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

LU SUN, University of California San Diego, USA
STONE TAO, University of California San Diego, USA
JUNJIE HU, University of Wisconsin-Madison, USA
STEVEN P. DOW, University of California San Diego, USA

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 designing intelligent scaffolding for writing meta-reviews for an academic peer review process. We prototyped a system called "MetaWriter" trained on five 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: academic peer review, meta-review, AI scaffolding, LLM

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

1 INTRODUCTION

Peer review is a key cornerstone of academic research [66]. The peer review process helps improve the research quality by providing feedback and an assessment of the paper [40, 72]. 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 [57, 72, 79]. Reviewers have to take on more submissions which can lead to a slower process, inconsistencies [76], and potential biases [49, 73].

Skyrocketing paper submissions also increase the burden and challenges for meta-reviewers [7]. In many academic communities, papers are evaluated by several reviewers and followed by a

Authors' addresses: Lu Sun, University of California San Diego, San Diego, CA, USA; Stone Tao, University of California San Diego, San Diego, CA, USA; Junjie Hu, University of Wisconsin-Madison, Madison, WI, USA; Steven P. Dow, University of California San Diego, San Diego, CA, USA.

ACM 2573-0142/2024/4-ART94
https://doi.org/10.1145/3637371

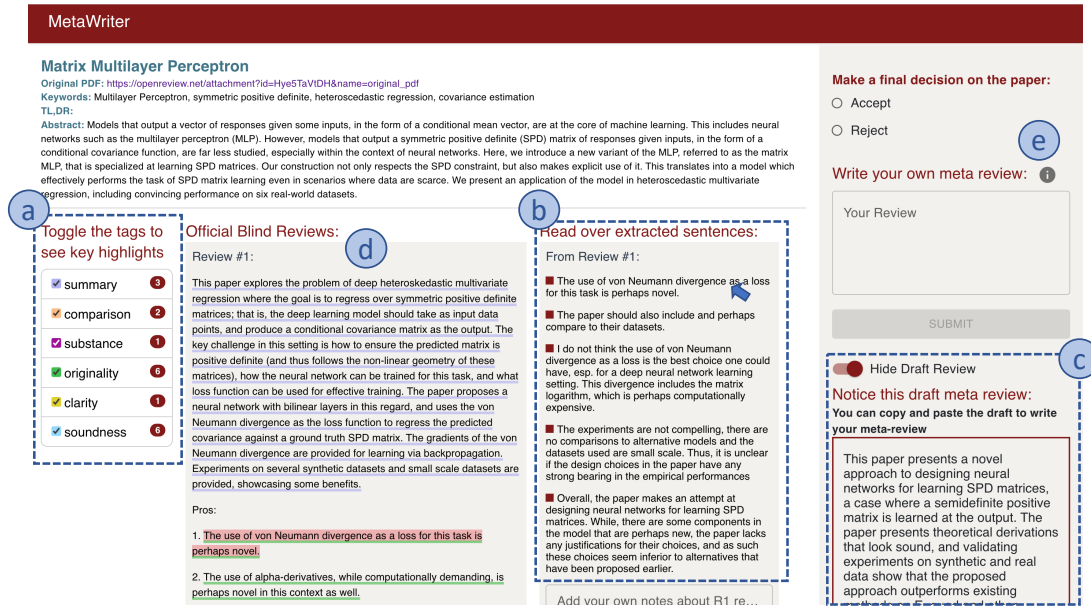


Fig. 1. MetaWriter interface. Users can: (a) toggle the tags to highlight specific aspects; (b) hover over the extracted sentences for each review and the corresponding sentence will be highlighted; (c) see or hide the generated draft using the toggle button; (d) review the three original independent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

"meta-review" that summarizes the reviews and offers a final decision. Meta-reviewers typically need to assign the reviewers, understand the core idea of the paper, make sense of each reviewer's opinions, resolve conflicts between reviewers, write a meta-review that synthesizes viewpoints across reviewers, and make a final decision with sufficient justifications for the paper's authors to understand the decision and iterate on their work. Writing a high-quality comprehensive meta-review with a well-justified decision is a challenging and complex work [72].

Considering the rapid growth of many academic communities, it is common for relatively new researchers to engage with the meta-review process. One strategy to help less experienced meta-reviewers is to provide scaffolding, an instructional strategy for guiding learners via examples, templates, and hints [19, 37, 69]. Prior work showed that scaffolding through examples and templates, if done effectively, can help learners perform similarly to experts, such as writing more effective feedback [95]. Many systems explored how to scaffold better writing, including details like grammar and spelling [1], but also higher-level thinking. For instance, the LetterSmith system used a "scaffolded annotation" approach to break down writing tasks into manageable key components together with annotated expert examples, in order to apply the best practices of professional writing in a new genre [36].

Advances in AI have the potential to enhance scaffolding for complex tasks like writing; as shown in recent research, text editors can now generate new ideas [27], create metaphors [26, 46] or story lines [92–94], summarize long text and rewrite sentences [3] or extend a seed of story into longer writing [4]. Despite recent successes, researchers have raised concerns about AI scaffolding trying to assist too much or creating risks for high-stakes contexts like job hunting [36]. Many modern AI systems leverage large language models (LLMs) that are known to have limitations,

such as a tendency to output contradictory or contrived information [55], which can lead to distrust or misuse [41].

Harnessing AI-based scaffolding in any domain requires a holistic understanding of the work context and potential trade-offs. Our research explores how AI scaffolding can impact the meta-reviewing experience: Can it increase the quality of meta-reviews while maintaining the reviewer's sense of agency and independence? Can it provide opportune guidance for inexperienced reviewers? What benefits and concerns are raised by reviewers and authors?

