MetaWriter: Exploring the Potential and Perils of AI Writing Support in Scientific Peer Review

LU SUN, STONE TAO, STEVEN P. DOW, University of California San Diego

Recent advances in Large Language models (LLMs) show the potential to significantly augment or even replace complex human writing activities. However, for complex tasks where people need to make decisions as well as write a justification, the trade offs between making work efficient and hindering decisions remain unclear. In this paper, we explore this question in the context of providing intelligent scaffolding for writing meta-reviews as part of the academic peer review process. We prototyped a system called "MetaWriter" trained on 5 years of open peer review data to support meta-reviewing. The system highlights common topics in the original peer reviews, extracts key points by each reviewer, and on request, provides a preliminary draft of a meta-review that can be further edited. To understand how novice and experienced meta-reviewers use MetaWriter, we conducted a within-subject study with 32 participants. Each participant wrote meta-reviews for two papers: one with and one without MetaWriter. We found that MetaWriter significantly expedited the authoring process and improved the coverage of meta-reviews, as rated by experts, compared to the baseline. While participants recognized the efficiency benefits, they raised concerns around trust, over-reliance, and agency. We also interviewed six paper authors to understand their opinions of using machine intelligence to support the peer review process and reported critical reflections. We discuss implications for future interactive AI writing tools to support complex synthesis work.

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

Additional Key Words and Phrases: conference peer review, meta-review, summarization, machine intelligence

ACM Reference Format:
Lu Sun, Stone Tao, Steven P. Dow. 2018. MetaWriter: Exploring the Potential and Perils of AI Writing Support in Scientific Peer Review. In *CSCW '23: The 26th ACM Conference On Computer-Supported Cooperative Work And Social Computing, October 13–18, 2023, Minneapolis, MN*. ACM, New York, NY, USA, 33 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Peer review is a key cornerstone of academic research [73]. The peer review process helps improve the quality of published research by providing feedback and assessment on the paper. [44, 80]. While the rapid increase in paper submissions can be viewed as a positive indicator of scientific progress, it has also created a burden on the peer review process [62, 80, 87]. Reviewers have to take on more submissions which can lead to a slow process, more inconsistencies [84], and potential biases [53, 81]. The skyrocket in paper submissions also leads to an increase in burden and challenges for meta-reviewers – who balance the reviewers' comments and recommend to the program chairs a decision about whether or not to accept the paper to the conference.

In many academic communities, papers are evaluated by several reviewers and followed by a "meta-review" that summarizes the reviews and provides a final decision. Meta-reviewers not only need to assign reviewers, but also need to understand the core idea of the paper, make sense of each

CSCW '24, October 13–18, 2024, San Jose, Costa Rica

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

reviewer's opinions, synthesize viewpoints across reviewers, resolve conflicts between reviewers, and make a final decision with sufficient explanations for the paper's author to understand the decision and iterate on their work. In this process, writing a high-quality comprehensive meta-review and delivering a fair decision with justification together are viewed as challenging and complex work [80]. In addition, an increase in paper submissions across many disciplines has placed an undue burden on peer review communities [4]. While the burden varies by the reviewer and depends on the paper, meta-reviewing is generally viewed as complicated and cognitively demanding. With the fast development of the research community, it is not uncommon for the academic community to look for new researchers to engage in the meta-review process.

To help less experienced meta-reviewers to gain expertise quickly, scaffolding is one approach that previous research showed as an efficient instructional strategy [38, 39, 99]. Scaffolding was used in educational settings to improve learners' problem-solving skills efficiently with guidance and examples [18, 77]. Recently, the scaffolded annotation as an instruction approach helped professional writing students improve the quality of their early-stage drafting in a new genre [38]. Scaffolded examples and templates can even help learners perform similarly to experts in terms of feedback quality [99]. Similarly, in the context of meta-reviewing, junior meta-reviewers may benefit from scaffolding by modeling on existing examples, while experienced meta-reviewers can reflect on their tasks[77].

In a complex writing scenario, existing research showed that AI scaffolding and machine intelligence offer tremendous potential to collaborate with humans by providing inspirations [28, 54, 98] or generating summaries [66]. Despite these successes, researchers have raised concerns about the AI directly conducting the task instead of assisting humans in a high stake context, such as job seeking [38]. The limitations of LLMs, such as the lack of content awareness [98], fact checking [60], or the tendency to produce contradictory or contrived information [45, 60], can limit human's trust of using them. Hence, in this context of conducting meta-review, we can foresee the trade-offs that AI scaffolding can expedite the process but can also take over the meta-reviewers' agency, resulting in an over-reliance on AI suggestions.

In a nutshell, harnessing machine intelligence to scaffold complex writing tasks requires a holistic understanding of its capabilities and trade-offs. We further studied the AI scaffolding in the context of meta-reviewing where meta-reviewers were tasked with making a fair decision while writing a meta-review at the same time. *Could AI scaffolding help improve the meta-review quality as well as maintain control and agency in the writing process? What are the benefits and concerns perceived by reviewers and authors?* A key goal of our research is to gain empirical insight into the trade-offs of weaving machine intelligence into complex socio-technical processes for writing meta-review while making a decision on the paper.

To answer our research questions, we developed a prototype called MetaWriter, which can support the meta-reviewing process using text-based machine-learning techniques. MetaWriter scaffolds and facilitates meta-reviewers in meta-reviewing while still maintaining a sense of control and oversight. To achieve these goals, we drew insights from previous literature on scaffolding and human-AI collaboration to specify three design goals: scaffold inexperienced meta-reviewers by embedding expert knowledge structures and examples, preserve agency through deliberative edits, and reduce bias through staging and masking. Correspondingly, the prototype uses (1) automated tagging to color code common topics in peer review (e.g., sentences that evaluate the originality of the paper), (2) extractive summarization to call attention to key sentences in the reviews, and (3) a natural language generation model to automatically generate a preliminary draft of the meta-review based on all three reviews, the paper title and paper abstract, that can be further edited, as shown in Figure 1.

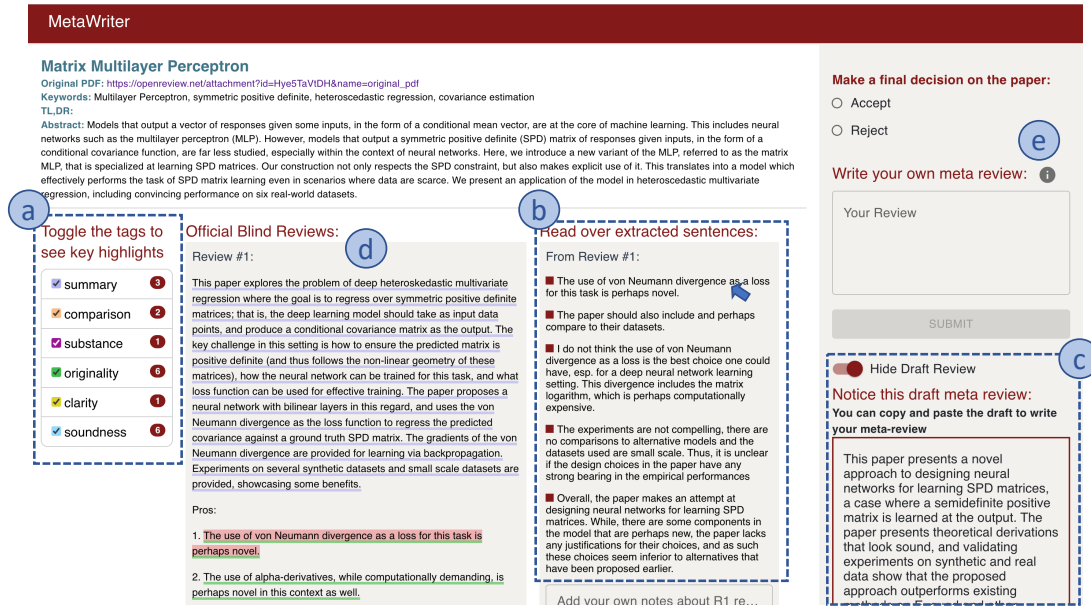


Fig. 1. MetaWriter interface. (a) users can toggle the tags to highlight specific aspects. (b) users can hover over the extracted sentences for each review and the corresponding sentence will be highlighted. (c) users can see or hide the generated draft using the toggle button. (d) users can review the three original independent reviews. (e) users can make a final decision and write their meta-review.

We designed a mixed method study to investigate whether machine intelligence provided objective performance benefits and to understand stakeholder perspectives, including reviewers and authors. In a within-subjects study, 32 participants who had prior experience writing reviews for ML conferences, wrote meta-reviews for two papers in a counterbalanced fashion: one using the MetaWriter system and one using the same writing interface but without machine support. We found that the MetaWriter system led participants to produce meta-reviews that cover more of the points raised by the independent reviewers, better summarized the paper idea and better justification on decisions. Meta-reviews written with MetaWriter are more similar and less diverse compared with meta-reviews written with Baseline Editor. MetaWriter also reduced the time for participants to generate final meta-reviews compared to the baseline editor with no machine support. We observed that participants used the three features of machine enhancements differently and all participants preferred to use the MetaWriter system over a baseline editor. However, some participants expressed concerns about over-reliance and the potential for bias to creep in when people think about how others might abuse the system. Participants reported that the machine-generated draft improved their meta-reviewing efficiency, but sacrificed some agency and flexibility in the writing process. When Comparing experienced meta-reviewers who have fruitful prior experience in meta-reviewing to inexperienced meta-reviewers who only had experience reviewing papers, we found that MetaWriter can better scaffold inexperienced meta-reviewers in identifying and covering more arguments and scaffold experienced meta-reviewers in justifying the paper content. Interviews with the paper authors also revealed similar concerns about meta-reviewers' potential over-reliance on machine intelligence and the risks of falling short of gate-keeping responsibilities in the conference review process.

Our paper offers several contributions to the CSCW and scientific peer review communities: Our mixed method study provides empirical data that carefully integrating ML techniques into an authoring environment can facilitate information synthesis and potentially expedite the meta-reviewing process. Our study brings to light concerns raised by meta-reviewers and authors alike indicating a need to increase trust, authorial control, and a sense of fairness and transparency before such systems can be adopted by peer-review communities. We further conduct critical reflections on ethical considerations of using AI in the peer review process. In addition, we fine-tuned several ML models and developed a prototype that facilitates the meta-review process by highlighting common review topics, extracting important sentences, and automatically generating an initial draft meta-review that can be further edited.

2 RELATED WORK

2.1 Conference peer review practices and challenges

Peer review is an essential step of academic research. As an essential step for ensuring the scientific quality of work, peer review has been adopted by most journals and conferences[73]. In a typical conference review process, each reviewer needs to provide reviews for their assigned papers. 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 may clarify any misunderstandings in the reviews. Based on all the information, the meta-reviewer then recommends to the program chairs a decision about whether or not to accept the paper to the conference [80, 81]. Note that conferences may have some differences in their peer-review process. For instance, some conferences or journals don't have meta-reviews and final decisions are made through discussion between all reviewers.

While the rapid increase in paper submissions can be viewed as a bloom of scientific progress, it has also created a burden on the peer review process [62, 80, 87]. This also brings more challenges for meta-reviewers in the decision-making, especially when there are conflicts among reviewers, as well as meta-review writing to provide a justification. To provide a high-quality review, reviewers need to fully understand the contribution of the paper and then provide an evaluation that covers multiple aspects, including the originality, clarity, validity, etc, along with a score on the submission [15, 24, 37, 44, 53, 84, 90, 96, 100].

