems, frame questions, or contextualize findings within broader disciplinary debates was not consistently represented. This tacit understanding, often unwritten but crucial to academic discourse, proved challenging for AI systems to simulate convincingly.

System designers building a research feedback tool could incorporate interactive prompting mechanisms allowing researchers to refine the AI's knowledge base. Users could specify key papers, methodologies, or ongoing debates, helping to fill gaps in the AI's procedural and implicit knowledge. This could involve describing common practices, unwritten rules, or typical interpretation frameworks used in their field. On the technical side, more research efforts are needed to correctly identify the related work of a paper and retrieve relevant domain knowledge to form a correct representation.

9.1.2 Support Navigating Repetitive Review Points. Our findings revealed a nuanced tension: while repetition in issues across perceived diverse reviews is generally beneficial (Section 7.2), repetitive points with the same wordings, often coupled with generic terms, that present in the synthetic feedback can sometimes lead to a sense of homogeneity and break the illusion of diversity, which undermines the perceived importance of critical issue convergence (Section 8.2). After all, are the repetitive points raised in synthetic reviews perceived to reflect a genuine consensus on the problem or a mere glitch by the LLMs?

To address these challenges, systems presenting synthetic feedback should support understanding repetitive comments. When facing a large number of unstructured LLMs responses, future work can incorporate sense-making design features to structure, organize, and potentially integrate automated analysis over the subpar review feedback [23]. For instance, a potential design feature in our context is categorizing repetition types between surface-level similarities and deeper conceptual overlap, and then presenting this analysis alongside reviews. The variation can be captured using color coding or tags to indicate "Shared concern with different rationales" versus "Similar phrasing but distinct points." On another note, when faced with diverse but long synthetic reviews, design features to summarize the right high-level feedback may help users effectively process this information. Summary features have been widely used for, e.g., reading papers [4] and conversing online [78]. In our context, the system may offer options to view the repetition as a summary. More concretely, a "Convergent Issues" section could list topics mentioned by multiple reviewers, with expandable details showing each reviewer's specific take on the issue. This would help users identify consensus areas while preserving nuancesMoreover, concise reviewer profiles may enhance the user engagement with customized "personas" [28] that show expertise, methodological preferences, and general attitudes.

9.1.3 Calibrating Expectations and Fostering Constructive Engagement with Synthetic Feedback. Our findings reveal that participants often approach synthetic feedback with preconception about the LLMs' capabilities. The persistent questions of "How does the LLM know this?" and "Do I trust that it really understands this?" create a cognitive burden, potentially limiting the benefits of synthetic diverse perspectives. While our study reveals increased scrutiny of synthetic feedback, further research is needed to determine whether this stems from participants' knowledge of the feedback source or their perceptions of LLM-generated content's characteristics. Future work should investigate how awareness of LLM authorship and specific output attributes influence users' reception of synthetic feedback.

This skepticism towards synthetic feedback also relates to the challenges in LLM interpretability research [66]. While LLMs can generate post-hoc natural language explanations to elucidate their rationale, these explanations may appear plausible yet be inconsistent with the model's actual outputs. Beyond the fundamental challenge of trust in generative models for expert-level tasks, the context of diverse research feedback presents unique difficulties. While expected to be beneficial, embracing different perspectives and addressing varied feedback demands additional cognitive effort from researchers. As some participants noted, while they value challenging feedback for improving their work, they are "happier" when seeing easy-to-fix suggestions (P1, P11, P12, P16). This highlights the need for reframing synthetic feedback from an evaluator to a collaborative thought partner that assist reflection [6]. This approach may invite creative engagement by presenting feedback as "potential avenues for exploration" rather than definitive critiques. Future work can explore how to use design interventions to shape the expectations of feedback. For example, designers could explore affording temporal flexibility that alleviates the initial overwhelm, removes the sense of urgency associated with traditional reviews, and assists reflective thinking upon the feedback.

