Designing interactions with AI to support the scientific peer review process

Lu Sun [1] **Stone Tao** [2] **Junjie Hu** [3 4] **Steven P. Dow** [1]

Abstract

Peer review processes include a series of activities from review writing to meta-review authoring. Recent advances in AI exhibit the potential to augment complex human writing activities. However, it is still not clear how to design interactive systems that leverage AI to scaffold the peer review process and what potential trade-offs may arise. In this paper, we prototype a system – *MetaWriter*, which uses three forms of AI to support meta-review authoring including review aspect highlights, viewpoint extraction, and hybrid draft generation. In a within-subjects experiment, 32 participants wrote meta-reviews using MetaWriter and a baseline environment with no machine support. We show that MetaWriter can expedite and improve the meta-review authoring process. However, participants raised concerns about trust, over-reliance, and agency. We further discuss general insights on designing interactions with AI to support the scientific peer review process.

1. Introduction

Peer review is the cornerstone of scientific research. As an essential step for ensuring the scientific quality of work, peer review has been adopted by most journals and conferences. While conferences may have different configurations in their peer-review process, they always involve multiple stakeholders, including paper authors, reviewers, and meta-reviewers, as shown in the Figure1. In a typical conference review process, each reviewer needs to provide reviews for their assigned papers. Then, a discussion for each paper then takes place between its reviewers and meta reviewer - who are intermediaries between reviewers and program chairs. In some conferences, the author may then provide a rebuttal to the review, which could clarify the misunderstandings in the reviews. Based on all information, the meta-reviewer then recommends to the program chairs a decision about whether or not to accept the paper to the conference.

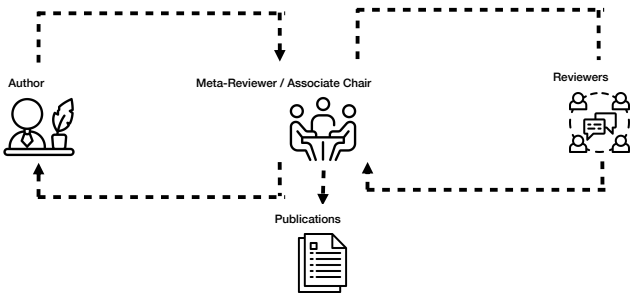

Figure 1. Peer-review ecosystem

While the rapid increase in paper submissions can be viewed as a positive indicator of scientific progress, it has also created burdens and challenges on the peer review process for multiple stakeholders (McCook, 2006; Shah et al., 2018). Reviewers have to take on more submissions which can lead to a slow process, more inconsistencies (Smith, 2006) and potential biases (Langford & Guzdial, 2015; Shah et al., 2018). Furthermore, as research continues to advance, peer review processes consistently involve novice reviewers who face challenges in understanding community standards and expectations. Additionally, meta-reviewers encounter more challenges, including maintaining review quality, ensuring a fair decision-making process, facilitating communications between reviewers and authors, and assembling the most relevant information from the reviews.

In this era, AI and LLMs show tremendous powers to support humans on writing tasks. Previous research has shown the promise of LLMs in a variety of use cases, such as legal document summaries (Norkute et al., 2021), scripts generations (Mirowski et al., 2023) or even scientific writing (Gero et al., 2022). However, less is explored in the context of academic peer review. In this project, *we explore potentials and risks of designing interactive systems that leverage AIs in scientific peer review processes.* Current science communities have adopted interfaces to support the review writing process and the discussion process for reviewers,

[*]Equal contribution [1]Cognitive Science Department, University of California San Diego, San Diego, USA [2]Computer Science Department, University of California San Diego, San Diego, USA [3]Department of Computer Science, University of Wisconsin-Madison, Madison, USA [4]Department of Biostatistics, University of Wisconsin-Madison, Madison, USA. Correspondence to: Lu Sun <l5sun@ucsd.edu>, Steven P.Dow <spdow@ucsd.edu>.

AI & HCI Workshop at the 40^{th} International Conference on Machine Learning (ICML), Honolulu, Hawaii, USA. 2023.

meta-reviewers and authors. During these processes, how might the interaction be integrated into the workflows while considering the main design needs and values of multiple stakeholders? AI interactions may provide power to scaffold novices, raise awareness of biases, and improve efficiency for experts (Gero et al., 2022; Shah et al., 2018).