Analyses of peer review practices illustrate the subjective nature of reviewing and indicate that variability between reviewers is the dominant factor that decides the fate of a submission [3, 10, 47, 71, 81]. This variability is evidenced by differences in reviewers' ratings as well as their judgments on detailed aspects of submissions [51, 57]. Previous research noticed that approximately 15% of ICLR conference publications have at least a pair of reviews that have a rating large difference – larger than 5 (note that reviewer ratings range from 1 as reject to 10 as seminal paper). The meta-reviewers have the difficult job of resolving disagreements between reviewers [57]. Sometimes these disagreements boil down to opposite opinions on the detailed aspects of the paper. For example, we observed that for one paper in the ICLR dataset [1], reviewer 2 commented that the paper is "The paper is generally well-written. The results and proof sketches are well presented and easy to follow". However, reviewer 3 commented that "The paper was a bit dense and hard to follow". Thus, to solve the conflicts and coordinate the peer review process, meta-reviewers play the role of an arbitrator and final decision maker [7, 47, 100]. To facilitate the meta-reviewing process, a prototype should simulate this meta-review scenario with high variability situation where meta-reviewers need to make decisions on borderline papers where three reviewers have disagreements with each other on the rejection or acceptance decision.

[1] https://openreview.net/forum?id=B1xxAJHFwS

Previous research has used computational methods to provide support to streamline several parts of the peer review process, such as matching submissions with appropriate peer reviewers, authoring more comprehensive and decisive reviews, as well as review quality assessment [4, 6, 37, 81, 90, 100]. However, as far as we know, there has been no previous work in the HCI community focused on supporting meta-reviewers. Meta-reviewers need to synthesize diverse multi-aspects information from authors and different reviewers and then provide a reasonable recommendation for the paper [75, 80, 82]. Specifically, they need to read and think holistically about the submission under review, evaluate reviewers' comments, and make a final decision on the conflicts raised by reviewers. This deliberation process can be time-consuming and cognitively demanding. With the rapid development of the research community, academic communities sometimes look for new senior researchers to engage in the meta-review process. However, inexperienced meta-reviewers may take some time to get used to conducting meta-review process. To help less experienced meta-reviewers to take over the meta-reviewing efficiently, scaffolding could be one approach to help them gain expertise [18, 39].

2.2 AI scaffolding for writing and information synthesis

Scaffolding was used as an instructional strategy to improve learners' problem-solving skills efficiently with guidance and examples [18, 77]. Previous research shows that effective scaffolding can help novices perform work nearly as good as experts [38, 39, 99]. Another writing support system used scaffolded annotation and examples as an instruction approach to professional writing students improve the quality of their early-stage drafting in a new genre [38]. In the context of writing introductory help requests, providing high-quality examples and expert-informed templates can increase learning and writing quality [38, 39]. Similarly, in the context of meta-reviewing, junior meta-reviewers may benefit from scaffolding by reflecting and modeling on existing examples, while experienced meta-reviewers can reflect on their tasks[77].

Existing research uses existing examples and templates to scaffold the learning process [38, 39, 99]. In the writing scenario, scholars found that scaffolding can help students learn about form and organization from analyzing the examples and templates [18, 22]. To build on existing research around scaffolding, providing meta-reviewers with structure and examples can potentially help them gain expertise more efficiently. Providing structures cal also help learners "transfer" skills into the current tasks [11, 25]

AI and machine intelligence can also provide efficient scaffolding. Previous research has developed multiple methods for using machine intelligence to scaffold complex tasks, including writing and researching [28, 66, 74, 98]. To help people understand the key points of the information [90, 100], researchers used argument mining or tagging techniques to identify important points from a massive amount of information, such as identifying key points from social media [69], online debates [31, 88], or student essays[70]. Another machine learning method that can help people understand the text is to summarize longer texts into shorter and more concise passages [20]. For example, extractive summarization techniques can identify the most relevant information from a long document and have been found to help humans make swifter decisions [20, 36, 94]. In the meta-review context, to help meta-reviewers understand each reviewer's points and facilitate the decision-making, this technique could extract important and relevant sentences from independent reviewers to helps meta-reviewers make swifter decisions [9, 36]. For example, Bhatia et al. [9] explored whether we can predict paper decisions by first creating an extractive draft using extractive summarization. Their results showed that extractive sentences increased the predictive power of paper decisions.

Furthermore, the recent development of NLP models brings the opportunity for human-AI co-creation where human and AI can collaboratively generate artifacts including poems, essays, and

figures.[17, 26–28, 41, 55, 64, 75]. For example, the Wordcraft project explores how to support users collaborating with generative language models to co-write a story [98]. Spark used a language model to generate prompts related to a scientific concept to facilitate scientific writing [28].

However, the use of models such as LLMs in these collaborative systems comes with a number of caveats that inhibit their usefulness. LLMs are highly context-dependent and their generated outputs can be subjectively interpreted [55, 98]. As a result, over-reliance of machine generated outputs from LLMs in meta-review writing processes can have the potential to sway the opinion of meta-reviewers and inject unwanted biases into the review process.

2.3 Concerns of AI scaffolding on agency, bias and over-reliance

Existing research showed that AI scaffolding and machine intelligence offer tremendous potential to collaborate with humans [28, 54, 66, 98]. One prior research project explored the possibility of generating meta-reviews automatically. After using extractive summary techniques to identify key sentences, Bhatia et al. fine-tuned a sequence-to-sequence model to generate meta-reviews [9]. Kumar et al. [50] used a deep neural network architecture to generate a meta-review that would account for paper decisions. Shen et al. [82] proposed a controllable meta-review generation method to generate meta-reviews according to the categories of reviews. For example, given the "strength" and "reviewer suggestion" categories, the model can generate a meta-review that summarized the strength of the work that reviewers mentioned with the suggestions reviewers provided. However, researchers have raised concerns about the AI directly conducting the task instead of assisting humans to gain expertise. In addition, the limitations of LLMs, such as the lack of content awareness [98] or the tendency to produce contradictory or contrived information [45, 60], can limit human's trust of using them. Hence, in this context of conducting meta-review, we need to explore the trade-offs that AI scaffolding can expedite the process but can also take over the meta-reviewers' agency and control.

Recent guidelines for human-AI interaction have also highlighted that systems should provide user controls to globally customize what the AI system monitors and how it behaves [2, 35, 83]. Some researchers have tried integrating agency and automation by creating effective "collaborative" interfaces using shared representations between humans and AI [46, 93].

For systems where humans collaborate with an AI system, researchers have noted that the efficiency gains from automation might also limit the agency and creativity of human users [33]. In the context of meta-review writing processes, if the machine automatically generates a high-quality draft, meta-reviewers might put forth less effort to make a decision and may miss key points that were not integrated by the machine. Understanding the tension between human agency and machine automation can help researchers design systems that successfully weave AI support into a workflow without sacrificing a user's agency and creativity without sacrificing decision-making accuracy [33, 89].

In addition, researchers raised concerns about trust and reliance when they used generated text[2, 40, 64, 92]. Researchers pointed out automation of creative work resulting in "cannibalizing" the work of creative artists engaged in script writing [64, 92]. Previous studies also reported that LLMs can exhibit stereotypes or biases on text generation tasks [64, 92]. When researchers apply LLMs to specific use scenarios, they need to further investigate the potential risks that can bring back to the users.

To summarize, meta-reviewers face several challenges: taking time to understand complex reviews, cross-comparing and making fair decisions, and authoring comprehensive and well-structured meta-reviews. AI scaffolding can potentially help meta-reviewers conduct meta-reviews efficiently. However, we also foresee the trade-offs of the efficiency and agency of using AI to facilitate this process. As far as we know, there is no existing interactive system to support the

meta-review process. More generally, the research community still lacks empirical data on how users should interact with machine intelligence and what are the potential risks.

3 METAWRITER SYSTEM

3.1 System design goals

From the meta-review guideline provided by ICLR, meta-reviewers need to "make a reasonable recommendation based on reasonable bases and to clearly and thoroughly convey this recommendation and reasoning behind it to the authors." Meta-reviewers usually perform the following key cognitive activities: 1) read and understand the arguments within long individual reviews, 2) identify points of agreement and conflict across peer reviewers, and 3) generate an initial editable draft that summarizes the independent reviews [80].

Drawing on prior work that grapples with issues of scaffolding and expertise from the learning sciences, as well as agency and bias from human-AI studies, we specified three primary design goals (DGs) for the MetaWriter prototype that support the entire meta-reviewing process:

- DG1: Scaffold inexperienced meta-reviewers by embedding expert knowledge structures
- DG2: Preserve agency and reduce over-reliance through deliberative edits
- DG3: Reduce potential bias through staging and masking:

3.1.1 ***DG1: Scaffold inexperience meta-reviewers.*** Previous research shows that effective scaffolding can help novices perform work on par with experts [38, 99] For example, Yuan et al. found that providing editable critique statements helps novices provide feedback that is rated nearly as valuable as expert feedback [99]. In the context of writing introductory help requests, Hui et al. show that providing high-quality examples and expert-informed checklists can increase learning and writing quality [38, 39]. We build on these insights in MetaWriter by visualizing reviewer content, through highlight tags, related to key considerations or rubric items that experienced reviewers have agreed are important to the academic community. MetaWriter also provides an example to convey wisdom about the potential structure for a meta-review. However, expanding on prior work, the system not only provides an example from some other context (which requires users to conceptually "transfer" insights into the current task [11, 25]), it generates a highly contextual example that uses language from the current context, potentially reducing the burden of adopting the expert structures.

3.1.2 ***DG2: Preserve agency and reduce over-reliance through deliberative edits.*** Prior work has raised concerns about the potential loss of agency or control when working with AI on complex tasks [28, 83]. Shneiderman and others argued for more direct manipulation interfaces to give users more control of information and decision [83]. Recent work in the context of AI writing support often gives users an opportunity to perform post-edits on the machine-generated output in order to improve quality [30, 65]. Towards more controllable AI support for writing, prior research has explored techniques for modifying the "chains" of prompts for LLMs, along with the intermediate results, to improve model transparency and controllability [95]. Cheng et al. offered a framework to outline the myriad of ways that users could interact with AI text generation, including guiding or rating the model decisions, post-editing the text output, and writing with real-time assistance (eg. autocomplete) [16]. That study evaluated the writer's perceptions of the post-editing in the text summarization task and found that users were satisfied with the editing control granted by the interface but they would like to see more information on how the AI generated the summary [16]. In MetaWriter, to preserve agency and provide a layer of user control, the system requires users to either compose their meta-review from scratch or copy and paste from

the extracted sentences and/or the generated draft meta-review. We explicitly chose *NOT* to give users a generated draft within an editable text area as it could lead lazy meta-reviewers to overly rely on the AI text output rather than critically reflect on the contents.

Researchers pointed out automation of creative work may result in "cannibalizing" the creativity of artists engaged in script writing tasks[64, 92]. To reduce the over-reliance on the draft, we pop up a text box to remind people that they cannot directly copy and paste if their submitted draft is very similar to the provided draft. To evaluate whether the system will influence participants' decision-making process, we selected a scenario where the reviews have high variability – a rejected draft that received two borderline accept, and one borderline reject, for participants to make decisions and write meta-reviews.

3.1.3 ***DG3: Reduce potential bias through staging and masking.*** Previous studies raise concerns that LLMs could exhibit biases in human-AI collaboration tasks [13, 34, 48, 72, 86, 91]. For example, a previous study on child protective services AI decision support showed that AI could be biased towards certain groups [48, 86]. Previous studies also found that providing information, such as explanations, generated by AI can potentially mislead users in decision making [13, 29, 52]. Users need a staging process that helps them reflect on the content to make unbiased decisions. Masking the uncertain content can also facilitate the thinking process. Correspondingly, to reduce the potential for biased decisions in MetaWriter, the system stages the workflow such that users are encouraged to first consume the independent reviews and make decisions before they see the draft meta-review; users must explicitly toggle a button to read the draft. Furthermore, when the user views the draft meta-review, we chose to cut out all sentences from the machine-generated meta-review draft that could give any indication of an accept-reject decision that might bias the user's decision. Finally, to encourage users to reflect on their own decision, MetaWriter pops up a notification if the user's submitted draft is too similar to the machine-generated draft.