9.1.4 Navigate the text-heavy output. While synthetic diverse feedback can offer novel ideas and uncover blind spots, the nature of variations often results in valuable insights scattered within a large volume of text. Our dataset of review annotations (Section 5.3) reveals that researchers engage with synthetic diverse feedback selectively, focusing on notably insightful, positive, or critically contested points. This approach differs from the obligation to address all comments when researchers receive human reviews. Of 496 highlighted synthetic review instances, only 56 prompted follow-up questions and 107 led to potential actions, indicating targeted engagement. Participants often made quick judgments based on alignment with their expertise, sometimes dismissing feedback after reading just the initial sentence. For instance, P11 immediately disregarded a comment about insufficient technical details as irrelevant. This selective engagement allows researchers to extract valuable insights from a large volume of text, but risks overlooking potentially useful feedback. The scattered nature of valuable insights within synthetic feedback presents both opportunities and challenges. While it can offer novel ideas and uncover blind spots, it requires researchers to carefully sift through the content, which was called out by P2 as "sparse".

Building on this observation, we can envision a fluid and dynamic approach to using synthetic diverse feedback, one that allows researchers to seamlessly transition between different purposes and levels of analyses. LLMs can facilitate this fluidity by generating high-level summaries of key themes across reviews, supporting quick navigation and comprehensive issue coverage. LLMs' ability to handle various queries could enhance this approach, allowing for more specific follow-up questions when initial feedback lacks sufficient justification. While our study didn't explore iterative questioning, our dataset captures potential follow-up questions that researchers might ask, pointing to future possibilities for more interactive feedback systems.

9.2 Subjectivity in Feedback Engagement and Leveraging Dataset Insight

Our study reveals that research feedback interpretation remains highly subjective and context-dependent, despite the line of work assessing the review quality through more generalized criteria [58]. Researchers, when processing feedback, draw upon their entire distilled knowledge base, project-specific experience, intentions, and understanding to evaluate the utility and relevance of advice. This context-rich engagement with feedback highlights the limitations of standardized approaches to feedback assessment. Given this subjectivity, system designers should consider integrating feedback tools into researchers' natural workflows. By embedding data collection within existing processes, we can capture the nuanced ways researchers interact with and interpret feedback. In our study, participants expressed interest for the annotation tool like our study UI, finding it intuitive for capturing their thoughts on reviews. This suggests an opportunity to adapt existing researcher workflows by providing a tool when researchers are parsing through human reviews they've received and simultaneously collecting valuable data to be fed into language models, tuning the models' representation of effective feedback patterns and researcher needs.

A corollary from our study is a dataset of comprising 496 sentence-level annotations from synthetic reviews and 362 from human reviews, each accompanied by researchers' rationales on how they perceive, interpret, and engage with feedback (See details in Section 5.3). While this dataset already enable us to qualitatively and systematically analyze users' perceptions on synthetic feedback, this may also help future quantitative analysis and model work to achieve some of the design features that we proposed in subsection 9.1. For instance, this dataset may enable analysis of specific characteristics of feedback that succeed or fail to account for the unique challenges of interdisciplinary research [47]. Also, by analyzing the patterns in sentences labeled as "Action" items, future work could develop models that prioritize concrete, implementable suggestions in synthetic feedback. Future work can explore the differences in textual signals between human and synthetic reviews. For instance, Lee et al. [36] has contributed a rich benchmark dataset that enables rich analysis between authors who are human and GPT-3 for features, e.g., spelling and grammatical errors. While these questions were beyond the scope of our current study, they represent promising avenues for developing nuanced synthetic feedback criteria to enhance LLMs for targeted and actionable research feedback. This dataset from experienced scholars is critical because most proposals for synthetic reviews from the NLP communities might focus the technical aspects but out of contexts [41].