We examine the challenges of research in the intersection of AI and HCI for scientific peer review processes through the example of a prototyping system - MetaWriter. MetaWriter is a prototype that facilitates meta-review authoring and fact-checking through aspect tagging and text generation. MetaWriter consists of three main functions: (1) **aspect tagging** that uses a pre-trained tagger to highlight crucial aspects in each review for decision-making in the meta-review process; (2) **extractive summarization** to call attention to key sentences in the reviews and (3) **hybrid meta-review generation** that uses a hybrid meta-generation model to generate a draft for meta-reviewers based on extracted sentences.

Recently, there have been a few research attempts that explore the possibility of generating meta-reviews automatically (Bhatia et al., 2020; Kumar et al., 2021; Shen et al., 2022). Notably, Shen et al. (2022) proposed a controllable meta-review generation method to generate meta-reviews according to the categories of reviews. However, meta-reviewers still have to briefly understand where the generated draft comes from and proofread the factuality of the generated draft. Hence, instead of a one-step generation, we used a hybrid model to generate a meta-review and visualize the extracted sentences from each review. Different from previous work mainly focusing on the meta-review generation task, MetaWriter system supplies a range of functions (e.g. aspect visualization, meta-review generation) to scaffold not only the writing but also the fact-checking in the entire meta-review process.

To systematically evaluate the effectiveness of MetaWriter, we compare our hybrid meta-review generation model with multiple state-of-art meta-review summarization models. Furthermore, we conducted a within-subject experiment where we invited 32 researchers who have review or meta-review experiences to write meta-reviews using MetaWriter versus a plain text editor. The experiment demonstrates that MetaWriter helps to shorten the meta-reviewing time and improves the meta-review quality post-edited by humans. We shed light on the design considerations of interactive systems that leverage AI to support peer review processes.

2. MetaWriter

2.1. System Overview

As shown in Figure 2, MetaWriter provides three main functions: (1) the system visualizes the results of an aspect tagger to facilitate reading and sense-making on each review; (2)the system highlights important sentences that users can hover on to cross-compare between reviews; (3) the system generates meta-review using a hybrid model and shows each step's results on the interface to improve transparency on meta-review generation and fact-checking.

F1: Visualizing tagged aspects We apply a pre-trained tagger (Yuan et al., 2021) to color-code each word according to its aspect. In our dataset, 30.4% of words are tagged with an aspect, including summary, substance originality, clarity, soundness, etc, as shown in Figure 3-A. MetaWriter color-codes these tags on each independent review to support proofreading. Yuan et al. (2021) released this aspect tagger to explore the possibility of paper review generation. Different from their use case, we run this aspect tagger to help meta-reviewers efficiently check the reviews.

F2: Highlighting extracted sentences Extractive summarization techniques have shown evidence of selecting important sentences for a summary. Here, we fine-tuned a pre-trained extractive summarization model to select key sentences from each review to shorten participants' time on reading the long reviews (Liu & Lapata, 2019).

F3: Generating meta-review draft using a hybrid model To capture the useful content of all reviews, we use a hybrid model that first extracts the candidate sentences from each review and then uses the extracted sentences, together with their ratings and the paper abstract to generate a meta-review. Unlike the previous studies that directly combined all reviews to fine-tune a generative model (Shen et al., 2022; Kumar et al., 2021), this hybrid model decomposes the writing procedure into an extraction-then-generation pipeline, which enhances the system transparency and fact-checking since meta-reviewers can easily identify which sentences are used to generate a draft. Our hybrid model is similar to Bhatia et al. (2020), but we use more recent pre-trained language models for extractive and abstraction summarizations. Moreover, different from the system MRed (Shen et al., 2022), where the meta-reviewers still may have to scan through each review and decide which aspects to control the generation, MetaWriter automatically predicts aspects and extracts a structured draft from each review, which enables faster proofreading. MetaWriter then provides meta-reviewers with an automatically generated draft for post-editing. The implementation details are described below.