We designed the MetaWriter system to seamlessly integrate machine intelligence into an authoring environment specifically for writing meta-reviews. Guided by these design goals, we created the MetaWriter system that enhances text authoring using three ML techniques: an automated tagging model that color codes common aspects in each review (e.g., originality of the paper), an extractive summarization model that identifies important sentences from each review, and a natural language generation model that provides an initial draft based on all three reviews, the paper title and the paper abstract only (but not the actual content of the paper) (see Fig 1).

3.2 ICLR conference dataset

To train the ML models that can automatically identify key points 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 machine learning conferences. We collected each submission's data using the OpenReview API and scraped reviews from all publicly accessible paper submissions from the year 2018 to 2022.[2] The submissions from earlier years were not collected because their meta-reviews were not released.

For each submission, we collected paper information including the title, keywords and abstract. We further collected all official reviews with reviewer ratings and confidence scores as well as the final meta-review with ratings and the final decision. Table 1 shows the descriptive statistics for the data collected each year. After filtering out submissions with fewer than 3 reviews and meta-review with less than 20 words, we retained 9803 submissions along with their corresponding

[1] https://openreview.net/

[2] The ICLR dataset includes 9803 submissions (paper abstracts, peer reviews, and final meta-reviews) https://openreview.net/group?id=ICLR.cc.

meta-reviews and 34,219 independent reviews. To simplify the meta-review process in the study, we only collected the original independent reviews and dropped the discussion comments during the rebuttal process. While prior research has examined the contents of the same ICLR submissions and independent peer reviews [47, 100], our dataset focuses on the meta-review passage for each submission.

Table 1 shows the average length of each individual review (508.9 words) and each meta-review (146.5 words). We observe that the average length of reviews and meta-reviews increase each year. That means meta-reviewers need to read and analyze more than 1,600 words from three reviews as well as the original submission paper, to make the decision. After conducting sentiment analysis using ROBERTA sentiment classifier from huggingface [3] on the dataset, we found that most reviews are framed in a negative sentiment [32]. In the independent reviews, 81.7% of sentences are negative, while 70.9 % of sentences in the meta-review exhibited negative sentiment.

year	submissions	accepted	rejected	review length (words)	meta-review length (words)
2018	910	335	575	431.1±273.5	100.0±69.0
2019	1419	502	917	470.4±309.0	139.1±109.7
2020	2213	687	1526	478.4±316.9	126.2±92.9
2021	2616	860	1756	567.8±360.3	180.0±144.0
2022	2645	1094	1551	596.8±346.7	187.0±154.8
Total	9,803	3,478	6,325	508.9	146.5

Table 1. Descriptive statistics of the ICLR peer review dataset from 2018-2022. Reviews and meta-reviews become longer each year. Total review length and meta-review length is an average where the lengths from each year are weighted equally.

3.3 Algorithms pipeline and evaluation

We used the constructed dataset to build the algorithm pipeline for MetaWriter. We first run a fine-tuned tagger on the dataset to color code each word according to their topic [100]. In the current dataset, 30.4% of words are tagged with an aspect, such as originality, clarity, etc, as shown in Fig 2 A. MetaWriter color coded these tags on each independent review to support reading. Next, we train an extractive summarization model that can identify important sentences automatically from each review [61], as shown in the Fig 2 B. Last, we combined all three reviewers' extracted key sentences, together with all reviewers' ratings and the paper submission's abstract, as input data to train a deep learning based natural language generation model that can automatically generate a draft meta-review [56], as shown in the Fig 2 C. MetaWriter provided the automatically generated draft as a start point for meta-reviewers to edit. The algorithm details are described below.

3.3.1 Tags that highlight multi-aspects in each reviewer's review. A good review not only contains a good summary of the paper but also consists of accurate and fair comments from multiple aspects to evaluate the paper's quality. As shown in Figure 2 - A, to highlight aspects raised by independent reviewers, MetaWriter directly adopts a fine-tuned tagger that is trained on a similar conference peer review dataset to color code reviews [100]. This tagger uses a pre-trained model BERT [21] and a multi-layer perceptron to classify the aspect of each token in the independent reviews. This tagger automatically annotates the following aspects of each review: summary, motivation, originality, soundness, substance, replicability, clarity and comparison [47, 90, 100]. Note that the original model predicts the sentiment of the tag as well but we opted not to include the sentiment

[3]https://huggingface.co/siebert/sentiment-roberta-large-english

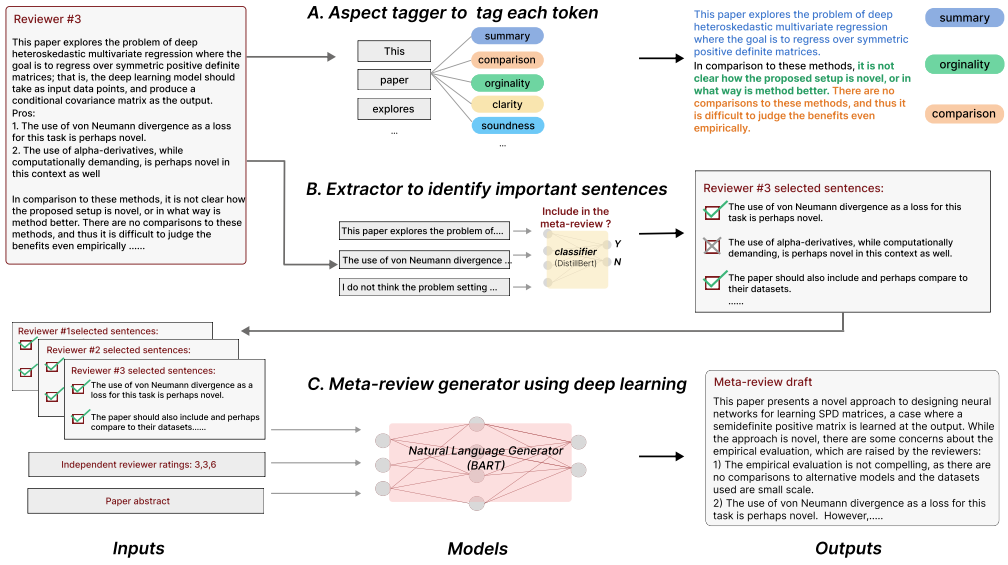


Fig. 2. Architecture of three machine-generated features. Left is the inputs for each model, middle is the three models, and right is the corresponding outputs, which are tags, extracted sentences and generated meta-review draft.

in MetaWriter as this could lead to reduced authorial agency amongst users, which is further investigated in 5.2.1. As reported in the previous study, the aspect tagger's precision score is 92.75% and recall is 85.19% [100] indicating relatively high accuracy. Among all independent reviews in our ICLR dataset, 30.4% of words received an aspect tag and the rest of words does not reach the probability of belonging to any of the tag. The research team download the review id from the training dataset of the tagging model to filter out them in the MetaWriter training data. Table 2 shows examples of each aspect tag and reports the percent distribution across the entire dataset. Among all tagged aspects, a summary of the paper has the highest frequency, which might be because every reviewer usually starts their review with a summary of the paper. Figure 1(a) shows how the UI for MetaWriter allows users to toggle on and off the six aspect tags which highlight passages in the original peer reviews.

Tagged aspect	Example	Freq (%)
Summary	This paper presents a neutral network model for machine translation using ...s...	53.4%
Motivation	The motivation of using the conditional prior is unclear.	5.1%
Originality	This paper presents a novel approach to cross-lingual language model learning.	7.3%
Soundness	This assumption is not true in practice	10.3%
Substance	The experiments are well-conducted.	7.1%
Replicability	The authors should provide more details about the hyperparameters.	2.0%
Comparison	The author should compare with [1,2,3] and [4].	4.5%
Clarity	The paper is well-written and easy to follow.	10.3%

Table 2. Tagged aspect category, examples sentences appear in the independent reviews and frequency. Among all tagged aspects, summary appears most frequently.

3.3.2 Extractive summarization algorithm to select important sentences. Extractive summarization techniques have shown evidence of being able to select strong candidate sentences for a summary [61]. Here, we utilized the extractive summarization technique and fine-tuned a pre-trained model to select key sentences from each independent review in order to shorten participants' time on reading the long reviews as shown in Figure 2 - B. To train the extractive summarization model, we first create an extractive summarization dataset from our ICLR dataset. The inputs are individual reviews and the labels are generated via a beam search procedure on the individual review [14, 61, 97]. The goal of this procedure is to label the sentences which were incorporated into the final meta-review. In particular, during beam search for each additional sentence we propose to add to the label, following [97] we compute a heuristic cost equal to the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) score of a given sentence with respect to the reference summary which is the meta-review written by ICLR meta-reviewers. ROUGE score is a widely used measure to automatically determine the quality of a summary by comparing it to original summaries written by human [58]. ROUGE-n scores count the number of overlapping n-grams and word sequences between the summary to be evaluated and the human written summary.

In this paper, we primarily use ROUGE scores to compare our fine-tuned models with that of previous works. Specifically, we use ROUGE-L as a measurement in this process as it considers sentence-level structures and finds similarities amongst sentences and n-grams via longest common subsequence statistics, making ROUGE-L ideal for complex, long-form content such as reviews [59, 82]. In this process, we iteratively loop over sentences from the review and only keep sentences when the ROUGE-L score between the selected sentences and the meta-review improves.

We fine-tuned the extractive summarization model PreSumm built on top of DistillBERT on our dataset of reviews and their beam-searched labels [61]. After following the PreSumm training setup and fine-tuning our dataset, we evaluated our extracted output against the real meta-reviews. We obtained the F1 scores of ROUGE-1 as 0.341, ROUGE-2 as 0.085 and ROUGE-L as 0.162. For each review, the model can automatically identify key sentences that have the highest probability of being in the meta-review. Figure 1(b) shows how the MetaWriter interface provides the participants a short summary to review.

3.3.3 Abstract summarization algorithm to generate a draft meta-review. Abstractive summarization techniques can generate a short and concise summary that captures the salient ideas of the source text. The generated summaries usually contain new phrases and sentences that may not appear in the source text [56]. Similarly, meta-reviews typically synthesize all reviews into one cohesive summary without directly copying sentences from each [82]. To ease the process of writing a meta-review, we generate a draft meta-review for participants to use as inspiration or a starting point. As shown in Figure 2 - C, to generate a realistic and natural draft, we used an abstractive summarization method that combines similar sentences to generate new sentences [9, 50, 82]. 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 used our dataset to fine-tune the bart-large-cnn model, one variant of the BART model [56]. More specifically, we use the PyTorch implementation in the open-source library Fairseq [67]. When we evaluate our generated meta-review against the real meta-reviews, we obtained the F1 scores of ROUGE-1 as 0.345, ROUGE-2 as 0.095 and ROUGE-L as 0.207. Our model performance based on ROUGE metrics is comparable with the controllable text generation method [82] and the model used in MetaGen [9]. More technical details are described in the Appendix. Figure 1(c) shows that MetaWriter provides a meta-review draft generated by the machine for participants to read and edit.

3.4 System implementation

Figure 1 shows the user interface (UI) for the MetaWriter system designed to guide the participant through performing a meta-review. After participants log in to the system and go through the tutorial, MetaWriter renders all three features on the interface: (1) color-coded tags that can be toggled on/off to highlight multiple aspects of reviews; (2) extracted key sentences that participants can hover over to locate the position within the review; (3) an automatically generated draft meta-review that users can copy, paste, and edit as their own initial draft if they wish. Participants can interact with these three features while forming a decision and delivering the final meta-review.