10 LIMITATIONS

In this work, we found that participants were able to perceive synthetic diversity in the set of reviews we generated using our pipeline. However, our participants mainly came from social science and computing backgrounds related to HCI, future work is needed to assess whether our approach and findings can generalize to broader academic domains. For instance, a recent large-scale quantitative study showed that NLP experts identify some marginal degree of novelty in LLM-generated research ideas [64]. This raises questions of how researchers in other fields perceive longer form of LLM-generated content, such as research feedback in our context. As such, a qualitative study like ours in other fields may add nuanced insights to some of our findings. Within HCI, on the other hand, the 18 experienced researchers might not fully represent the range of experience and potential perceptions toward LLM-generated feedback, whether critical or accepting. Regardless, much of our findings can be viewed as a starting point to motivate discussions on incorporating LLMs in the HCI research process, a topic that has received growing attention in our field [3], in which paper review is an integral part of.

In addition, our study can pose ethical and privacy concerns. Using LLMs to generate feedback requires us to incorporate real papers into our prompts (see Section 4), but papers at this stage are often confidential. Sharing this information to closed models may risk leaking users' private information, which can be subject to idea misappropriation. We checked these services' privacy statements to make sure all inputs to the models during the generation process are not used to train, retrain, or improve models. We also took additional precautions when recruiting our participants by (1) obtaining an IRB of this study, (2) explicitly asking for consent before and during our user study, and (3) deleting the users' papers after generating the review. However, if scaling up, this approach might bear the same privacy concerns. The centralization of LLMs behind the veil of a proprietary API offers virtually no transparency, into how they leverage or store the user data [8]. While we wished to use open models, we decided that the GPT-4 and Claude-3.5-Sonnet models could generate high-quality feedback to preserve the ecological validity of our study results. Meanwhile, it is of community interest to study the potential consequences of using LLMs in synthetic review as well as researchers' perceptions of easily accessible models that has been proposed and even used by researchers.

REFERENCES

[1] MURAL 2024. *MURAL*. MURAL. https://mural.com/ Accessed: 2024-08-21.

[2] Otter.ai 2024. *Otter*. Otter.ai. https://otter.ai/ Accessed: 2024-08-21.

[3] Marianne Aubin Le Quéré, Hope Schroeder, Casey Randazzo, Jie Gao, Ziv Epstein, Simon Tangi Perrault, David Mimno, Louise Barkhuus, and Hanlin Li. 2024. LLMs as Research Tools: Applications and Evaluations in HCI Data Work. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–7.

[4] Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A Hearst, Andrew Head, and Kyle Lo. 2023. Paper plain: Making medical research papers approachable to healthcare consumers with natural language processing. *ACM Transactions on Computer-Human Interaction* 30, 5 (2023), 1–38.

[5] Christian Baden and Nina Springer. 2014. Com (ple) menting the news on the financial crisis: The contribution of news users' commentary to the diversity of viewpoints in the public debate. *European journal of communication* 29, 5 (2014), 529–548.

[6] Eric PS Baumer. 2015. Reflective informatics: conceptual dimensions for designing technologies of reflection. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 585–594.

[7] Abeba Birhane, Atoosa Kasirzadeh, David Leslie, and Sandra Wachter. 2023. Science in the age of large language models. *Nature Reviews Physics* 5, 5 (2023), 277–280.

[8] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[9] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201* (2023).

[10] Kwangsu Cho, Tingting Rachel Chung, William R King, and Christian Schunn. 2008. Peer-based computer-supported knowledge refinement: An empirical investigation. *Commun. ACM* 51, 3 (2008), 83–88.

[11] Kwangsu Cho and Charles MacArthur. 2010. Student revision with peer and expert reviewing. *Learning and instruction* 20, 4 (2010), 328–338.

[12] Kwangsu Cho and Christian D Schunn. 2007. Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. *Computers & Education* 48, 3 (2007), 409–426.

[13] Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T Rogers. 2023. Simulating opinion dynamics with networks of llm-based agents. *arXiv preprint arXiv:2311.09618* (2023).