2.2. Dataset

To train the ML models that can automatically extract key sentences and generate a meta-review draft, we collected a large peer review dataset from the online peer-reviewing platform OpenReview [1] for ICLR, one of the largest ma-

[1] https://openreview.net/

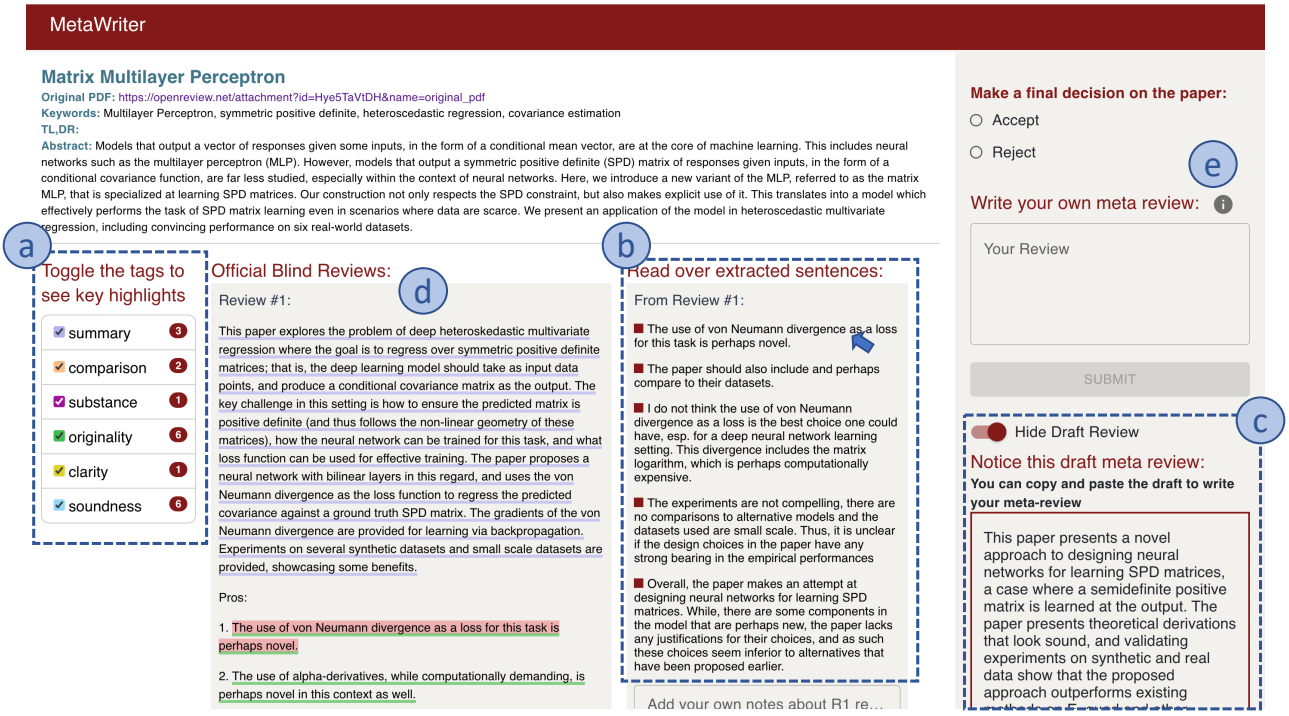

Figure 2. MetaWriter interface. (A) users can toggle the tags to highlight specific aspects. (B) users can hover over the extracted sentences to locate the corresponding positions highlighted in each review. (C) users can see or hide the generated draft using the toggle button. Other basic function includes that: Users can review the three original independent reviews and users can make a final decision and write their meta-review on the right.

chine learning conferences. We collected each submission's data using the OpenReview API and scraped reviews from all publicly accessible submissions from the year 2018 to 2022. For each submission, we collected the paper title and abstract, all official reviews with reviewer ratings and confidence scores as well as the final meta-review with ratings and the final decision. After filtering out submissions that had fewer than 3 reviews and meta-review that had less than 20 words, we retained 9,803 submissions along with their corresponding meta-reviews and 34,219 independent reviews. To simplify the meta-review authoring process in the study, we only collected the original independent reviews and dropped the discussion comments.

2.3. Hybrid Summarization

With the design goal of maintaining transparency and facilitating fact-checking for the generated meta-review draft, MetaWriter used hybrid summarization approach.

Extractive Summarization: We fine-tuned an extractive summarization model to extract important sentences from each review, as shown in Figure 3-B. We create an extractive summarization dataset from our ICLR dataset to train the extractive summarization model. The inputs are individual reviews and the labels are generated via a beam search procedure on the individual review (Carbonell & Goldstein, 1998; Xu & Durrett, 2019; Liu & Lapata, 2019). The goal of this procedure is to label the sentences which got incorporated into the final meta-review. In particular, during beam search for each additional sentence we propose to add to the label, we compute a heuristic cost equal to the ROUGE-L score of a given sentence with respect to the reference summary written by ICLR meta-reviewers (Lin, 2004; Lin & Och, 2004). Here, we iteratively loop over sentences from each review and only keep sentences when the ROUGE_L score between the selected sentences and the original ICLR meta-review improves.