As shown in Figure 1, MetaWriter simulates the OpenReview platform which provides the paper abstract, a link to the paper pdf, keywords, and the three original independent reviews. Figure 1 (a) shows the output of the aspect tagger algorithm as a set of six categories along with the frequency that they appear in the reviews. When users toggle the checkbox before the category, the categories will get underlined with the corresponding color. Figure 1 (b) lists all sentences extracted by MetaWriter for the independent review. When participants hover over an extracted sentence, the corresponding sentence will be highlighted in red to help participants locate the key point in the original review. In Figure 1 (b), when the user hovers on the sentence "The user of vonNeumann divergence as a loss for this task is perhaps novel", MetaWriter highlighted the corresponding sentence in red in area d. Figure 1 (c) provides the generated meta-review draft based on the abstractive summarization technique described in Section 3.3.3. Users can select to hide the meta-review if they think the draft might influence their judgment on the paper. Instead of predicting the paper decision, we asked participants to make their own choice and we remove the sentences that indicated the decision in the meta-review draft. In the upper right-hand corner, participants need to make a decision and then write the final meta-review. We also provided instruction and guidance for both conditions.

In terms of implementation details, the client-side UI was developed using React and Typescript. We used a Firebase server to set up user accounts and store data. Offline, we run our trained language models to generate tags and drafts which are stored on our server and served to users. The client-side program is responsible for rendering the user interface and monitoring user actions on the webpage. We style the components and overall theme to be similar to that of the OpenReview platform. Whenever the user toggles the highlight tags, hovers over the extracted sentences, or performs edits on the text box, the UI will send the content and time stamps to the server-side program through HTTP requests. The Firebase server stores all reviews and metadata, including tags, extracted sentences and drafts that are generated from the Python server.

4 METHOD

We designed a mixed-method study, which includes a within-subject study that invited 32 participants to play the role of meta-reviews to use the MetaWriter system, and an interview study with six paper authors. In the within-subject study, we aim to understand users' reactions to the MetaWriter system and to measure performance differences in terms of quality and time spent. Peer review is a complex process that involves multiple stakeholders, including paper authors, reviewers, and meta-reviewers [76, 80]. To understand perspectives from multiple stakeholders, we conduct interviews with six paper authors to gain their perceptions on using machine intelligence in the peer review process.

This mixed-method study focused on four key research questions:

- RQ1: How do the machine-generated enhancements in MetaWriter affect the overall quality of meta-reviews and time-on-task compared to a baseline editor?
- RQ2: How do participants perceive the overall value of MetaWriter compared to a baseline editor?

- RQ3: How does the meta-reviewing experience affect the performance and perspectives?
- RQ4: How do paper authors react to the meta-review assisted by the MetaWriter?

4.1 Within-subjects Experiment

To evaluate MetaWriter compared to a standard meta-review editor, we conducted a with-in-subjects experiment with 32 participants. We simulated a real peer review process where participants played the role of a meta-reviewer with the task of summarizing independent reviews and writing meta-reviews. Each participant completed two meta-reviews: one with the MetaWriter system that includes the machine-generated highlights and drafts (MetaWriter condition) and one with a plain text editor (Baseline condition). We counterbalanced the order of each condition through random assignment.

4.1.1 Within-subjects Experiment Participants. Before the study, we sent out a pre-study survey to participants to collect information about their expertise, research topics, review experience, ML knowledge level, and demographics (e.g. age, gender, race). We recruited 32 participants who have experience in ML as well as with peer reviewing for ML conferences. We advertised recruitment messages to colleagues, mailing lists, communication channels, and social media and recruit participants from six universities and two companies across the US. Using a snowball sampling approach, we asked participants to refer their friends and colleagues. All participants reported that they have experience in reviewing ML papers and six had been a meta-reviewer or an associate chair for ML conferences. Participants have on average 4.7 years of writing and submitting machine learning papers (with a standard deviation of 2.8). Participants are mainly senior Ph.D. candidates, postdocs, professors, or professionals who worked as researchers and had fruitful experiences with conference peer review.

The two research papers that are selected in the study include the keywords as "Neural Networks, Machine Learning, and Reinforcement Learning". In order to protect the author's anonymity in the author interview, we decide not to reveal the title of those selected papers. To ensure participants had enough expertise to understand the paper's content, we only recruited participants who covered the keyword including "machine learning theory, neural networks, and reinforcement learning" and filtered out participants whose research topic did not cover any of the keywords, such as keywords that are only about NLP or Computer Vision. All participants were compensated for $20 per hour and five participants who wrote the highest quality meta-reviews rated by experts received a $20 bonus.

4.1.2 Within-subjects Experiment Procedure. Before the user study, participants filled out a pre-survey that captured their previous experience in reviewing and meta-reviewing papers on ML conferences and their expertise in ML to join the study through an online teleconferencing tool. We asked participants whether they had read the paper before to make sure they all were seeing the work for the first time. In the beginning, participants read and virtually signed our IRB-approved consent form. Then they conducted meta-reviews on two papers using different versions of our interface (MetaWriter or Baseline). To counterbalance the order effect, we randomized the order of the control condition and the experiment condition for each participant, so some participants encountered MetaWriter in their first session, and some others experienced it in their second session. For each session, we followed the meta-review process used at the ICLR conference, where we provided the paper draft and all three individual reviews with their ratings and confidence scores.

For both sessions, participants were told to spend around 30 minutes – and no less than 10 minutes – on the meta-review, but they could take as much time as needed. In the pilot study with 4 participants who had reviewing experience prior to the study, we found that participants can finish

writing meta-reviewing within 30 minutes. Before each condition session, we gave the participants a quick 2-minute demo of the interface, while in the experiment session demo, we also introduced the functionality of MetaWriter. After two sessions, the research team asked the participants to fill out a post-survey to evaluate the system, such as their satisfaction with each feature and their trust in the system. The research team then conducted a 15-minute semi-structured post-interview to ask open-ended questions about their experience, perceptions, and feedback. The post-interviews were video recorded with participants' permission and were transcribed into text for later analysis.

4.1.3 Submissions Selection Process. As the primary materials for the study, we selected two comparable submissions from ICLR 2020 that had three independent reviews of similar length and ratings. We chose two papers using several criteria to ensure that the two papers are similar for meta-reviewers to review. First, they both had borderline scores – which have 2 reviews selected weak accept and 1 review selected weak reject (both between the 20 to 30 percentile of the peer review scores) – and ultimately got rejected from ICLR. As described in Section 2.1, one challenging scenario that meta-reviewers face is the presence of disagreements, conflicts and variabilities in the independent reviews' evaluation. Here our goal is to simulate scenarios where the meta-reviewer would need to deal with reviewer conflicts and make a tough decision. Second, the length of individual reviews (average of 337.3 words) between the two papers is similar and they both fall into the middle range of the individual review length across the entire ICLR 2020 reviews (between the 40 and 60 percentile of the distribution). This ensures that participants spend a similar amount of time reading independent reviews that have similar lengths. Third, the generated meta-reviews for two papers need to have similar lengths (1037 words and 1301 words), to control the time that participants may spend on reading the generated meta-review draft. Fourth, the generated draft has similar high ROUGE_L scores. Combining all criteria above, we selected two comparable papers that were used for the within-subject experiment.

In the simulated scenarios, we only provided the original review and did not include all the discussions that happened during the rebuttal phase. In the ICLR peer review process, after the original review is delivered to the author, there is a lengthy discussion panel that takes place on OpenReview that meta-reviewers should take into consideration. In addition, a previous study analyzed reviewers' decisions before and after the rebuttal phase and found that a reviewer's final score is largely determined by their initial score and the distance to the other reviewers' initial scores [24]. Hence, we decided to control the study length and reduced the complexity by only presenting the original reviews to participants.

4.1.4 Within-subjects Experiment Measures. We collected a mix of quantitative and qualitative data, including each participant's log data that captured their interactive behaviors with the system, the final meta-review artifacts for each paper (N=64) that recorded the final meta-review, the post-survey (N=32) and the interview transcripts (N=32). The research team analyzed these combined sources of data to reveal insights towards our research questions.

Quality of final meta-review. To evaluate MetaWriter's effect on quality, we collected the final meta-review artifacts and each participant's decision on whether to accept or reject the paper. We calculated the alignment between the participants' decisions (accept or reject the paper) and the real final decision (reject the paper) on the two papers in both conditions. We measure the accuracy of their meta-review decision and observe whether MetaWrite will influence users' judgment on the paper decision.

To measure the quality of the meta-review, we asked two experts who have more than 5 years of machine learning research background and have experience being a meta-reviewer for ICLR

conference to rate the quality of all final meta-review (N=64) with a 3-dimension rubric based on meta-reviewer guideline for ICLR conference [4] and previous research [90, 100]:

- Summative (summarizes key points from the submission): A good meta-review captured the paper submission's main content, including scientific claims, and points out what is missing from the submission.
- Coverage (covers the reviews): A good meta-review should concisely summarize and cover independent reviewers' opinions.
- Justified (rationalizes the decision): A good meta-review should be clear about and provide specific reasons for its decision.

Two experts rated these 3 dimensions on a simple three-point scale (0-2). Because the reliability and validity are independent of the number of scale points used for Likert-type items, the three-point scale is adequate [42]. The simple three-point scale is also used in recent HCI studies to measure levels of degree [79, 85]. The research team first provided five examples and provide instructions for them to rate and discuss until they reached a consensus on ratings. Then they rate each dimension of the meta-review independently. The final rating of each dimension of the meta-reviews is the average of two experts' ratings.

We measured the lexical diversity of the written meta-review in each condition by calculating the BLEU score for meta-reviews in each condition. We used the average BLEU score to be the Self-BLEU of the meta-review [103]. We also measured the similarity between the drafted meta-review with the original meta-review provided by the real ICLR conference reviewers using cosine similarity. To calculate the semantic similarity, we first used BERT encoding (length = 512) to encode each review and then calculated the cosine similarity between the written review and the original review [21]. This approach can potentially reflect the semantic distance between two documents. In addition, we measured the length of each meta-review written by participants in both conditions.

User interaction data. To measure participants' interaction with the tool, we instrumented the interface in order to collect a range of user activity logs data. We collected two timing measures – how long each participant took to finish the review session and how long each participant spent on editing the meta-review within the textbox. We collected how much time they spent on authoring the draft from the time stamp they started editing the meta-review in the text box as well as the entire session time, including time spent on reading and writing.

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

User preferences. To evaluate users' preferences for the MetaWriter experience compared to a plain meta-review Editor, we asked participants to fill out a short post-study survey. The survey asked participants to directly compare the perceived usefulness, enjoyment, easiness, and sense of control between the MetaWriter and baseline system. Then we specially asked participants to evaluate each of the machine-generated features. The survey collected their 5-point Likert scale ratings on perceived usefulness, perceived accuracy, and trust for each feature.

User reactions. After the post-survey, we conducted a 15-minute semi-structured interview with all participants to capture their overall thoughts as well as specific perceptions of machine-generated highlights and summaries. For example, the research team asked "What do you think of

[4] https://iclr.cc/Conferences/2021/MetareviewGuide

the difference between the task with and without the support of MetaWriter", and "Which function did you use the most? Which one did you like the most?", and "What concerns did you have when using MetaWriter?".

Control variables - Users' knowledge on each paper's topic. After the post-survey, participants described their knowledge of each paper's topic from not familiar as 1 to very familiar as 4. We found that 11 participants are more familiar with one paper A than paper B, while 5 participants conducted paper A in the MetaWriter condition and 6 participants conducted paper A in the controlled condition. Participants' average familiarity with each paper topic is 2.1 with a standard deviation of 1.4. All participants reported that they had never read or remembered the papers before.

4.1.5 Data Analysis.

Quantitative data anlaysis. To measure the effect of the MetaWriter system on each dimension of the meta-review quality (eg. summative, coverage, justified), we conducted repeated measure ANCOVA tests. We used paper id, the order of experiment conditions (whether MetaWriter was used for the participant's first or second paper), the knowledge level of each paper topic, and the length of each meta-review as co-variants.