[14] Jocalyn Clark and Reshma Jagsi. 2021. Peer review: economy, identity, diversity. *European Science Editing* 47 (2021), e76284.

[15] Victoria Clarke, Virginia Braun, and Nikki Hayfield. 2015. Thematic analysis. *Qualitative psychology: A practical guide to research methods* 3 (2015), 222–248.

[16] Ed Collins, Isabelle Augenstein, and Sebastian Riedel. 2017. A supervised approach to extractive summarisation of scientific papers. *arXiv preprint arXiv:1706.03946* (2017).

[17] Petra De Saá-Pérez, Nieves L Díaz-Díaz, Inmaculada Aguiar-Díaz, and José Luis Ballesteros-Rodríguez. 2017. How diversity contributes to academic research teams performance. *R&d Management* 47, 2 (2017), 165–179.

[18] Rochelle DeCastro, Dana Sambuco, Peter A Ubel, Abigail Stewart, and Reshma Jagsi. 2013. Mentor networks in academic medicine: moving beyond a dyadic conception of mentoring for junior faculty researchers. *Academic Medicine* 88, 4 (2013), 488–496.

[19] Gina Dokko, Aimée A Kane, and Marco Tortoriello. 2014. One of us or one of my friends: How social identity and tie strength shape the creative generativity of boundary-spanning ties. *Organization Studies* 35, 5 (2014), 703–726.

[20] Steven Dow. 2011. How prototyping practices affect design results. *Interactions* 18, 3 (2011), 54–59.

[21] Siamak Faridani, Ephrat Bitton, Kimiko Ryokai, and Ken Goldberg. 2010. Opinion space: a scalable tool for browsing online comments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1175–1184.

[22] C Ailie Fraser, Tricia J Ngoon, Ariel S Weingarten, Mira Dontcheva, and Scott Klemmer. 2017. CritiqueKit: A mixed-initiative, real-time interface for improving feedback. In *Adjunct Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*. 7–9.

[23] Katy Ilonka Gero, Chelse Swoopes, Ziwei Gu, Jonathan K. Kummerfeld, and Elena L. Glassman. 2024. Supporting Sensemaking of Large Language Model Outputs at Scale. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 838, 21 pages. https://doi.org/10.1145/3613904.3642139

[24] Google. 2024. *Google Sheets: Free Online Spreadsheets for Personal Use*. Google. https://www.google.com/sheets/about/ Accessed: 2024-08-21.

[25] Michael D Greenberg, Matthew W Easterday, and Elizabeth M Gerber. 2015. Critiki: A scaffolded approach to gathering design feedback from paid crowdworkers. In *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition*. 235–244.

[26] Ken Gu, Ruoxi Shang, Tim Althoff, Chenglong Wang, and Steven M Drucker. 2024. How Do Analysts Understand and Verify AI-Assisted Data Analyses?. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–22.

[27] Lars Guenther, Claudia Wilhelm, Corinna Oschatz, and Janise Brück. 2023. Science communication on Twitter: Measuring indicators of engagement and their links to user interaction in communication scholars' Tweet content. *Public Understanding of Science* 32, 7 (2023), 860–869.

[28] Juhye Ha, Hyeon Jeon, Daeun Han, Jinwook Seo, and Changhoon Oh. 2024. CloChat: Understanding How People Customize, Interact, and Experience Personas in Large Language Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 305, 24 pages. https://doi.org/10.1145/3613904.3642472

[29] Shirley Anugrah Hayati, Minhwa Lee, Dheeraj Rajagopal, and Dongyeop Kang. 2023. How Far Can We Extract Diverse Perspectives from Large Language Models? Criteria-Based Diversity Prompting! *arXiv preprint arXiv:2311.09799* (2023).

[30] Mohammad Hosseini and Serge PJM Horbach. 2023. Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other large language models in scholarly peer review. *Research integrity and peer review* 8, 1 (2023), 4.