Abstractive Summarization: Meta-reviews typically synthesize all reviews into one cohesive summary without directly copying sentences from each (Shen et al., 2022). As shown in Figure 3 - C, to generate a draft that is realistic and natural, we used an abstractive summarization method that combines similar sentences together to generate a draft (Bhatia et al., 2020; Shen et al., 2022; Kumar et al., 2021; Lewis et al., 2019). To train the abstractive summarization model, for each submission, we combined the extracted sentences from all three reviewers using the extractive summarization above, along with their ratings and the paper abstract as inputs, and then use the real meta-review as the output target. We fine-tuned the bart-large-cnn model, one variant of the BART model, on our dataset (Lewis et al., 2019).

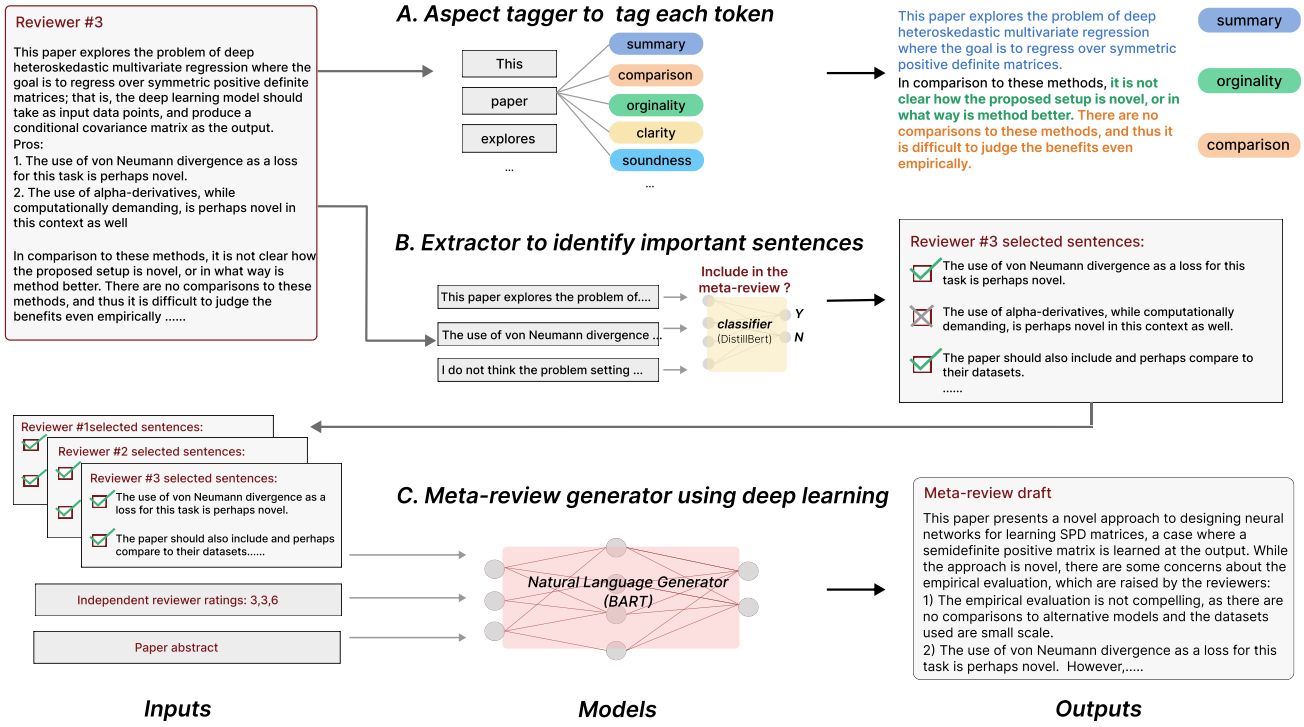

Figure 3. Platform that can run trained tagger, extractive summarization model and abstractive generation model to facilitate the meta-review process.