To measure the effects of the experiment condition on the time they spent reading independent reviews and writing meta-reviews taking the consideration of the differences between the paper topic and order effects, we ran a repeated measure ANCOVA using the paper id, the order of experiment condition and the knowledge level of each paper topic as co-variants.

Qualitative data anlaysis. All semi-structured interviews with participants are recorded and transcribed. Two researchers conducted iterative open coding on the transcripts using Dovetail[5] following the grounded theory approach [19]. They open-coded the data by identifying topics mentioned by the participants. Initial codes were combined into preliminary themes, which were discussed among the research team. Finally, after iteratively discussing the code themes, researchers derived the final themes around: participants' reactions to each feature, their overall perceptions of the MetaWriter system, and their concerns about using the system.

4.2 Interview Study

4.2.1 Interview Study Participants. To evaluate the author's reactions to the meta-write written by the MetaWriter, we reached out to six ICLR paper authors, including one author who is the original ICLR paper for the paper that got used in the user study. All paper authors have experience in submitting ICLR papers and obtaining reviews and meta-reviews.

4.2.2 Interview Study Procedure. We conducted 30 minutes interviews with the paper authors. We began the interview by demonstrating the MetaWriter interface and showing three meta-reviews examples written by participants in the within-subjects experiment. We first asked the question about their expectation of high-quality meta-reviews. To understand their concern about using the MetaWriter tool, we asked them "how much would you concern about the meta-review if you are aware that meta-reviewers are using this MetaWriter tool to write a meta-review for your own paper?". Then we asked them to elaborate on what kind of concern they have. We then asked them to cross-compare three machine-generated features to see which feature they had the most concern on. Last, we asked them about general opinions on using AI in the peer review process.

[5]https://dovetailapp.com/

4.2.3 Interview Data Analysis. All semi-structured interviews with paper authors are recorded and transcribed. Two researchers conducted iterative open coding on the transcripts using Dovetail following the grounded theory approach [19]. They open-coded the data by identifying topics mentioned by the paper authors and then combined initial codes into themes, which were discussed among the research team. Finally, after iteratively discussing the code themes, researchers derived the final themes around: the authors' concerns about using AI to support the meta-review process, the authors' perceptions of each feature, and their general opinions of using AI to support the conference peer review writing.

5 RESULTS

We report our results from the within-subjects experiment and the interview study. In the within-subject experiments, across both conditions, participants spent an average of 18.9 minutes writing a meta-review with an average length of 164 words written. 79.7 % of participants chose rejection decisions, which is consistent with the original decisions for the two papers. Our findings suggested that MetaWrite can make the meta-review process more efficient, enjoyable, and less cognitively demanding. However, participants surfaced their potential concerns about the MetaWriter. In the interview study, the paper authors who received the meta-review expressed their strong concerns about AI usage in meta-review writing. **For the ease of distinguishing different participants in two studies, we use "paper authors" to represent the ICLR authors we invited in the second interview study, while "participants" means the people we invited in the within-subject experiments.**

5.1 RQ1: MetaWriter helps participants write more informative meta-reviews in a shorter time

5.1.1 MetaWriter helps participants write more informative meta-reviews. To assess the quality of meta-reviews, we invited two experts to judge to what extent each final meta-review is summative, coverage, and well justified on a simple three-point scale (0-2). In the ANOVA test on each dimension of the meta-review in two conditions, we found that meta-reviews written with MetaWriter contain significantly more detailed summaries of the paper contribution and have significantly more coverage on reviewer perspectives compared with the meta-reviews written in the baseline editor. This indicates that MetaWriter can help participants write more informative meta-reviews.

	Baseline	MetaWriter	p	F
summative ratings (0-2)	1.49 (0.45)	1.90 (0.18)	0.02 *	5.01
coverage ratings (0-2)	1.23 (0.50)	1.64 (0.39)	0.03 *	4.01
justified ratings (0-2)	1.08 (0.49)	1.52 (0.39)	0.03 *	3.81
length of meta-review (words)	146.0 (59.7)	182.0 (65.3)	0.02 *	4.95
similarity across participant drafts (self-BLEU)	0.13	0.22	-	-
cosine similarity with the original meta-review	0.16	0.39	***	18.63

Table 3. Characteristics of meta-review written by participants in two conditions. Average length and ratings are shown, along with the standard deviations. ANCOVA test showed that the participants wrote significantly longer and more comprehensive meta-reviews with MetaWriter. p-value significance codes: 0.001 * * *, 0.01 **, 0.05 *

To compare the length of meta-reviews written in both conditions, we performed repeated measures ANCOVA to examine the effect of the two conditions on the length of the meta-review and control the two papers, the order of experiment conditions, and the knowledge level of each

paper topic as co-variates. As shown in Table 3, we found that participants wrote statistically significantly longer meta-reviews with the MetaWriter system. We found no significant interaction effect between the order of the two paper meta-review tasks and the conditions, no statistically significant differences between the two papers, and no statistically significant differences between the knowledge of the two papers. Because we randomly assigned participants to one of the groups.

We further calculated the lexical similarity across participants' drafts using self-BLEU scores in each condition to compare the diversity of meta-reviews written in each condition [103]. Higher self-BLEU indicates that less diversity between written meta-reviews in the group. As shown in Table 3, meta-reviewers in the baseline editor condition wrote less similar and more diverse meta-reviews, but no significant test is conducted given that we only have aggregated data in each condition group. This reflects that providing a template for draft meta-review can potentially lead to homogeneous meta-reviews and more uniformity.

We conducted an ANCOVA test to examine the effect of the two conditions on the cosine similarity between the written meta-review and the original ICLR meta-review and control the two papers and the order of the experiment conditions. As shown in Table 3, we found the cosine similarity between the draft and the original ICLR metareview in MetaWriter is significantly higher than the baseline condition. Hence, with MetaWriter, participants can author a more similar meta-review to the original ICLR meta-reviewers. On the contrary, this leads to more uniformity and may limit the diversity and creativity of meta-reviewers who used MetaWriter

5.1.2 MetaWriter expedited the meta-review reviewing process. Our experiment showed that MetaWriter reduced meta-review writing time and total meta-review completion time compared to the baseline editor. We performed repeated measures ANCOVA to examine the effect of the two papers and the two conditions (with vs without MetaWriter) on writing time and control the length of the meta-review as a co-variate. As shown in Table 4, participants spent significantly less time on writing meta-reviews when using the MetaWriter than the baseline condition ($p < .000$). In addition, they spent significantly less time on the total meta-reviewing process ($p < .05$). Moreover, there is no significant interaction effect between the order of the two seed papers and condition on writing time and total completion time. We distinguished the writing time and total time here because there is more text and content provided in the MetaWriter condition, which includes the extracted sentences and the generated draft. Participants may spend more time reading these extra texts, but these texts could potentially be helpful with writing the meta-review draft.

	Baseline	MetaWriter	p	F
total time (minutes)	20.05 (5.92)	17.76 (5.12)	.04 *	4.20
writing time (minutes)	10.92 (3.74)	7.93 (3.71)	.000 ***	13.10

Table 4. Time on tasks in both conditions. ANCOVA test showed that participants spent significantly less time on writing a meta-review in the MetaWriter condition. p-value significance codes: 0.001 ∗ ∗ ∗, 0.01 ∗∗, 0.05 ∗

As shown in Table 5, each feature is used by more than 80% of users (24 out of 32 participants), and the generated draft meta-review feature is used the most frequently (27 out of 32 participants). We observed that participants frequently interact with the extracted sentences (average frequency of 12.19) by hovering over these important sentences to seek the location in the individual reviews. Participants further described how the MetaWriter system made their meta-review process faster in the qualitative review, which is discussed in the sections below.

5.1.3 Comparison between experienced meta-reviewer versus inexperienced meta-reviewer. Participants consist of 6 experts who have more than one meta-review experience and inexperienced

	# of participants	Average frequency
Used highlighted tags	24	3.81
Copied or used the extracted sentences	24	12.19
Copied or used the generated draft meta-review	27	0.84

Table 5. Usage of three machine-generated features. Frequency is the number of times the feature was used before submitting a meta-review. Generated draft meta-review was used by most participants and participants frequently engaged with the extracted sentences.

meta-reviewers with no meta-review experience. The different background experiences of the participants provided us with more opportunities to explore the factor of expertise while using MetaWriter. Previous literature suggested that by providing enough scaffolding and support to novices, they can almost perform as well as experts [99]. To explore the potential novelty effects in the study, we separate participants into two groups: the experienced meta-reviewers group consists of 6 experienced meta-reviewers and the inexperienced meta-reviewers group consisted of 26 inexperienced meta-reviewers who only have review experience and no meta-review experience.

Participants	**Condition**	**time (mins)**		**quality**			**length**
		total	writing	summative	coverage	justified	
experienced	MetaWriter	17.0(7.0)	7.7(4.0)	2.0(0.0)	1.8(0.41)	1.8(0.45)	225.7 (63.0)
	Baseline	17.0 (9.1)	9.3 (4.2)	1.33(0.52))	1.5 (0.84)	1.33(0.52)	157.0 (44.6)
				*	-	*	.
inexperienced	MetaWriter	17.9(4.7)	8.0(3.7)	1.88 (0.35)	1.6 (0.50)	1.46(0.56)	172.0 (62.7)
	Baseline	20.8 (4.9)	11.3 (3.6)	1.53(0.21)	1.17(0.68)	1.02(0.67)	143.4 (61.8)
			*	*	**	*	.

Table 6. Means with std calculated for each group on length, time, and meta-review quality. Mann-Whitney U Test results of the effects of expertise and experiment condition on the review length, total time, writing time, and review quality including summative, coverage, and justified. Mann-Whitney U Test p-value significance codes: 0.001 * * *, 0.01 **, 0.05 *

Table 6 below compared each group's performance between MetaWriter and controlled condition. Given that the experienced meta-reviewer group had a very small size, we conducted the Mann-Whitney U Test instead of the ANOVA test, which is a non-parametric statistic rank test between two samples [63]. MetaWriter led both experienced and inexperienced participants to write significantly longer meta-reviews compared to the baseline system. Interestingly, in terms of Time on Task, we see that MetaWriter helps inexperienced participants perform the writing task more efficiently (and nearly as fast as the experienced meta-reviewers), while there was no significant change in writing time for more experienced participants. One reason might be that the experienced participants may already have familiarity with the basic structure for meta-reviews, allowing them to focus more on quality improvements. MetaWriter seemed to provide more support to inexperienced participants by scaffolding an initial structure which can be daunting for the first-timer.

In terms of the expert ratings on the quality of meta-reviews, the MetaWriter system helped experienced and inexperienced participants significantly improve the overall paper summary and the justification of the decision. Only inexperienced participants saw significant improvement in their coverage of the points raised by the independent reviewers. Hence, MetaWriter seems helped inexperienced meta-reviewers more compared with experienced meta-reviewers.

We also compared the cosine similarity between each written meta-review and the original ICLR meta-review in both conditions. Through Mann Witney U tests, only for inexperienced meta-reviewers, meta-reviews written in MetaWriter(cosine similarity as 0.15) have significantly more similarity with the original ICLR meta-review compared with the baseline condition. One reason might be inexperienced meta-reviewers rely more on the provided draft to justify the arguments.

5.2 RQ2: Participants revealed a preference towards MetaWriter but shared concerns

5.2.1 Participants preferred the MetaWriter even though it provided less control. After participants had experienced the two systems, we asked participants to compare them overall directly. As shown in Figure 3, participants highly prefer MetaWriter over the baseline Editor. All participants indicated that they would like to use this interface in the future meta-review process, some of which reflected that the machine-enhanced interface could make the meta-review process "faster and more interesting". MetaWriter is perceived as not only easier to use but also more useful and enjoyable.