To answer our research questions, we developed a prototype called MetaWriter [1], as shown in Figure 1, to support the meta-reviewing process using text-based machine-learning techniques. We draw insights from the prior literature on scaffolding and human-AI collaboration to motivate three design goals: scaffold inexperienced meta-reviewers, preserve agency, and mitigate bias. To address these design goals, Metawriter implements: (1) automated tagging to highlight review content related to expert criteria for evaluation (e.g., passages that speak to the originality of the research), (2) extractive summarization to stage a process to compare and synthesize key sentences in the independent reviews, and (3) a generative model that automatically writes a preliminary draft of a meta-review tuned based on the reviews, as well as the paper title and abstract; to preserve a sense of agency around this feature, users must explicitly request the draft and borrowing any text from the draft requires manual copy and paste.

To evaluate MetaWriter, we conducted a mixed-method study to understand how the system impacts performance and to uncover the perspectives of primary and secondary stakeholders (i.e., potential meta-reviewers and authors). In a within-subjects study, 32 participants who had prior experience writing reviews but little experience with meta reviewing, 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 significantly covered more points raised by independent reviewers, summarized the paper idea better, and provided better justifications for their decisions. MetaWriter also reduced the time for participants to generate final meta-reviews compared to the baseline editor. While participants used the key features in MetaWriter differently, they all preferred it over the baseline editor.

However, some participants expressed concerns about over-reliance and the potential for bias to creep in, especially when they considered how others might abuse the system. Participants reported that the machine-generated draft improved their meta-reviewing efficiency, but sacrificed some agency in the writing process. When comparing participants with some prior experience in meta-reviewing to those who never played this role before, we found that inexperienced meta-reviewers benefited significantly more in terms of covering reviewer comments, while even experienced reviewers improved the justifications for their decisions. Interviews with the paper authors also revealed similar concerns about the potential risks of being negligent in gate-keeping responsibilities and eroding trust in the peer review process.

Our paper offers several contributions: We fine-tuned several ML models and developed a prototype that scaffolds the meta-review process by highlighting common review topics, extracting important sentences from each review, and automatically generating an initial draft (only upon request) that can be further edited. Our mixed method study provides empirical data that carefully integrating AI techniques into an authoring environment can facilitate sense-making 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

[1]Code, scripts and data: https://github.com/LusunHCI/MetaWriter-Interface.git

transparency before such systems can be adopted by peer-review communities. We discuss critical reflections on ethical considerations for integrating AI into the peer review process.

2 RELATED WORK

2.1 Conference peer review practices and challenges

Peer review is an essential step of academic research, ensuring the scientific rigor of scholarly work [66]. In a typical conference or journal review process, each reviewer needs to provide reviews for their assigned papers. A discussion for each paper then takes place between its reviewers and the meta-reviewer - who acts as an intermediary between reviewers and program chairs. Sometimes, the author then provides 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 and writes a comprehensive meta-review [72, 73]. Note that conferences or journals may have variations 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.

Meta-reviewers need to synthesize diverse multi-aspects information from authors and different reviewers and then provide a reasonable recommendation for the paper [68, 72, 74]. 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. 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 [7, 9, 35, 73, 84, 96]. However, as far as we know, there has been no previous work in the HCI community focused on supporting meta-reviewers.

For meta-reviewers, navigating conflicts among reviewers while making final decisions and crafting meta-reviews can present substantial challenges [10, 43, 96]. We calculated that 15% of the International Conference on Learning Representations (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 [52]. 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 [2], 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".

Furthermore, though the rapid increase in paper submissions can be viewed as a bloom of scientific progress, it has created a burden on the peer review process [57, 72, 79]. Academic communities sometimes seek new senior researchers to engage in the meta-review process. However, inexperienced meta-reviewers may take some time to become accustomed to the nuances of conducting the meta-review process. To facilitate the transition of less experienced meta-reviewers into their roles, scaffolding could offer a valuable approach to help them build expertise [19, 37].

2.2 AI scaffolding for writing and information synthesis

Scaffolding was used as an instructional strategy to improve learners' problem-solving skills [19, 69]. Prior research showed that effective scaffolding can help novices perform work on par with experts [36, 95]. For example, Yuan et al. founded that providing rubric of design principles helps novices provide feedback that is rated nearly as valuable as expert feedback. In the writing scenario,

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

scholars found that scaffolding can help students learn about form and structure by analyzing the examples and templates [19, 22]. Hui et al. showed that providing high-quality examples and expert-informed checklists can increase student learning and improve writing quality [37]. Another writing support system used "scaffolded annotation" approach that decomposes writing tasks into key components and offers annotated expert examples to help writers apply the best practices of professional writing in a new genre [36].

Recent advances in AI unlocked myriad possibilities for providing efficient scaffolding on complex tasks, including writing and researching [27, 60, 67, 94]. To enhance comprehension of key information [84, 96], researchers employed argument mining or tagging techniques to annotate important points from a massive amount of information, such as social media posts [63], online debates [29, 82], or student essays [64]. Another AI approach can summarize lengthy texts into 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, 34, 88]. 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 [12, 34].

Recent research shows more potential for generative LLMs in scaffolding writing. Writing support tools can now help researchers brainstorm new ideas [2, 27], create metaphors [26, 46], generate storylines [92–94], summarize long text and rewrite sentences [3], suggest arguments [50]or extend a seed of story into longer writing [4]. For example, the Wordcraft project explores how to support users collaborating with generative LLMs to co-write a story [94]. Spark used a language model to generate prompts related to a scientific concept to facilitate scientific writing [27]. VISAR used a generated LLMs to help writers brainstorm goals, organize argument structure, and revise argument phrasing in the argument writing scenario [97]. CoAuthor further explored the capabilities of LLM working as a writing collaborator and reasoned about its ability to generate fluent text, generate new ideas and work jointly with writers [50]. They found that sentences written with GPT model contained more diverse name identities and less grammar errors.