[31] Kristen Intemann. 2009. Why diversity matters: Understanding and applying the diversity component of the National Science Foundation's broader impacts criterion. *Social Epistemology* 23, 3-4 (2009), 249–266.

[32] Joseph Kahne, Ellen Middaugh, Nam-Jin Lee, and Jessica T Feezell. 2012. Youth online activity and exposure to diverse perspectives. *New media & society* 14, 3 (2012), 492–512.

[33] Hyunwoo Kim, Haesoo Kim, Kyung Je Jo, and Juho Kim. 2021. StarryThoughts: facilitating diverse opinion exploration on social issues. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–29.

[34] Michèle Lamont. 2009. *How professors think: Inside the curious world of academic judgment*. Harvard University Press.

[35] Michèle Lamont, Grégoire Mallard, and Joshua Guetzkow. 2006. Beyond blind faith: overcoming the obstacles to interdisciplinary evaluation. *Research Evaluation* 15, 1 (2006), 43–55.

[36] 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

[37] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems* 36 (2023), 51991–52008.

[38] Ming Li, Jiuhai Chen, Lichang Chen, and Tianyi Zhou. 2024. Can llms speak for diverse people? tuning llms via debate to generate controllable controversial statements. *arXiv preprint arXiv:2402.10614* (2024).

[39] Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Yi Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, et al. 2024. Can large language models provide useful feedback on research papers? A large-scale empirical analysis. *NEJM AI* (2024), AIoa2400196.

[40] Chien Hsiang Liao. 2011. How to improve research quality? Examining the impacts of collaboration intensity and member diversity in collaboration networks. *Scientometrics* 86, 3 (2011), 747–761.

[41] Q. Vera Liao and Ziang Xiao. 2023. Rethinking Model Evaluation as Narrowing the Socio-Technical Gap. arXiv:2306.03100 [cs.HC] https://arxiv.org/abs/2306.03100

[42] Alex Liu and Min Sun. 2023. From Voices to Validity: Leveraging Large Language Models (LLMs) for Textual Analysis of Policy Stakeholder Interviews. *arXiv preprint arXiv:2312.01202* (2023).

[43] Ryan Liu and Nihar B Shah. 2023. Reviewergpt? an exploratory study on using large language models for paper reviewing. *arXiv preprint arXiv:2306.00622* (2023).

[44] Yiren Liu, Si Chen, Haocong Cheng, Mengxia Yu, Xiao Ran, Andrew Mo, Yiliu Tang, and Yun Huang. 2024. How ai processing delays foster creativity: Exploring research question co-creation with an llm-based agent. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–25.

[45] Yu Lu Liu, Su Lin Blodgett, Jackie Cheung, Q. Vera Liao, Alexandra Olteanu, and Ziang Xiao. 2024. ECBD: Evidence-Centered Benchmark Design for NLP. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 16349–16365. https://aclanthology.org/2024.acl-long.861

[46] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292* (2024).

[47] Miles MacLeod. 2018. What makes interdisciplinarity difficult? Some consequences of domain specificity in interdisciplinary practice. *Synthese* 195, 2 (2018), 697–720.

[48] Nora Madjar. 2005. The contributions of different groups of individuals to employees' creativity. *Advances in developing human resources* 7, 2 (2005), 182–206.

[49] Elizabeth Mannix and Margaret A Neale. 2005. What differences make a difference? The promise and reality of diverse teams in organizations. *Psychological science in the public interest* 6, 2 (2005), 31–55.

[50] Angela Barron McBride, Jacquelyn Campbell, Nancy Fugate Woods, and Spero M Manson. 2017. Building a mentoring network. *Nursing outlook* 65, 3 (2017), 305–314.

[51] Shweta Mishra. 2020. Social networks, social capital, social support and academic success in higher education: A systematic review with a special focus on 'underrepresented'students. *Educational Research Review* 29 (2020), 100307.