In the survey, 90.6% of participants believed it is faster to write the meta-review using the MetaWriter interface. During interviews, 12 participants explicitly mentioned that MetaWriter saved them time in meta-reviewing. Participants reflected that the generated meta-review contains details similar to their own meta-review writing format, saving them time on structuring and drafting the meta-review. However, participants also pointed out that they took some time to read the additional text extracted by the machine and then further verify the information provided by machine intelligence. For example, P1 mentioned "I don't know which one makes it faster. I still need to go back and forth to verify that information".

Results show that MetaWriter reduced the cognitive load in the meta-review process and made the meta-review process easier. Interestingly, we observe that this increase in efficiency seems to trade-off with the feelings of agency as only around 28.1% of participants said MetaWriter gave them more sense of control. Some participants also discussed these trade-offs between efficiency and the feeling of agency explicitly during interviews (n = 3): "... because the extracted sentences and draft meta-review provided to me are like steering me towards one particular way of writing your review. So there's a trade-off between flexibility and time-saving. It saved more time, but it means less flexibility. (P5)"

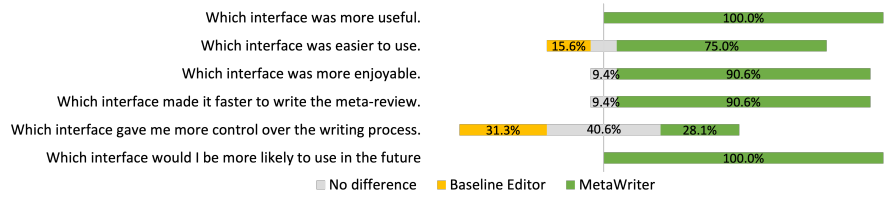


Fig. 3. Participants' comparison between MetaWriter and baseline Editor. Participants preferred MetaWriter overall.

5.2.2 Participants shared diverse perspectives on the relative value on three machine assistance approaches. We conducted an in-depth analysis of how participants used and perceived three types of machine support: highlighted tags, extracted sentences and generated meta-review.

In the post-survey, we asked users to compare each feature in terms of its usefulness, accuracy, and their trust in each feature. As shown in Figure 4, we found that most participants (78.1%) perceived highlighted tags as accurate and 68.7% of them trusted the tags. Interestingly, most participants

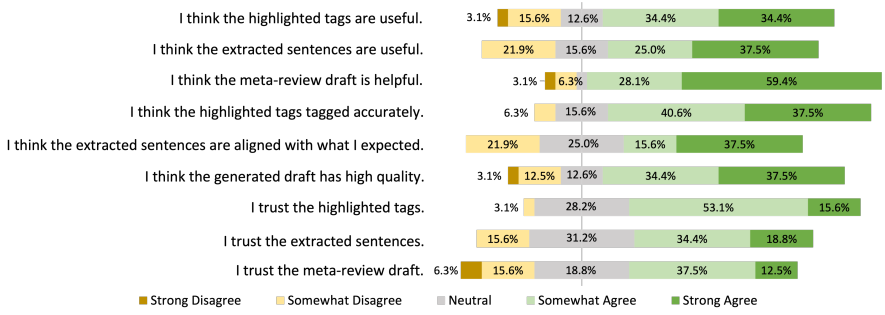


Fig. 4. User feedback on the machine-generated features. Among three features, most participants think that the generated draft is useful but only half of the participants trusted the machine-generated draft.

(87.5%) think that the generated draft is useful, but only 50.0% of them trusted the machine generated draft. In the post-interview, participants explained their concerns on the potential bias in the generated draft. Among the three features, users perceived that the extracted sentence feature has the lowest accuracy, whereas only 53.1% think they are accurate. From the qualitative interview data, we also noticed that participants used these three features slightly differently. We discuss how users reflect on their usage of three features below:

Highlighted tags guide reading and help cross-compare reviews. Participants used highlighted tags as visual guides or anchors that can draw their attention while reading reviews (N = 10). As P16 said " Tags can hint where I was reading. It actually helped me to locate the information that I need to focus on during the review". A similar function of tags is also revealed in the context of making sense of group chat. For example, Tilda system asked users to create tags to enrich the chat conversations [101]. Participants also used tags to cross-compare individual reviews and identify conflicts (N=6). P15 said that "I found that is very powerful when I want to cross-compare the different opinions among the reviewers"

However, participants mentioned the disadvantages of tags – when the review tags have too many categories, it is hard to distinguish the subtle differences between tags and potentially biased participants. For example, P3 mentioned that, "I don't have a clear distinction about the difference of the tags, like the validity versus soundness. I basically treat highlighted sentences as important sentences and read through them."

Extracted sentences ease the process of identifying major issues but may lose context. Participants used extracted sentences to identify major issues and locate reviewers' viewpoints within individual reviews (N = 9). P27 mentioned that "I used it to look back into the reviewers through the extracted summarization. There are some places I need to refer back and forth between the reviewers' comments and my writing to see which reviewers' comments I can potentially combine." They also reflected on how despite some reviews being long, the extracted sentences saved them time and helped them skim the reviews and make decisions (N=4).

However, participants pointed out that the extracted sentences might lack context and reasoning (N = 4). P14 explained that " sometimes, the extractive version only has one sentence and misses all the details. And usually, you want some details to make the meta-review more convincing from there". Participants also raised concerns about the accuracy of extracted sentences and mentioned that the extractor might miss some important points (N=5). For example, P3 said that "the obvious issue is the accuracy of the extractor. Some of the sentences might not be extracted and those

sentences are still very relevant, especially for making decisions on the paper." Some participants also perceived the extracted sentences feature as having high coverage but low precision. P26 mentioned that "Not every extracted sentence useful, but in terms of coverage, it does a good job. I would say the recall rate is pretty high, but the precision rate is not that high."

The auto-generated drafts help kick start the writing process. Participants reflected that the generated summary provided a good starting point from which they can add their own viewpoints and judgments. Seven participants explicitly mentioned that the meta-review draft is very similar to what they will write and the draft looks very natural – as P23 mentioned "I feel like most of the sentences look very natural to me, and some of them hit the point". A previous system Spark also raised a similar notion of initiation using generated text and prompts [28]. Different from the Spark system that shows prompts generated from the wild, the generated draft shown in our system contains the meta-review structure and important points based on the given independent reviews.

However, participants raised concerns about the bias in the generated draft (N = 11). They were worried that the tone of the generated draft might influence their judgment on the paper. "if it's not accurate, it might mislead the meta-reviewer or potentially put a bias there (P11)." In the experiment, we observed among some participants that when their judgment is not aligned with the tone of the meta-review draft, instead of following the draft, they changed the tone of the draft and authored a new version of the meta-review based on the existing draft. In the interview, P24 mentioned that "I fine-tuned the tone of the draft actually since I want to accept this paper. I tried to change it so that it sounds more positive for those issues pointed out by reviewers."

Some participants explained that they did not trust the generated meta-review since some points generated by the machine are not mentioned by any of the reviewers. As P8 mentioned, "I was not very happy about like generating something that is non-existent in the reviews, such as the experimental validation point did not appear in the original reviews". Similarly, P6 also elaborated that "highlighted tags and extracted sentences are purely based on the existing information so they cannot really produce too much harm, but no one knows what ML models can generate and where it come from". Participants also expressed concerns over the generated draft when there were conflicts in reviews and the draft was not coherent with some reviewers' viewpoints. P15 pointed out "I remember it says all reviewers, agreed with the paper that it's a little bit dense, but according to what I saw, I think only reviewer 3 mentioned about, you know, the paper is not really easy to follow and it's a little bit dense. This summarization might mislead ACs"

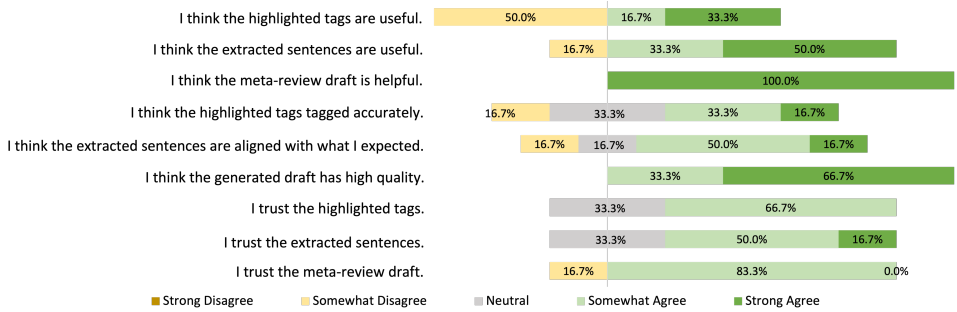
5.2.3 Participants expressed concerns about using MetaWriter to write the system. Participants who play the role of meta-reviewers elaborated on some concerns about the potential bias of the MetaWriter system in the interview. P9 mentioned that the class imbalance of tags across different individual reviews can influence the meta-reviewer's perception of the paper, "I saw the two reviewers(R1 and R2) have some tags related to originality but the other reviewer did not have any. Then I kind of paying more attention to them(R1 and R2) to see their judgment."

Participants also mentioned that the extracted sentences' accuracy needs to be improved (N = 5). There might be potential biases in them, one example being if the extracted sentences consist mostly of weaknesses and ignore the contributions of the paper. P4 explained that "when it is extracting pieces from text, it cannot like really understand or grasp what words are conveying. It will mislead users if it only selects some points that are more or less problematic."

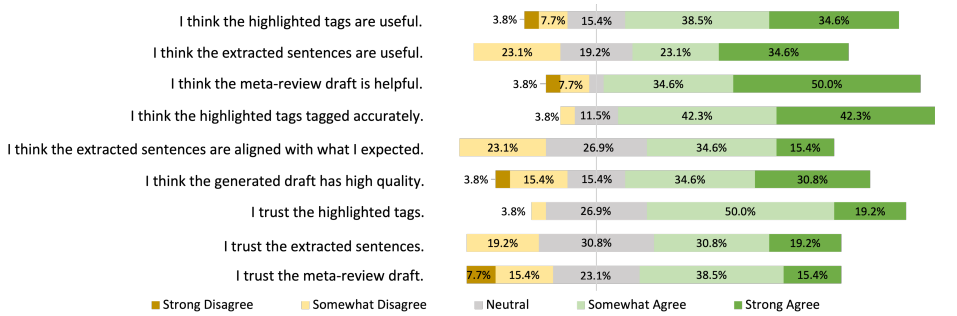
In the survey, we found that only 50.0% of participants trusted the machine-generated draft even though most participants perceived it as useful. Participants explained why the meta-review draft might mislead meta-reviewers in their decision-making process (N = 11). They worried that the generated draft contained inclinations towards certain paper decisions and provided more weaknesses in the draft. In the current MetaWriter, we removed the sentences that contain a

strong indication of the paper decision from the generated text (eg. I recommend rejection) and we provided users a toggle button to hide the draft meta-review. Instead, a better approach might be showing the draft meta-review after the user has read the individual reviews and made their decision. The system can then generate a decision-aware meta-review draft that is aligned with the user's decision.

Participants also have concerns that if the draft meta-review quality is high, it might make meta-reviewers over-reliant on the draft and potential for lackluster efforts of meta-review (N=7). P14 mentioned that "if meta-reviewers are lazy or short on time, they might just copy the generated one and won't check any technical details. This is not fair for the authors.."



(a) Experienced meta-reviewers' perceptions on each machine support feature



(b) Inexperienced meta-reviewers' perceptions on each machine support feature

Fig. 5. Comparison of each machine-supported feature by inexperienced meta-reviewers and experienced meta-reviewers.

5.2.4 Perceptions from inexperienced meta-reviewers and experienced meta-reviewers.

As shown in Figure 5a and 5b, more inexperienced meta-reviewers perceived the highlighted tags as useful than experienced meta-reviewers.For each survey question, we converted responses into a 5-point numerical scale and conducted Mann-Whitney U tests to compare the perceptions of experienced and inexperienced groups. Inexperienced meta-reviewers perceived the highlighted tags as slightly more useful (but not significantly more) than experienced meta-reviewers. All experienced meta-reviewers strongly agreed that the meta-review draft was useful, while only half of the inexperienced participants provided this rating (Mann-Whitney U test shows a significant

difference at p=0.03). Also, more experienced meta-reviewers trusted the meta-review draft than inexperienced meta-reviewers, but this trend does not result in a significant effect.