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

Prior work has raised concerns about the potential loss of agency or control when working with AI on complex tasks [6, 33, 75]. Shneiderman and others argued for more direct manipulation interfaces to give users more control of information and decision [75]. 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 [89]. 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 (e.g. autocomplete) [18]. 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 [18].

Some researchers have tried integrating agency and automation by creating effective "collaborative" interfaces using shared representations between humans and AI [42, 87]. 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 [31]. Prior research 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 [12]. Kumar et al. used a deep neural network architecture to generate a

meta-review that would account for paper decisions [47]. Shen et al. proposed a controllable meta-review generation method to generate meta-reviews according to the categories of reviews [74]. However,researchers have raised concerns about the AI directly conducting the task instead of assisting humans to gain expertise. Researchers pointed out automation of creative work may result in "cannibalizing" the creativity of artists engaged in script writing tasks[59, 86]. In the context of meta-review writing processes, if the machine automatically generates a draft, inexperienced meta-reviewers may lose the opportunity to gain the expertise of writing a high quality meta-review. Understanding the tension between human agency and machine automation can help researchers design systems that successfully weave AI scaffolding into the task without sacrificing a user's agency [31, 83].

Previous studies also raise concerns that LLMs could exhibit biases in human-AI collaboration tasks [15, 32, 44, 65, 78, 85]. For example, a previous study on child protective services AI decision support showed that AI could be biased towards certain groups [44, 78]. Previous studies found that providing information, such as explanations, generated by AI can potentially mislead users in decision making [15, 28, 48]. Faced with the challenge, users need a staging process that helps them reflect on the content to make unbiased decisions. During the process, masking the uncertain content can potentially slow down the decision making process. In addition, the limitations of LLMs, such as hallucination or the tendency to produce contradictory or contrived information [41, 55], can limit human's trust of using them. Hence, we need to further explore stakeholders' trust and concerns of using AI scaffolding.

To summarize, meta-reviewers face several challenges: taking time to understand complex reviews, cross-comparing independent reviews to make fair decisions, and authoring comprehensive and well-structured meta-reviews. AI scaffolding can potentially help novices gain the expertise 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 scaffold novices for the meta-review process. More generally, the research community still lacks empirical data on how could AI scaffolding play a role in the peer review and what the potential risks are.

3 METAWRITER SYSTEM

3.1 System design goals

Drawing on the prior literature on scaffolding and expertise from the learning sciences, as well as, studies on human-AI collaboration that explore issues of agency and bias, we specified three primary design goals (DGs) for the MetaWriter prototype that support the entire meta-reviewing process: (1) DG1: Scaffold inexperienced meta-reviewers; (2) DG2: Preserve agency and reduce over-reliance; (3) DG3: Reduce potential bias.

3.1.1 ***DG1: Scaffold inexperience meta-reviewers by embedding expert knowledge structures.***

Prior work indicates that scaffolding techniques like checklists [95], editable examples [36], and annotations [81] can help novices perform more like experts. In MetaWriter, we use visualization techniques to scaffold how participants read and assess the independent reviews. First, color-coded tags highlight content related to considerations on the rubric agreed upon by the academic community. Second, MetaWriter provides an example meta-review structure to convey wisdom about the potential structure for a meta-review. Third, the system not only provides an example from some other context (which requires users to conceptually "transfer" insights into the current task [13, 25]), MetaWriter also 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.*

Collaboration between humans and AI has raised concerns among researchers, as they worry that the increased efficiency resulting from automation may limit users' agency [31] and lead to over-reliance on AI-generated artifacts [27]. In MetaWriter, to preserve agency and provide a layer of user control, we require 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. To reduce the over-reliance on the draft, if their submitted draft is very similar to the provided draft, we pop up a text box to remind people that they cannot directly copy and paste.

3.1.3 *DG3: Increase coverage and reduce bias through process support.*

Prior research found that providing information, such as explanations, generated by AI can potentially mislead users in decision-making and make users vulnerable to cognitive biases [15, 48]. Users need a staging process that helps them reflect on the content to make unbiased decisions. To increase coverage of reviewer points and to reduce the potential for bias, MetaWriter stages a process, such that meta reviewers must first consume each unique point raised by reviewers, and make a decision, before they can see a draft meta-review. If and when the user requests to see the draft, the generated meta-review excludes any sentences that might bias the user's decision.

3.2 System Design

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 (as shown in Figure 1) to guide the participant through performing a meta-review: (1) color-coded tags that can be toggled on/off to highlight multiple aspects of reviews (Figure 1-a); (2) extracted key sentences that participants can hover over to locate the position within the review (Figure 1-b); (3) a generated draft meta-review that users can copy, paste, and edit as their own initial draft if they wish (Figure 1-c). Participants can interact with these three features while forming a decision and delivering the final meta-review. MetaWriter simulates the OpenReview platform[1] which provides the paper abstract, a link to the paper pdf, keywords, and the three original independent reviews.

3.2.1 *Tags that color-code multi-aspects in each reviewer's review.*

A good review not only contains a good summary of the paper but also consists of accurate comments with high coverage from multiple aspects to evaluate the paper's quality. 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 [17, 24, 35, 40, 49, 76, 84, 90, 96]. 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.

3.2.2 *Extracted key sentences that identify main points from each review.*

Figure 1-b lists key sentences extracted by MetaWriter from each independent review. When participants hover over one sentence, the same sentence within the review paragraph will be highlighted in red to help participants locate the source. For example, when the user hovers over the sentence "The user of vonNeumann divergence as a loss for this task is perhaps novel", MetaWriter highlighted the corresponding sentence in red in the original individual review area.