2.4. User Interface Design

As shown in Figure 2, MetaWriter simulates OpenReview which provides the paper abstract, a link to the paper pdf, keywords, and three original independent reviews. Figure 2-A shows the output of the aspect tagger as a set of six categories and their frequency of appearing in the reviews. When users toggle the checkbox before the category, the corresponding categories will get underlined with the colors. Figure 2-B lists all sentences extracted by MetaWriter for each review. When participants hover over an extracted sentence, the corresponding sentence will be highlighted in red. In Figure 2-B, when the user hovers over the sentence "The user of vonNeumann divergence as a loss for this task is perhaps novel", MetaWriter will highlight the sentences in red in the middle area. This allows meta-reviewers **cross-check** the location of the original sentences. Figure 2-C provides the generated meta-review draft by the abstractive summarization model based on extracted sentences. Users can select to show or hide the meta-review if they think the draft may influence their judgment. The meta-reviewers still have to make their own choice in the upper right corner. To avoid the strong biases that can be contained in the meta-review draft, we also remove the sentences that indicate the decision in the provided draft. Participants can post-edit the meta-review. The system is implemented using React, Typescript, and backend servers.

3. Evaluation and Case Study

3.1. Hybrid Meta-review Generation Model

Baselines We employ two representative extractive summarization baselines: (1) **LexRank** computes a graph-based centrality score to select important sentences (Erkan & Radev, 2004); (2) **TF-Extract** is a Transformer-based extractive model that follows PreSumm (Liu & Lapata, 2019) to use a fine-tuned BERT model for extractive summarization. In our experiment, we use both methods to extract the five most important sentences from each review and concatenate all reviewers' extracted sentences as a meta-review.

TF-Abstract is a Transformer-based abstractive model that is fine-tuned from a bart-cnn-large model (Lewis et al., 2019) using all reviewer's original reviews without extraction. Our two-step **Hybrid** model first uses a fine-tuned BERT model to perform extractive summarization on each review, then uses a fine-tuned BART model to generate a meta-review draft based on all extracted sentences as well as the predicted decision and reviewers' ratings. During training, we fine-tune the BERT model in the same way as PreSum, and we fine-tune the bart-cnn-large model to generate the final meta-review draft. To evaluate our hybrid model, we compare its performance with the Transformer-based extractive summarization model (Liu & Lapata, 2019) and abstractive summarization model (Lewis et al., 2019) alone.

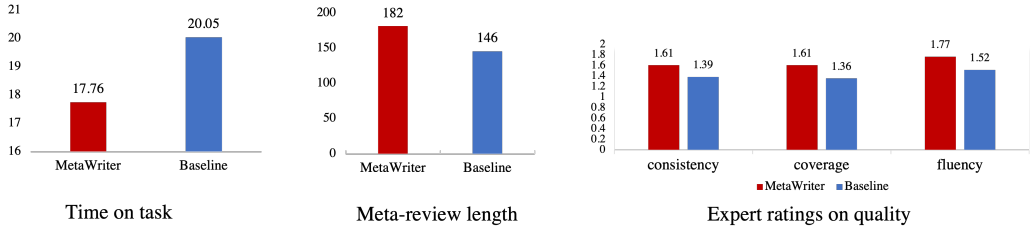

Figure 4. User evaluation study

Model	$\mathbf{R}_1$	$\mathbf{R}_2$	$\mathbf{R}_L$
LexRank	0.202	0.031	0.173
TF-Extract	0.271	0.068	**0.234**
TF-Abstract	0.334	0.091	0.203
Hybird	**0.329**	**0.099**	**0.215**

Table 1. Meta-review generation results on the Rouge_1, Rouge_2, and Rouge_L F_1 scores

Automatic Evaluation As shown in Table 1, we evaluate the meta-review quality by the ROUGE F_1 scores between the generated meta-review and the ground-truth meta-review (Lin, 2004). Our two-step hybrid generation approach can generate comparable or slightly better meta-reviews than the Transformer-based extractive and abstractive summarization models.

3.2. Case Study: Controlled Experiment

To evaluate the efficiency of writing a meta-review with and without the aid of MetaWriter in terms of time and quality, we conduct a with-in-subjects experiment with 32 participants who have review experience or meta-review experience. We simulate a realistic meta-reviewing process and invite participants with ML conference reviewing experience to play the role of a meta-reviewer to summarize independent reviews and write meta-reviews. Each participant completes two meta-reviews: one with the MetaWriter system (MetaWriter condition) and another with only a plain text editor (Baseline condition). We counterbalance the order of each condition through random assignments. Our study selects two borderline rejected papers that (1) have their reviews and meta-review of average word length and (2) have high ROUGE_L scoring meta-reviews generated by our hybrid system.