As shown in Figure 3 of all users' feedback on the system, users perceive control as a concern since only 28.1% of users think MetaWriter provided more control. Breaking down to detailed features, interestingly, while experienced meta-reviewers perceived the draft to be useful, they disagreed that MetaWriter provided them enough control over the meta-review authoring process. Interviewees mentioned that they still have control over what they can edit, but they are worried about other reviewers may lose control and fully rely on the draft. Experienced meta-reviewers highlighted that "The draft looks similar to what I usually wrote as a meta-reviewer and it contains a bullet list of cons, so I am satisfied with the quality [P26]". Slightly more experienced meta-reviewers disagree that the MetaWriter provided them enough control over the meta-review authoring process. P27 reflected that "I guess I still have control over what I can do and write there. I am just worried about other reviewers, like, if they see there is a draft, they may try to incline to make decisions that are more aligned with what the draft described... so some sort of less control here"

5.3 RQ3: Paper authors expressed concerns on using machine intelligence in the peer review process

All paper authors rated that they have concerns about using this system in the real peer review process and expressed their reasons (ratings bigger than 3 out of 5, where 5 represents extremely concerned and 1 equals not concerned at all). They are concerned that meta-reviewers may over-rely on the generated draft and not put enough effort into writing constructive feedback or justifying the reviews. P3 mentioned: "I feel it is ok to use the MetaWriter tool to help summarize the reviews, but I do need to see evidence that the meta-reviewer has put his/her own thoughts into it. I would be upset if my paper is rejected solely based on the output of MetaWriter." P1 also echoed that point and express concerns about the bias in the generated draft "there is a potential that meta-reviewers might be too busy to review the paper, then they may just use the generated draft". The concern of over-reliance is always raised in the context of AI-assisted decision-making [5, 13]. Interestingly, in this peer review context, over-reliance on AI potentially reduced paper authors' trust in the meta-review written by a human meta-reviewer.

In addition, paper authors asked for more transparency in using this tool in the peer review process. P2 reflected that more information about the process is needed –"whether the decision is made by the tool? Or only the meta-review is generated by the tool."

Further, they mentioned that meta-reviewers may not play a vital role in gatekeeping the quality of the reviews and verifying the correctness of the review if the reviewers are not experts in the corresponding area. The tool can ease the process but may reduce the opportunity for meta-reviewers to filter out low-quality reviews.

Among the three machine intelligence features, paper authors are most concerned about the generated meta-review draft. "There is a risk that the meta-reviewer copies directly from the generated review without looking at the paper/reviews. I want to see evidence that this is not the case"(P1, P6). Paper authors are less concerned about the highlighted tags and extracted sentences since they think these features can mainly enhance reading instead of authoring. Paper authors also suggested that this system can also be used for authors to write rebuttals or reflect on the reviews.

6 DISCUSSION

We conducted a mixed-method study that includes a within-subjects experiment and an interview study to explore the potential and perils of using AI in the meta-review process. In the within-subject experiment, we found that meta-reviewers not only prefer MetaWriter, but they also wrote longer

and more informative meta-reviews in a shorter amount of time than when using a baseline editor. MetaWriter improves the quality of written meta-review as rated by experts. Meta-reviews written in the MetaWriter condition are less diverse and more similar compared with the meta-reviews written in the Baseline Editor. This reflects that providing AI scaffolding may result in delivering homogeneous meta-reviews.

Both experienced and inexperienced meta-reviewers benefit from the scaffolding in terms of our quality measures. A previous study found that novices who don't have enough background knowledge improves learning through scaffolding [38]. In our study, we found that scaffolding not only improves junior meta-reviewers' performance but also helped experienced meta-reviewers, from different aspects. Scaffolding improves experienced meta-reviewers in summarizing content while it supports inexperienced meta-reviewers in identifying important arguments. We found that inexperienced meta-reviewers mostly benefit in terms of time but also benefit from getting a sense of meta-review structure. One open question this study raises is whether the adaptive example – the generated draft – can scaffold inexperienced meta-reviewers to perform better on the meta-reviewing task compared with showing an example from a prior meta-review of another paper. Future studies can explore the longitudinal effect of machine support in the peer-review context. One goal of scaffolding is that users can eventually be able to perform the task without the support of the tool [18]. Then, researchers can discern from the novelty effect versus the diminishing need for expert scaffolding as novices gained more expertise on the task.

From the user study, we observed everyone either modified or did not use the draft. The system did maintain some amount of agency, as described by the previous study on the summarization task [16], but participants reflected that they lacked some control, even though participants, especially experienced meta-reviewers, perceived the generated draft as being helpful. One participant(P6) indicated that the risks involved with machine supports can be a big factor: simply highlighting information in the original reviews does not change the underlying meaning, but generating entirely new blocks of text has the potential of infusing unintended meaning or leaving out key bits of information. The tension between efficiency and user agency is perceived differently by experienced meta-reviewers and inexperienced people. Experienced meta-reviewers have the confidence to control the machine intelligence and put more emphasis on efficiency, while inexperienced meta-reviewers preserved more concerns and trust the draft less. To provide participants with more control in the collaborative writing scenario, the tool can potentially update the generated artifacts according to users' input [16]. For example, participants expressed that they want the draft to update and the extracted sentences to be changed according to their decisions - accept or reject.

By rethinking the design goals, we further discuss the potential open questions around creating AI scaffolding and building trust in AI systems as a community:

6.1 Creating AI scaffolding that supports users but prevents uniformity

Prior work often considers human-AI systems have potential trade-offs between efficiency and creativity [16, 28, 64]. In the meta-review writing scenario, both experienced and inexperienced meta-reviewers benefit in terms of our quality measures. We found that inexperienced meta-reviewers mostly benefit in terms of time but also benefit from quality through learning on the structure and adaptive example. While we also saw efficiency gains for inexperienced meta-reviewers, it's inexperienced meta-reviews that are likely to get the most benefits from writing support. To produce more learning gains, the human-AI collaborative system should provide the right amount of scaffolding that can motivate learners to drive the process as well as provide more diverse and creative learning outcomes. In this way, the AI scaffolding can potentially promote the transfer of knowledge for learners or users. In the MetaWriter condition, we found that participants wrote more similar meta-reviews to each other compared to participants with no scaffolding. This could indicate that

the MetaWriter will lead to homogeneous work across papers, or it could be an indication that the users are more consistently following standards set by the research community which are scaffolded by the MetaWriter tool. Future research can potentially explore whether the similarity between participants is linked to better adherence to community practices and expectations of quality, or if it might be an indication that participants are just blindly following the writing scaffolding without thinking about whether/how it matches with community standards.

6.2 Building trust in AI systems for individuals versus trust in systems of people in the community using AI

Previous research showed that writing with the opinionated language model can affect participants' attitude on social topics [43] One potential explanation is that LLM suggestions may intervene with participants' thought processes and drive them to spend time evaluating the suggested content [8]. In our study, very few participants were concerned that their own decision or writing was being biased by the AI, instead they were more concerned about the AI influencing the other people's bias. By unpacking users' concerns, especially experienced meta-reviewers, we found that participants find value in this AI support and trust their own use of it, but that they might not feel comfortable with how the whole community appropriates the technology. Their concerns about machine intelligence partially come from the limits of machine intelligence, but they worry more about the over-reliance and the potential for bias to creep in when people think about how others might abuse the system. Hence, this brings open questions on how to build trust not only for individuals who are using the system but also help users gain confidence and trust in community usage.

6.3 Making AI systems transparent without adding cognitive complexity

Increasing transparency so that users can understand how and why the machine outputs certain information can potentially improve the users' trust in high automation[23, 68]. Prior research provides explanations for AI decisions or visualizes the origin of words/sentences in text suggestions [23, 78, 91], but this potentially bogs down the user trying to complete a task. For example, to explain how the meta-review draft is generated, MetaWriter needs to add more visualizations on the source of each sentence as well as explain how the model training works and show the model performance. Then this leads to extra work for meta-reviewers to read and understand. Existing research argues the importance of providing explanations and transparency for AI systems [23, 68], but the extra cognitive complexity brings the question of when is the extra work to understand the genesis of AI output worth it? A big open question is how and when to make AI reasoning transparent without getting in the way of users conducting the task.

6.4 Provide accountability and promote communication of LLM usage between multiple stakeholders

Peer review is a complex process that involves multiple stakeholders, including authors, reviewers, and meta-reviewers [80]. There is always tension between multiple stakeholders in this high-stake context. Previous studies highlighted the importance of keeping multiple stakeholders in the loop while designing AI systems [102]. One hard problem is how to design a system if there are conflicts between stakeholders. The within-subject experiments found that the generated draft feature in the MetaWriter system is the most useful and effective feature that meta-reviewers would like to use in the peer review process. However, during the interview with the paper authors, they had strong concerns about incorporating this feature in the real meta-reviewing process as they worried that meta-reviewers may over-rely on this feature and cannot provide fair and justified decisions. They asked for full transparency and more communications about how meta-reviews make decisions and write meta-reviews. One potential approach is to communicate through more

visualizations. For example, a future LLM support system for peer review could visualize the entire process while meta-reviewers interact with the interface and highlights the words that are inspired by LLM. Again, this brings back the discussion point of balancing the cognitive complexity and trust.

6.5 Towards human-machine hybrid collaboration for academic review

What role should machine intelligence play in this meta-review process? Our 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 they used machine-generated artifacts as inspiration in the process of conducting their meta-review. We consider their interaction with machine intelligence as one type of partnership relationship. In the recent discussion of human-AI collaboration, researchers proposed that AI and humans should maintain a "partnership relationship" where AI is designed to fit into the existing human task workflow and assist parts of tasks according to human needs [89]. Our study demonstrated different means to achieve human-AI collaboration in the meta-review writing process. This human-AI collaboration approach could help users to learn more from the AI suggestions and offload task to AI.

Machine intelligence 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 example, machine intelligence can potentially help reviewers to evaluate the paper and write a high-quality paper review [90, 100]. For paper authors, machine intelligence can help them analyze the reviews, write a persuasive rebuttal [24], and later edit the submission more effectively. In the future, we plan to extend our work to design more human-AI collaboration systems to support users in academic peer review. In addition, while the current study is conducted in a machine learning conference peer review context, we imagine that MetaWriter can be used for a broader range of academic communities.

7 LIMITATIONS

Our study has several limitations. First, in the study materials, we only provided the original review and not all the discussions that happened during the rebuttal phase. The reason is that we would like to control the study length and reduce the complexity. Notably, the reviewers of the two papers selected for this study did not change their ratings after the rebuttal. However, a rebuttal phrase is one important step in the peer review process and the discussion content can change the paper's final decision. Future studies can embed the discussion text into the study and train the model such that it takes the discussion structure into consideration [49].

Second, we created a mock scenario and encouraged users to spend about 30 minutes on each meta-review session which might be different than what meta-reviewers actually spent on writing meta-review. In a real meta-review scenario, from our interview, meta-reviewers take at least one hour to conduct meta-review. However, meta-reviewers spent large amount of time to read and understand the discussion and rebuttal. In our study, for simplicity, we only provide the original review and eliminate the discussion phase.

Third, we selected as core study material two borderline papers that got rejected. Our goal is to control the difficulty of making decisions and explore the tough situation in that meta-reviewers solve the conflicts. However, we did not explore the other situations, such as when all reviewers indicated a clear acceptance of the paper. When all reviewers clearly reject or accept a paper, meta-reviewers might perceive the value of machine support and the overall decision-making process differently. The limit size of scenarios may bring limitations in generalizing the results to other communities. In future studies, we could explore machine-generated features for all different types of review situations and different research areas.