[1] https://openreview.net/

3.2.3 *Generated meta-review draft with structure.*

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 [51]. Similarly, meta-reviews typically synthesize all reviews into one cohesive summary without directly copying sentences from each [74]. MetaWriter offers a draft meta-review for participants to use as inspiration or a starting point, as shown in Figure 1-c. Users need to toggle to read or hide the meta-review if they think the draft might influence their judgment of the paper. Instead of predicting the paper decision, we asked participants to make their own choice and we removed the sentences that indicated the decision in the meta-review draft. In Figure 1 (d), participants need to make a decision and then write the final meta-review. We also provided instruction and guidance for both conditions.

3.3 System implementation

3.3.1 *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 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.[3]

The ICLR dataset we collected includes 9803 submissions. For each submission, we collected paper information including the title, keywords and abstract. All official reviews with reviewer ratings and confidence scores, the final meta-review with ratings, and the final decision were collected. 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. The average length of each individual review is 508.9 words and each meta-review is 146.5 words. We observe that the average length of reviews and meta-reviews increases 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. Further details are in Appendix (Sec. 9).

3.3.2 *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, such as originality, clarity, etc [96], as shown in Fig 2-A. MetaWriter color-coded these tags on each independent review to support reading. Next, we fine-tuned an extractive summarization model [56] that can identify key sentences automatically from each review, 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 fine-tune a language generation model that can automatically generate a draft meta-review [51], as shown in the Fig 2-C. Details on each evaluation can be found in the Appendix (Sec. 9).

Tags that highlight multi-aspects. 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 [96]. 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 [84, 96]. Note that the original model predicts the sentiment of the tag as well but we opted not to include the sentiment in

[3]https://openreview.net/group?id=ICLR.cc.

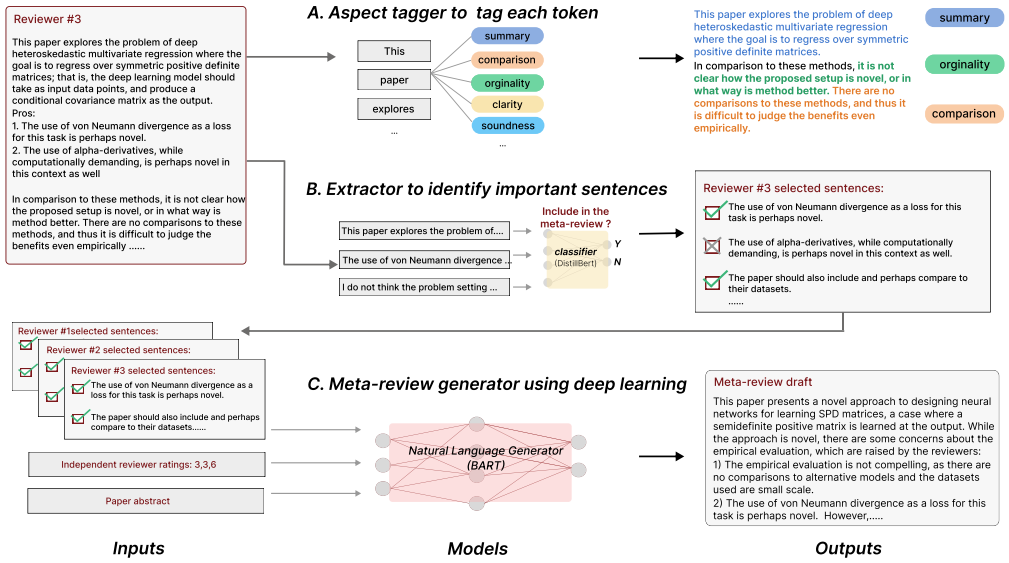


Fig. 2. Architecture of three machine-generated features. Left is the inputs for each model, the middle is the three models, and the right is the corresponding outputs, which are tags, extracted sentences, and generated meta-review draft.

MetaWriter as this could lead to reduced authorial agency amongst users. Among all independent reviews in our ICLR dataset, 30.4% of words received an aspect tag and the rest of words did not reach the probability of belonging to any tags.

Extractive summarization algorithm to select important sentences. Extractive summarization techniques have shown evidence of being able to select strong candidate sentences for a summary [56]. Here, we fine-tuned a pre-trained model to select key sentences from each independent review to shorten participants' time on reading the long reviews as shown in Figure 2-B. To train the extractive summarization model, we fine-tuned the model PreSumm on our dataset of reviews [56]. After following the PreSumm training setup and fine-tuning our dataset, we evaluated our extracted output against the ground truth meta-reviews. We obtained the F1 scores of ROUGE-1 as 0.341, ROUGE-2 as 0.085 and ROUGE-L as 0.162 [53]. For each review, the model can automatically identify key sentences that have the highest probability of being in the meta-review.

Abstract summarization algorithm to generate a draft meta-review. To fine-tune the abstractive summarization model, for each submission, we combined the extracted sentences from all three reviewers, along with their ratings and the paper abstract as inputs, and then used the real meta-review as the output target, as shown in Figure 2-C. We used our dataset to fine-tune the bart-large-cnn model, one variant of the BART model [51]. The fine-tuned model obtained the F1 scores of ROUGE-1 as 0.345, ROUGE-2 as 0.095, and ROUGE-L as 0.207.

3.3.3 System Architecture. 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. We run our trained language models to generate tags and drafts which are stored on our server and sent to the Firebase server. 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 conducted a mixed-method study with potential users—who role-played as meta-reviewers—and authors. Through a within-subjects experiment, we invited participants (N=32) to write meta reviews for two papers, one with MetaWriter and one with a baseline editor, counterbalancing the order. We collected qualitative data on participants' reactions to the MetaWriter system and measured differences in terms of quality and time spent. To gain perspective from another key stakeholder, we conducted semi-structured interviews with six paper authors.