Our results show that MetaWriter reduces meta-reviewing time compared to the baseline editor. We use the analysis of covariance (ANCOVA) test to examine the effect of the two conditions on writing time and control the length of the meta-review as a co-variate. As shown in Figure 4, participants spent significantly less time writing meta-reviews when using the MetaWriter than the baseline condition. We find that participants write statistically significantly longer meta-reviews with the MetaWriter system. To further measure the quality, we recruit two experts who have more than five years of machine learning research background and have experience being an ICLR meta-reviewer to rate the quality of all final meta-review (N=64) with a 3-dimension rubric: coverage, consistency and fluency, using a simple three-point scale (0-2). Figure 4 shows that MetaWriter helps reviewers write more consistent, fluent meta-reviews with higher coverage.

4. Discussion

To explore the potential and risks of designing interaction with AI to support peer review processes, we designed a MetaWriter system and run a within-subject case study. Our experiment found the benefits of leveraging AI to support meta-review writing, but meta-reviewers also expressed concerns about using AI in the peer review process and raise the issue around trust, agency, and over-reliance.

4.1. Risks posed by AI in peer review writing

Previous researchers have raised concerns about the limitations of AI and LLMs (Yuan et al., 2022). Prior studies have also reported that LLMs can exhibit stereotypes or biases on text generation tasks (Weidinger et al., 2021). Based on our experiments, we emphasize the risks of using AI in the peer-review writing context: LLMs are not transparent and may potentially have biases when they generate text (Weidinger et al., 2021). Proving the generated draft may have the risk of misleading meta-reviewers in their decision-making.

Based on our experiments, we claim that an automatic meta-review generation system can not replace human meta-reviewers. Instead, AI should scaffold inexperienced meta-reviewers in this meta-review process to guide them to write better meta-reviews. Meta-reviewers should not fully rely on MetaWriter's support and directly rely on the MetaWriter to submit their meta-review without reading reviews and paper content and make a fair judgement.

4.2. Designing interactions that balance the trade-offs between efficiency and agency

The MetaWriter system provided participants with different levels of automated support, from simple extraction methods to visually highlight common aspects and key arguments, to a deep learning-based natural language generation model

that generates an initial draft. One interesting result we found is that even though participants perceived the generated draft as being helpful, they also felt that this feature contributed to a sense of less control. The tension between efficiency and user agency in hybrid human-machine systems raises the question of how to effectively design to balance this trade-off.

In the peer review context, participants expressed the importance of maintaining **control** of the final decision and justification. One method to preserve a sense of agency is to give users the ability to adjust the degree of machine support they receive. Maybe interactive systems can dynamically tune their level of support for different types of participants (Kocielnik et al., 2019), such as only showing the highlighted tags and generated drafts for users with little or no reviewing experience. Another strategy would be to make the AI models themselves controllable and editable (Yuan et al., 2022) so that the draft will generate accordingly based on the meta-reviewer's own decisions.

4.3. Engage multiple stakeholders when designing AI systems to support peer review

Peer review is a complex process that involves multiple stakeholders, including authors, reviewers, and meta-reviewers (Shah, 2022). There is always tension between multiple stakeholders in this high-stake context. In the within-subject experiments, participants pointed out the potential problem of over-reliance on the generation feature and cannot provide fair and justified decisions. They are also concerned about the potential trust-breaking between reviewers and authors. Our findings suggest that future designs that intend to incorporate AI in a similar high-stake context should incorporate multiple stakeholders and provide more transparency. For example, authors and other reviewers need to know which part of the meta-review is incorporated from the generated draft and how much influence the AI brings to the decision-making.

4.4. Designing interactions with AI for academic peer review

How might we design interactions leveraging AI in the academic peer review process? Our case study of MetaWriter indicates that instead of using the automatically generated draft, participants prefer to use multiple hybrid methods to write meta-review. Participants elaborated that the interactions with AI artifacts can provide inspiration in the process of conducting their meta-review, but they were also concerned about trust and over-reliance. We consider their interaction with AI as one type of partnership relationship, where AI is designed to fit into the existing human task workflow and assist parts of tasks according to human needs (Wang et al., 2019). Designing interactions with AI for the peer review workflow should also consider providing enough scaffolding for novices, maintaining agency for experts, and raising awareness of biases.