[52] Michael D Mumford and Sigrid B Gustafson. 1988. Creativity syndrome: Integration, application, and innovation. *Psychological bulletin* 103, 1 (1988), 27.

[53] Michèle B Nuijten, Chris HJ Hartgerink, Marcel ALM Van Assen, Sacha Epskamp, and Jelte M Wicherts. 2016. The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior research methods* 48 (2016), 1205–1226.

[54] Vishakh Padmakumar and He He. 2023. Does Writing with Language Models Reduce Content Diversity? *arXiv preprint arXiv:2309.05196* (2023).

[55] Vimla L Patel, Jose F Arocha, and David R Kaufman. 1999. Expertise and tacit knowledge in medicine. In *Tacit knowledge in professional practice*. Psychology Press, 89–114.

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

[57] 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*. 1–16.

[58] Charvi Rastogi, Ivan Stelmakh, Alina Beygelzimer, Yann N Dauphin, Percy Liang, Jennifer Wortman Vaughan, Zhenyu Xue, Hal Daumé III, Emma Pierson, and Nihar B Shah. 2024. How do authors' perceptions of their papers compare with co-authors' perceptions and peer-review decisions? *Plos one* 19, 4 (2024), e0300710.

[59] Zachary Robertson. 2023. Gpt4 is slightly helpful for peer-review assistance: A pilot study. *arXiv preprint arXiv:2307.05492* (2023).

[60] Bryan Semaan, Heather Faucett, Scott P Robertson, Misa Maruyama, and Sara Douglas. 2015. Designing political deliberation environments to support interactions in the public sphere. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 3167–3176.

[61] Jacqueline Senker. 2008. *The Contribution of Tacit Knowledge to Innovation*. Springer London, London, 376–392. https://doi.org/10.1007/978-1-84628-927-9_20

[62] Nihar B Shah. 2022. Challenges, experiments, and computational solutions in peer review. *Commun. ACM* 65, 6 (2022), 76–87.

[63] Nikhil Sharma, Q Vera Liao, and Ziang Xiao. 2024. Generative Echo Chamber? Effect of LLM-Powered Search Systems on Diverse Information Seeking. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17.

[64] Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2024. Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers. *arXiv preprint arXiv:2409.04109* (2024).

[65] Tejpalsingh Siledar, Swaroop Nath, Sankara Sri Raghava Ravindra Muddu, Rupasai Rangaraju, Swaprava Nath, Pushpak Bhattacharyya, Suman Banerjee, Amey Patil, Sudhanshu Shekhar Singh, Muthusamy Chelliah, et al. 2024. One Prompt To Rule Them All: LLMs for Opinion Summary Evaluation. *arXiv preprint arXiv:2402.11683* (2024).

[66] Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. 2024. Rethinking interpretability in the era of large language models. *arXiv preprint arXiv:2402.01761* (2024).

[67] 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*. 120–137.

[68] David Tinapple, Loren Olson, and John Sadauskas. 2013. CritViz: Web-based software supporting peer critique in large creative classrooms. *Bulletin of the IEEE Technical Committee on Learning Technology* 15, 1 (2013), 29.

[69] Rayden Tseng, Suzan Verberne, and Peter van der Putten. 2023. ChatGPT as a commenter to the news: can LLMs generate human-like opinions?. In *Multidisciplinary International Symposium on Disinformation in Open Online Media*. Springer, 160–174.

[70] Marjo Van Zundert, Dominique Sluijsmans, and Jeroen Van Merriënboer. 2010. Effective peer assessment processes: Research findings and future directions. *Learning and instruction* 20, 4 (2010), 270–279.

[71] Jeroen PH Verharen. 2023. ChatGPT identifies gender disparities in scientific peer review. *Elife* 12 (2023), RP90230.

[72] Linlin Xu and Tiefu Zhang. 2023. Engaging with multiple sources of feedback in academic writing: postgraduate students' perspectives. *Assessment & Evaluation in Higher Education* 48, 7 (2023), 995–1008.