We focused on four key research questions:

- RQ1: How does MetaWriter affect the quality of meta-reviews and the 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 prior experience as a meta reviewer affect the performance and perspectives on MetaWriter?
- 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 baseline editor, we conducted a within-subjects experiment with 32 participants. We simulated a peer review process where participants played the role of a meta-reviewer with the task of reading the independent reviews, writing a meta-review, and making an accept-reject decision. Each participant completed two meta-reviews: one with the MetaWriter system that includes the machine-generated features described above and one with a plain text editor (Baseline). We counterbalanced the order of each condition through random assignment.

4.1.1 Paper 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 used several criteria to ensure that the two papers were similar for meta-reviewers to review. First, both papers had borderline scores (2 weak accepts and 1 weak reject) and ultimately got rejected from ICLR. As described in Section 2.1, many meta-reviewers face the challenge of dealing with disagreements, conflicts and variability between the independent reviewers. Our goal was to simulate a scenario where the meta-reviewer would need to deal with reviewer conflicts and make a tough decision. Second, the length of individual reviews on the two papers was similar (averaging 337.3 words) and fell within the middle range of all ICLR 2020 reviews (between the 40 and 60 percentile). This aimed to ensure that participants spent a similar amount of time reading the independent reviews. Third, the generated meta-reviews for both papers were of similar lengths (1037 words and 1301 words). Fourth, the generated meta-review drafts for each paper had similarly high ROUGE_L scores [53]. Combining all the 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 often a lengthy discussion panel that takes place on OpenReview that meta-reviewers take into consideration. 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.

4.1.2 Participants. Participants filled out a pre-study survey to collect information about their domain expertise, research topics, review experience, prior knowledge about machine learning, and demographics (e.g. age, gender). 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 32 participants had prior experience as ML researchers (on average 4.7 years of writing and submitting ML papers). All of them had done peer reviewing for ML conferences, but only six had been a meta-reviewer or an associate chair for ML conferences. Participants were predominantly senior PhD students, postdocs, professors, or industry researchers.

To ensure participants had enough expertise to understand the selected papers, we only recruited participants who selected relevant keywords in the pre-study, 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.We also asked participants whether they had already read the two selected papers, and if so, we excluded them so that all participants were reading these papers for the first time. All participants were compensated for $20 per hour for this 90-minutes study. The five participants who wrote the highest quality meta-reviews (as rated by experts) received a $20 bonus.

4.1.3 Procedure. First, 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 select papers and the study conditions for each participant. 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 at least 10 minutes on the meta-review, but they could take as much time as needed. Pilot studies with four ML experts found that participants could read and write a meta-review on these short papers within about 30 minutes. Before each session, we gave participants a quick 2-minute demo of the interface, including the basic features in the baseline editors and the ML-enhanced features of MetaWriter. At the end of both 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.

4.1.4 Data Collection and Analysis. We collected a mix of quantitative and qualitative data, including each participant's log data that captured their interactive behaviors with the system, their final meta-reviews for both papers (N=64) post-survey responses (N=32) and interview transcripts (N=32). The research team analyzed all these 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 text and each participant's decision on whether to accept or reject the paper. We tabulated whether the participants' decisions aligned with the original "Reject" decisions on both papers. While the decision is subjective, it provides a measure of accuracy and allows us to observe whether MetaWriter influences participants' decisions. To measure task performance, we asked

two experts with more than five years of experience being a meta-reviewer for ICLR conference to rate the quality of all final meta-reviews (N=64) using the following rubric informed by community guidelines [4] and previous research [84, 96]:

- Summative (summarizes key points from the submission): A good meta-review captures 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 comprehensively 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 three dimensions on a simple scale from 0-2. Three-point scales have been used in recent HCI studies to measure relative degrees [71, 77] and show similar reliability and validity to Likert scales with more points [38].

The research team first provided experts five examples and instructions for rating. The experts rated each dimension of the meta-review independently. The final rating for each dimension is the average of the two experts' ratings. We also measured the lexical diversity of the written meta-review in each condition by calculating the Self-BLEU score for meta-reviews in each condition [99]. 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. 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 to log a range of user activity. 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 text box. 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 include in their own meta-review.

User preferences. To evaluate participants' preferences for the MetaWriter experience compared to the baselines 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. After using MetaWriter, participants were asked to evaluate the usefulness, accuracy, and trust of each of the machine-generated features on a 5-point Likert scale.

User reactions. After the post-survey, we conducted a 15-minute semi-structured interview with all participants to capture their overall thoughts as well as specific perceptions of machine-generated highlights and summaries. For example, the research team asked "What do you think of the difference between the task with and without the support of MetaWriter", and "Which feature did you use the most?", and "What concerns did you have when using MetaWriter?".

Control variables. On the post-survey, participants self-reported their knowledge of each paper's topic from 1=not familiar to 4=very familiar. Participants' average familiarity was 2.1 with a standard deviation of 1.4. The familiarity was slightly higher for one of the paper topics than the

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

other, so we included this as a control variable in our statistical analyses comparing the conditions. All participants reported that they had never read or remembered the specific papers before.

4.1.5 *Data Analysis.*

Quantitative data analysis. To measure the effect of the MetaWriter system on each dimension of quality (eg. summative, coverage, justified), we conducted repeated measure ANCOVA tests. We used paper ID, the order of experiment conditions (whether MetaWriter was used first or second), the self-reported 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, we again ran a repeated measure ANCOVA using the paper ID, the order of experiment condition, the self-reported knowledge level of each paper topic and the meta-review word count as co-variants. For each survey question (e.g. I think the highlighted tags are useful), we converted responses into a 5-point numerical scale and conducted Mann-Whitney U tests to compare the perceptions of each group.