Fourth, the machine learning algorithms we used here to tag the reviews, extract sentences and generate paragraphs can be improved. In the study, we used state-of-the-art models to create machine-enhanced features. Given the rapid development of ML and NLP, more powerful models such as GPT-3 [12] could be used and fine-tuned to our data. Moreover, for the meta-review draft feature, we used the ML model to automatically generate the draft offline before the experiment. Ideally, users can engage in a co-creation experience with the meta-review generation model where the MetaWriter can automatically generate the next sentence while the meta-reviewer writes the meta-review, similar to other co-creation systems on writing task in a human-in-the-loop manner [98]. Recent advances such as ChatGPT [1] present a co-creative experience in addition to training for human-alignment which could be integrated to improve the meta-reviewing system.

In addition, there is potential bias that comes from using models trained at a static fixed timestamp, which we trained using 5 years of data. These training data may potentially contain bias and as data becomes increasingly diverse, model performance might become worsen. Also reviewing culture is different for each venue and it can change largely, using fixed models could inhibit cultural or norm changes.

8 CONCLUSION

In this paper, we have designed a mixed-method experiment and built MetaWriter to support information synthesis in academic meta-review through machine-generated highlights and summaries. We empirically explored how users perceived the value of machine intelligence within an interactive tool for writing academic meta-reviews. A within-subjects experiment with 32 participants evaluated how MetaWriter affected the time and quality of the meta-reviewing process and how participants engaged with the three types of machine-generated features. We found that MetaWriter significantly shortened the time spent on meta-reviewing and helped participants write longer and more informative meta-reviews. However, our analysis reveals several risks of using LLMs in peer-review writing.

REFERENCES

[1] Open AI. 2022. *ChatGPT: Optimizing Language Models for Dialogue.* https://openai.com/blog/chatgpt/

[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.

[3] Thomas Anderson. 2009. Conference reviewing considered harmful. *ACM SIGOPS Operating Systems Review* 43, 2 (2009), 108–116.

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

[5] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

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

[7] Dominik Beese, Begüm Altunbaş, Görkem Güzeler, and Steffen Eger. 2022. Detecting Stance in Scientific Papers: Did we get more Negative Recently? http://arxiv.org/abs/2202.13610 arXiv:2202.13610 [cs].

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

[9] Chaitanya Bhatia, Tribikram Pradhan, and Sukomal Pal. 2020. Metagen: An academic meta-review generation system. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1653–1656.

[10] Lutz Bornmann, Rüdiger Mutz, and Hans-Dieter Daniel. 2010. A reliability-generalization study of journal peer reviews: A multilevel meta-analysis of inter-rater reliability and its determinants. *PloS one* 5, 12 (2010), e14331.

[11] John D Bransford and Daniel L Schwartz. 1999. Chapter 3: Rethinking transfer: A simple proposal with multiple implications. *Review of research in education* 24, 1 (1999), 61–100.

[12] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

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

[14] Jaime Carbonell and Jade Goldstein. 1998. 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*. 335–336.

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

[16] Ruijia Cheng, Alison Smith-Renner, Ke Zhang, Joel R Tetreault, and Alejandro Jaimes. 2022. Mapping the design space of human-ai interaction in text summarization. *arXiv preprint arXiv:2206.14863* (2022).

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

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

[19] Juliet M Corbin and Anselm Strauss. 1990. Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative sociology* 13, 1 (1990), 3–21.

[20] Dina Demner-Fushman, Wendy W Chapman, and Clement J McDonald. 2009. What can natural language processing do for clinical decision support? *Journal of biomedical informatics* 42, 5 (2009), 760–772.

[21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

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

[23] Jaimie Drozdal, Justin Weisz, Dakuo Wang, Gaurav Dass, Bingsheng Yao, Changruo Zhao, Michael Muller, Lin Ju, and Hui Su. 2020. Trust in AutoML: exploring information needs for establishing trust in automated machine learning systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 297–307.

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

[25] Dedre Gentner, Jeffrey Loewenstein, and Leigh Thompson. 2003. Learning and transfer: A general role for analogical encoding. *Journal of educational psychology* 95, 2 (2003), 393.

[26] Katy Ilonka Gero and Lydia B Chilton. 2019. How a Stylistic, Machine-Generated Thesaurus Impacts a Writer's Process. In *Proceedings of the 2019 on Creativity and Cognition*. 597–603.

[27] Katy Ilonka Gero and Lydia B Chilton. 2019. Metaphoria: An algorithmic companion for metaphor creation. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.

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

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

[30] Spence Green, Jeffrey Heer, and Christopher D Manning. 2013. The efficacy of human post-editing for language translation. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 439–448.

[31] Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional lstm. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1589–1599.

[32] Kishaloy Halder, Alan Akbik, Josip Krapac, and Roland Vollgraf. 2020. Task-aware representation of sentences for generic text classification. In *Proceedings of the 28th International Conference on Computational Linguistics*. 3202–3213.

[33] Jeffrey Heer. 2019. Agency plus automation: Designing artificial intelligence into interactive systems. *Proceedings of the National Academy of Sciences* 116, 6 (2019), 1844–1850.

[34] Kenneth Holstein, Vincent Aleven, and Nikol Rummel. 2020. A conceptual framework for human–AI hybrid adaptivity in education. In *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part I 21*. Springer, 240–254.

[35] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 159–166.

[36] Chao-Chun Hsu and Chenhao Tan. 2021. Decision-Focused Summarization. *arXiv preprint arXiv:2109.06896* (2021).

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

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

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

[40] Laura Illia, Elanor Colleoni, and Stelios Zyglidopoulos. 2023. Ethical implications of text generation in the age of artificial intelligence. *Business Ethics, the Environment & Responsibility* 32, 1 (2023), 201–210.

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

[42] Jacob Jacoby and Michael S Matell. 1971. Three-point Likert scales are good enough.

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

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

[45] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *Comput. Surveys* (2022).

[46] Sean Kandel, Andreas Paepcke, Joseph Hellerstein, and Jeffrey Heer. 2011. Wrangler: Interactive visual specification of data transformation scripts. In *Proceedings of the sigchi conference on human factors in computing systems*. 3363–3372.

[47] Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine Van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (peerread): Collection, insights and nlp applications. *arXiv preprint arXiv:1804.09635* (2018).

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

[49] Neha Nayak Kennard, Tim O'Gorman, Akshay Sharma, Chhandak Bagchi, Matthew Clinton, Pranay Kumar Yelugam, Rajarshi Das, Hamed Zamani, and Andrew McCallum. 2021. A Dataset for Discourse Structure in Peer Review Discussions. *arXiv preprint arXiv:2110.08520* (2021).

[50] Asheesh Kumar, Tirthankar Ghosal, and Asif Ekbal. 2021. A Deep Neural Architecture for Decision-Aware Meta-Review Generation. In *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, 222–225.

[51] Sandeep Kumar, Hardik Arora, Tirthankar Ghosal, and Asif Ekbal. 2022. DeepASPeer: towards an aspect-level sentiment controllable framework for decision prediction from academic peer reviews. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*. 1–11.

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

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

[54] Mina Lee, Percy Liang, and Qian Yang. 2022. CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. In *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

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

[56] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461* (2019).

[57] Miao Li, Jianzhong Qi, and Jey Han Lau. 2022. PeerSum: A Peer Review Dataset for Abstractive Multi-document Summarization. http://arxiv.org/abs/2203.01769 arXiv:2203.01769 [cs].

[58] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[59] Chin-Yew Lin and Franz Josef Och. 2004. 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)*. 605–612.

[60] Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958* (2021).

[61] Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345* (2019).

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

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

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

[65] Francesco Moramarco, Alex Papadopoulos Korfiatis, Aleksandar Savkov, and Ehud Reiter. 2021. A preliminary study on evaluating consultation notes with post-editing. *arXiv preprint arXiv:2104.04402* (2021).

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

[67] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038* (2019).

[68] Andrea Papenmeier, Gwenn Englebienne, and Christin Seifert. 2019. How model accuracy and explanation fidelity influence user trust. *arXiv preprint arXiv:1907.12652* (2019).

[69] Joonsuk Park and Claire Cardie. 2014. Identifying appropriate support for propositions in online user comments. In *Proceedings of the first workshop on argumentation mining*. 29–38.

[70] Isaac Persing and Vincent Ng. 2015. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 543–552.

[71] Elizabeth L Pier, Markus Brauer, Amarette Filut, Anna Kaatz, Joshua Raclaw, Mitchell J Nathan, Cecilia E Ford, and Molly Carnes. 2018. Low agreement among reviewers evaluating the same NIH grant applications. *Proceedings of the National Academy of Sciences* 115, 12 (2018), 2952–2957.

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

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

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

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

[76] 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. 2022. How do Authors' Perceptions of their Papers Compare with Co-authors' Perceptions and Peer-review Decisions? *arXiv preprint arXiv:2211.12966* (2022).

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

[78] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[79] Laura Scholes, Kathy A Mills, and Elizabeth Wallace. 2022. Boys' gaming identities and opportunities for learning. *Learning, Media and Technology* 47, 2 (2022), 163–178.

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

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

[82] Chenhui Shen, Liying Cheng, Ran Zhou, Lidong Bing, Yang You, and Luo Si. [n.d.]. MReD: A Meta-Review Dataset for Controllable Text Generation. *arXiv preprint arXiv:2110.07474* ([n. d.]).

[83] Ben Shneiderman and Pattie Maes. 1997. Direct manipulation vs. interface agents. *interactions* 4, 6 (1997), 42–61.

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

[85] Raimel Sobrino-Duque, Juan Manuel Carrillo-de Gea, Juan José López-Jiménez, Joaquín Nicolás Ros, and José Luis Fernández-Alemán. 2022. Usevalia: Managing Inspection-Based Usability Audits. *International Journal of Human–Computer Interaction* (2022), 1–25.

[86] Logan Stapleton, Min Hun Lee, Diana Qing, Marya Wright, Alexandra Chouldechova, Ken Holstein, Zhiwei Steven Wu, and Haiyi Zhu. 2022. Imagining new futures beyond predictive systems in child welfare: A qualitative study with impacted stakeholders. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1162–1177.

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

[88] Henning Wachsmuth, Nona Naderi, Ivan Habernal, Yufang Hou, Graeme Hirst, Iryna Gurevych, and Benno Stein. 2017. Argumentation quality assessment: Theory vs. practice. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 250–255.

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

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

[91] Daniel Karl I Weidele, Justin D Weisz, Erick Oduor, Michael Muller, Josh Andres, Alexander Gray, and Dakuo Wang. 2020. AutoAIViz: opening the blackbox of automated artificial intelligence with conditional parallel coordinates. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 308–312.

[92] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359* (2021).

[93] Kanit Wongsuphasawat, Dominik Moritz, Anushka Anand, Jock Mackinlay, Bill Howe, and Jeffrey Heer. 2015. Voyager: Exploratory analysis via faceted browsing of visualization recommendations. *IEEE transactions on visualization and computer graphics* 22, 1 (2015), 649–658.

[94] T Elizabeth Workman, Marcelo Fiszman, and John F Hurdle. 2012. Text summarization as a decision support aid. *BMC medical informatics and decision making* 12, 1 (2012), 1–12.

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

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

[97] Jiacheng Xu and Greg Durrett. 2019. Neural extractive text summarization with syntactic compression. *arXiv preprint arXiv:1902.00863* (2019).

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

[99] 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.

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

[101] Amy X Zhang and Justin Cranshaw. 2018. Making sense of group chat through collaborative tagging and summarization. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–27.

[102] Haiyi Zhu, Bowen Yu, Aaron Halfaker, and Loren Terveen. 2018. Value-sensitive algorithm design: Method, case study, and lessons. *Proceedings of the ACM on human-computer interaction* 2, CSCW (2018), 1–23.

[103] Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in*

information retrieval. 1097–1100.