AI can potentially support the entire peer review ecosystem and our findings can further guide the design of AI to support other tasks in the academic peer review cycle. For paper reviewers, AI interactions can potentially provide scaffolding for novice reviewers to deliver a high-quality paper review (Yuan et al., 2021; Wang et al., 2020). AI interactions can also raise their awareness of the potential biases and the tone of voices. For paper authors, AI interactions can help them analyze the reviews and draft a more persuasive rebuttal (Gao et al., 2019). In the future, we plan to extend our work to design more human-AI collaboration systems to scaffold novice stakeholders in the academic peer review ecosystem. In addition, the current study has limitations of only conducting experiments in a machine learning conference peer review context. We imagine that these interactions can have the potential for a broader range of academic communities.

5. Ethics

MetaWriter is a design prototype with the goal of understanding the potentials and risks of using AI to support the peer review process. We claim that an automatic metareview generation system can not replace human meta-reviewers. Instead, it only plays a role in scaffolding reviewers to write higher quality meta-reviews.

References

Bhatia, C., Pradhan, T., and Pal, S. Metagen: An academic meta-review generation system. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1653–1656, 2020.

Carbonell, J. and Goldstein, J. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 335–336, 1998.

Erkan, G. and Radev, D. R. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479, 2004.

Gao, Y., Eger, S., Kuznetsov, I., Gurevych, I., and Miyao, Y. Does my rebuttal matter? insights from a major nlp conference. *arXiv preprint arXiv:1903.11367*, 2019.

Gero, K. I., Liu, V., and Chilton, L. Sparks: Inspiration for science writing using language models. In *Designing Interactive Systems Conference*, pp. 1002–1019, 2022.

Kocielnik, R., Amershi, S., and Bennett, P. N. Will you accept an imperfect ai? exploring designs for adjusting end-user expectations of ai systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2019.

Kumar, A., Ghosal, T., and Ekbal, A. A deep neural architecture for decision-aware meta-review generation. In *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pp. 222–225. IEEE, 2021.

Langford, J. and Guzdial, M. The arbitrariness of reviews, and advice for school administrators. *Communications of the ACM*, 58(4):12–13, 2015.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Lin, C.-Y. and Och, F. J. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pp. 605–612, 2004.

Liu, Y. and Lapata, M. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*, 2019.

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

Mirowski, P., Mathewson, K. W., Pittman, J., and Evans, R. Co-writing screenplays and theatre scripts with language models: Evaluation by industry professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–34, 2023.

Norkute, M., Herger, N., Michalak, L., Mulder, A., and Gao, S. Towards explainable ai: Assessing the usefulness and impact of added explainability features in legal document summarization. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–7, 2021.

Shah, N. B. An overview of challenges, experiments, and computational solutions in peer review (extended version). *Communications of the ACM*, 2022.

Shah, N. B., Tabibian, B., Muandet, K., Guyon, I., and Von Luxburg, U. Design and analysis of the nips 2016 review process. *Journal of machine learning research*, 2018.

Shen, C., Cheng, L., Zhou, R., Bing, L., You, Y., and Si, L. MReD: A meta-review dataset for structure-controllable text generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2521–2535, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.198. URL https://aclanthology.org/2022.findings-acl.198.

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

Wang, D., Weisz, J. D., Muller, M., Ram, P., Geyer, W., Dugan, C., Tausczik, Y., Samulowitz, H., and Gray, A. Human-ai collaboration in data science: Exploring data scientists' perceptions of automated ai. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–24, 2019.

Wang, Q., Zeng, Q., Huang, L., Knight, K., Ji, H., and Rajani, N. F. ReviewRobot: Explainable Paper Review Generation based on Knowledge Synthesis, December 2020. URL http://arxiv.org/abs/2010.06119. arXiv:2010.06119 [cs].

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

Xu, J. and Durrett, G. Neural extractive text summarization with syntactic compression. *arXiv preprint arXiv:1902.00863*, 2019.

Yuan, A., Coenen, A., Reif, E., and Ippolito, D. Wordcraft: Story writing with large language models. In *27th International Conference on Intelligent User Interfaces*, pp. 841–852, 2022.

Yuan, W., Liu, P., and Neubig, G. Can we automate scientific reviewing? *arXiv preprint arXiv:2102.00176*, 2021.