Qualitative data analysis. All semi-structured interviews with participants were recorded and transcribed. Two researchers conducted iterative open coding on the transcripts using Dovetail [5] following the thematic analysis approach [14]. 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 MetaWriter, we reached out to six ICLR paper authors. All interview participants had prior experience in submitting ICLR papers and obtaining reviews and meta-reviews.

4.2.2 *Interview Study Procedure.*

We conducted 30-minute interviews with the paper authors. We began the interview by demonstrating the MetaWriter interface and showing three meta-review examples written by participants in the within-subjects experiment. We first asked participants to reflect on their quality expectations for meta-reviews. To understand their concerns about using the MetaWriter tool, we asked them "what concerns would you have if meta-reviewers used this tool to write a meta-review for your paper?". We then asked them to cross-compare our three machine-generated features in terms of pros and cons. Lastly, we asked them about their general opinions on using AI in the peer review process.

4.2.3 *Interview Data Analysis.*

The semi-structured interviews with the authors were recorded and transcribed. Two researchers conducted iterative open coding on the transcripts using Dovetail following the thematic analysis approach [14]. 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] https://dovetailapp.com/

5 RESULTS

We report our results from the within-subjects experiment and the interview study. In the within-subject experiment, across both conditions, participants spent an average of 18.9 minutes writing a meta-review with an average length of 164 words. 79.7 % of participants chose rejection decisions, consistent with the original decisions for the two papers. Our findings suggested that MetaWriter can make the meta-review process more efficient, enjoyable, and less cognitively demanding. However, both the experiment participants and the authors we interviewed surfaced potential concerns about potential bias and over-reliance when using MetaWriter. For the ease of distinguishing different subjects in two studies, we use "participants" to refer to people in the within-subject experiment (P1-P36), while "authors" to represent the ICLR authors (A1-A6).

5.1 RQ1: How does MetaWriter affect the quality of meta-reviews and the time-on-task compared to a baseline editor?

5.1.1 MetaWriter helps participants write better meta-reviews.

To assess the quality of meta-reviews, two experts judged each final meta-review on a three-point scale (ie. summative, coverage, and justified). We ran ANCOVA tests on each dimension, accounting for co-variants, to compare the two conditions. We found that meta-reviews written with MetaWriter were significantly better at summarizing the paper contribution, covering the independent reviewers, and justifying the decision compared with the meta-reviews written with the baseline editor (see Table 1).

	Baseline	MetaWriter	p	F
summative ratings (0-2)	1.37 (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)	6.85	9.28	-	-
cosine similarity with the original meta-review	0.70	0.80	***	18.63

Table 1. Data on meta-reviews written by participants by condition. The average ratings on each quality dimension and the length are shown, along with the standard deviations. ANCOVA tests show that participants wrote significantly longer, more summative, more comprehensive, and better justified 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 with the order of conditions, the order of the papers, and the knowledge level of each paper topic as co-variates. As shown in Table 1, we found that participants wrote significantly longer meta-reviews with the MetaWriter system. We found no significant interaction effects between the order of the two tasks and conditions, and no differences between the two papers or due to the topic knowledge on two papers.

We calculated the lexical similarity across participants' drafts using self-BLEU scores to compare the diversity of meta-reviews written in each condition [99]. Higher self-BLEU indicates less diversity within a collection of documents. As shown in Table 1, participants in the baseline editor condition wrote less similar, and therefore, more diverse meta-reviews. We did not conduct a statistical test since we only had aggregated data in each condition group. We also 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, controlling for the two papers and

the order of the experiment conditions. We found the cosine similarity between the draft and the original ICLR meta-review in MetaWriter was significantly higher than the baseline condition. Hence, with MetaWriter, participants landed closer to the original ICLR meta-reviewers. The fact that MetaWriter led to meta-reviews that were less diverse and more aligned with the original meta-reviews could represent a tradeoff: the scaffolding seems to help people adhere to community standards but perhaps constrains their creativity.

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, controlling for the length of the meta-review as a co-variate. As shown in Table 2, 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 ($p < .05$). There was no significant interaction effect between the order of the two papers and condition on writing time and total completion time. We distinguished between 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.

We performed repeated measures ANCOVA to examine the effect of the two papers and the two conditions (with vs. without MetaWriter) on writing time, controlling for co-variates including the paperID, the order of experiment condition, self-reported knowledge level, and the meta-review word count. As shown in Table 2, participants spent significantly less time on writing meta-reviews when using the MetaWriter than the baseline condition ($p < .001$). In addition, they spent significantly less time on the total meta-reviewing process ($p < .05$). There was no significant interaction effect between the order of the two papers and condition on writing time and total completion time. We distinguished between the writing time and total time here because MetaWriter involved more upfront reading, including the extracted sentences and the generated draft. While MetaWriter participants appear to spend slightly more time absorbing this extra text content, they gain time benefits when writing the meta-review draft.

	Baseline	MetaWriter	p	F
total time (minutes)	20.08 (5.92)	17.86 (5.12)	.04 *	4.20
writing time (minutes)	10.93 (3.74)	7.94 (3.71)	.000 ***	13.10

Table 2. Time on tasks in both conditions. An ANCOVA test showed that participants spent significantly less time writing a meta-review draft using MetaWriter. (p-value significance codes: 0.001 ∗ ∗ ∗, 0.01 ∗∗, 0.05 ∗)

5.2 RQ2: How do participants perceive the overall value of MetaWriter compared to a baseline editor?

5.2.1 Participants preferred MetaWriter, despite giving up some agency.
After participants had experienced the two systems, we asked participants to compare them overall directly. As shown in Figure 3, participants highly preferred 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"(P13). MetaWriter was perceived as not only being easier to use but also more useful and enjoyable.