[73] Ilan Yaniv. 2004. The benefit of additional opinions. *Current directions in psychological science* 13, 2 (2004), 75–78.

[74] Yu-Chun Grace Yen, Joy O Kim, and Brian P Bailey. 2020. Decipher: an interactive visualization tool for interpreting unstructured design feedback from multiple providers. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[75] Steven Bethard Ryan Cotterell Yichao Zhou, Iz Beltagy and Tanmoy Chakraborty. 2024. *ACL pubcheck.* https://github.com/acl-org/aclpubcheck Accessed: 2024-08-21.

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

[77] Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2022. Can we automate scientific reviewing? *Journal of Artificial Intelligence Research* 75 (2022), 171–212.

[78] Amy X Zhang, Lea Verou, and David Karger. 2017. Wikum: Bridging discussion forums and wikis using recursive summarization. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 2082–2096.

[79] Jiayao Zhang, Hongming Zhang, Zhun Deng, and Dan Roth. 2022. Investigating fairness disparities in peer review: A language model enhanced approach. *arXiv preprint arXiv:2211.06398* (2022).

[80] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems* 36 (2023), 46595–46623.

A SAMPLE REVIEWER PERSONA

Synthetic Reviewer 1 (P10):

- Background: Associate Professor in Human-Computer Interaction
- Research domain: Technology-enhanced learning, Computer-supported cooperative work
- Expertise: Qualitative research methods, Educational technology design
- Related personal experience: Has conducted several studies on collaborative learning tools and has some familiarity with counselor training processes

Synthetic Reviewer 2 (P10):

- Background: Senior Researcher at a major tech company
- Research domain: AI-assisted learning, Intelligent tutoring systems
- Expertise: Machine learning, Quantitative evaluation methods, System architecture
- Related personal experience: Has worked on several large-scale learning platforms, primarily in STEM fields

Synthetic Reviewer 3 (P10):

- Background: Assistant Professor in Digital Health
- Research domain: Health informatics, Computer-supported therapy
- Expertise: Interaction design for healthcare, Mixed methods research
- Related personal experience: Has collaborated with mental health professionals on technology interventions and has first-hand experience with counseling training

B SAMPLE VIEWPOINTS

Synthetic Reviewer 1 (P18):

- Skeptical about the time constraint findings, believing real-time responses are still crucial in high-stakes situations.
- Enthusiastic about multimodal understanding, seeing it as essential for future fact-checking systems.
- Advocates for comprehensive corrections, arguing that thorough explanations are necessary for long-term belief change.
- Cautiously optimistic about AI surpassing humans in fact-checking, but emphasizes the need for human oversight.
- Concerned about relying too heavily on external credibility ratings, preferring a more dynamic, context-aware approach to source evaluation.

Synthetic Reviewer 2 (P18):

- Agrees with the paper's findings on time constraints, arguing that quality matters more than speed in most cases.
- Skeptical about the importance of multimodal understanding, believing text-based fact-checking is sufficient for most scenarios.
- Prefers concise corrections, arguing that brevity is key to capturing and maintaining audience attention.
- Strongly opposed to AI surpassing human fact-checkers, emphasizing the importance of human judgment and contextual understanding.

- Supportive of the source credibility evaluation method, seeing it as a necessary step to combat misinformation from unreliable sources.

Synthetic Reviewer 3 (P18):

- Neutral on time constraints, believing the importance of speed varies greatly depending on the type and potential impact of misinformation
- Moderately supportive of multimodal understanding, but questions whether the 33% improvement justifies the additional complexity.
- Advocates for adaptive approaches to correction length, tailoring comprehensiveness to the specific misinformation and audience.
- Intrigued by AI's potential to surpass humans in certain aspects, but emphasizes the need for AI-human collaboration rather than replacement.
- Critical of relying on a single source (Media Bias/Fact Check) for credibility ratings, preferring a more diverse and transparent evaluation system.