In the survey, 90.6% of participants believed it was faster to perform the activity using the MetaWriter interface. During interviews, 12 of 32 participants explicitly mentioned that MetaWriter saved them time when 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 and verify the additional text generated 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 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 explicitly mentioned trade-offs between efficiency and feelings of agency during the interviews (n = 3). P12 said "... the extracted sentences and draft meta-review are like steering me towards one particular way of writing ... So there's a trade-off ... It saved more time, but it means less flexibility. (P5)"

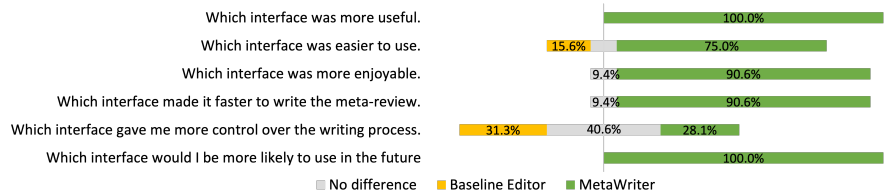


Fig. 3. Participants' comparison between MetaWriter and baseline Editor. Participants preferred MetaWriter overall.

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 paid more attention to them (R1 and R2) to see their judgment."

5.2.2 Participants shared diverse perspectives on the relative value of MetaWriter's scaffolding features.

We conducted an in-depth analysis of how participants used and perceived MetaWriter's three scaffolding features: highlighted tags, extracted sentences, and a generated draft meta-review. In the post-survey, we asked participants 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 (87.5%) thought that the generated draft was 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, participants perceived the extracted sentence feature to have the lowest accuracy, with only 53.1% saying they thought it was accurate. From the qualitative interview data, we also observed that participants used these three features slightly differently. We discuss how participants 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 focus their attention while reading reviews (N = 10). As P16 said " Tags can hint where I was reading. It helped me to locate the information that I need to

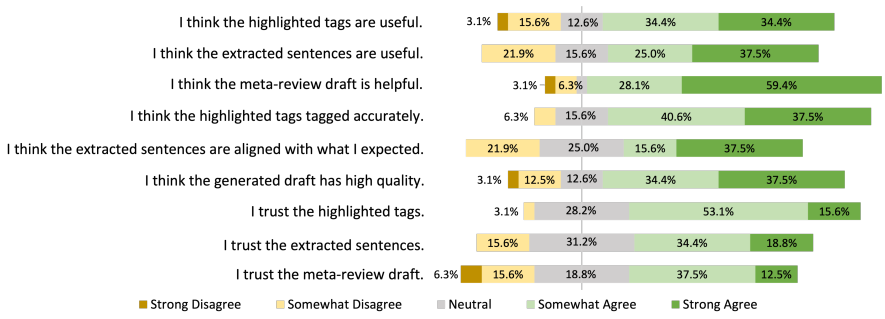


Fig. 4. Participants' feedback on the machine-generated features. Among three features, most participants think that the generated draft is useful but only half of the participants trusted the machine-generated draft.

focus on from each review". Participants also clicked on each tag, eg, originality, to cross-compare individual reviews and identify conflicts (N=6). P15 said "I found that is very powerful when I want to cross-compare the different opinions among the reviewers" However, a few participants mentioned one disadvantage of tags – when the review tags have too many categories, it is hard to distinguish the subtle differences between tags. For example, P3 mentioned "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 surface major issues but may leave out context. Participants used the extracted sentences feature to identify major issues and locate reviewers' viewpoints within individual reviews (N=9). P27 mentioned that "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. So I was using that to locate these sentences." 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=6).

However, some participants pointed out that the extracted sentences might lack context and reasoning (N=4). P14 explained that "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". Participants also raised concerns about the accuracy of extracted sentences and mentioned that the extractor might miss some important points (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 understand or grasp what words are conveying. It will mislead participants if it only selects some points that are more or less problematic". Other participants 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 draft helps kick start the writing process, but raises concerns. Participants reflected that the generated summary provided a good starting point from which they could add their own viewpoints and judgments. Seven participants explicitly mentioned that the meta-review draft is very similar to what they would write and the draft looked very natural – as P23 mentioned "most of the sentences look very natural to me and some hit the point". P8 also reflected that the

draft contains the structure of a meta-review and lists important points that they can use as a starting point.

However, 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). Participants also raised concerns about potential bias in the generated draft (N=11). Some worried that the tone of the generated draft might influence their judgment on the paper. "if it's not accurate, it might potentially put a bias there (P11)." In the experiment, we observed that when participants' judgment did not align with 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."

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. 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 participants a toggle button to hide the draft meta-review. Some participants were concerned that the generated meta-review may contain some points that were not mentioned by any reviewers. As P6 mentioned "highlighted tags and extracted sentences are purely based on the existing information so they cannot produce too much harm, but no one knows what ML models can generate and where it comes from". Participants also expressed concerns when the draft was not coherent with some reviewers' viewpoints. P15 pointed out "it says all reviewers agreed that [the paper] is a little bit dense, but according to what I saw, only the Reviewer 2 mentioned that. The paper is not easy to follow and a bit dense, but this summarization might mislead ACs"

5.3 RQ3: How does prior experience as a meta reviewer affect the performance and perspectives on MetaWriter?

Our design for MetaWriter explicitly builds on theories and techniques for scaffolding, an approach for guiding novices based on expert knowledge and practices. This raises the question of whether the tool is equally valuable for different levels of task expertise, or would more experienced reviewers prefer to see the scaffolding fall away over time. Our participant pool consisted of 6 people with one or more prior experience writing meta-reviews, and 26 with no meta-review experience. This variation provided us with an opportunity to explore how expertise impacts the performance and perspectives when using MetaWriter. As a post hoc analysis, we separated participant data into two groups and added a binary variable for whether the person had prior meta review experience or not.

5.3.1 MetaWriter helped inexperienced meta-reviewer more than experienced meta-reviewer.

Table 3 below compares the performance of experienced and inexperienced participants when using MetaWriter and the Baseline condition. Given that the experienced meta-reviewer group had a smaller size, we conducted the Mann-Whitney U Test, a non-parametric statistic rank test between two samples [58], instead of the ANOVA test. 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 helped inexperienced

Participants	Condition	time (mins)		quality			length
		total	writing	summative	coverage	justified	
experienced	MetaWriter	17.0(7.0)	7.7(4.0)	2.00(0.0)	1.80(0.41)	1.80(0.45)	225.7 (63.0)
	Baseline	17.0 (9.1)	9.3 (4.2)	1.33(0.52))	1.50 (0.84)	1.33(0.52)	157.0 (44.6)
				*	-	*	.
inexperienced	MetaWriter	17.9(4.7)	8.0(3.7)	1.88 (0.35)	1.60 (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 3. 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(word count), 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 ∗

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 adopt a basic structure for meta-reviews, allowing them to focus more on quality improvements. MetaWriter seemed to provide most value to inexperienced participants in terms of scaffolding an initial structure, which can be daunting for first-timers. In terms of the expert ratings on the quality of meta-reviews, MetaWriter helped participants significantly improve the overall summary and the justification of the decision, regardless of prior experience with the task. Only inexperienced participants saw significant improvement in their coverage of the points raised by the independent reviewers.

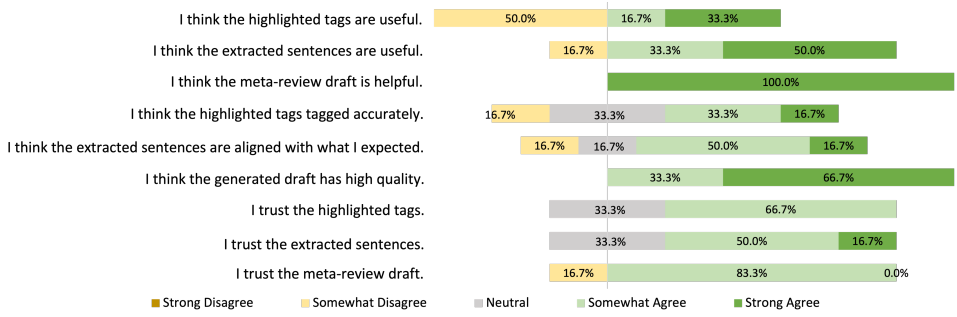
5.3.2 Inexperienced and experienced meta-reviewers view the machine support differently.

As shown in Figure 5a and 5b, for each survey question, we conducted Mann-Whitney U tests to compare the perceptions of experienced and inexperienced groups. Interestingly, 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). We also find that the more experienced meta-reviewers trusted the meta-review draft slightly more than inexperienced participants, although this trend is not a significant effect.

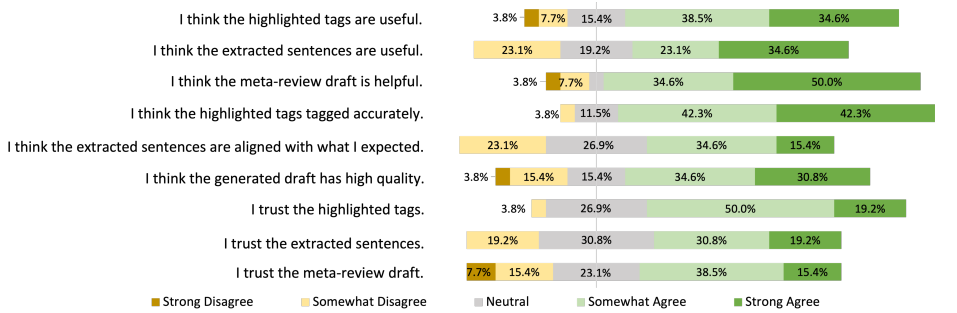
Breaking perceptions down by detailed features and experience level (see Fig 5a and 5b), we see that all experienced meta-reviewers perceived the draft to be helpful, while only 84.6% of inexperienced reviews agreed with that point. Interestingly, this seems at odds with the data in Figure 3 where, across all participants, we see that only 28.1% think MetaWriter provided more control than the Baseline interface. Interviewees mentioned that they still have control over what they can edit, but they were more worried about other reviewers may lose control and fully rely on the draft. One experienced meta-reviewer 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). 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 be inclined to make decisions that are more aligned with what the draft described."

5.4 RQ4: How do paper authors react to the meta-review assisted by the MetaWriter?

We interviewed six ICLR paper authors. Half of authors reflected this tool could be helpful as a training tool or help "visualize" aspects in the meta-review process. The authors also provided a



(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.

rating on a 5-point scale (from 1= not concerned at all to 5= extremely concerned) on how they feel about reviewers leveraging machine intelligence. The average rating was 3.6, with all authors giving a three or higher.

While most authors were fine with using machine intelligence to summarize the reviews, they were concerned that meta-reviewers may over-rely on the generated draft and not put enough effort into writing constructive feedback or justifying the reviews. A3 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." A1 also echoed that point and expressed concerns about leaning too much on 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". This concern around over-reliance also appears in other domains of AI-assisted decision-making [8, 15]. In addition, paper authors asked for more transparency around how the tool would be used in the peer review process. A2 wanted more information about the role of machine intelligence: "whether the decision is made by the tool? Or they only use this tool to write the meta-review."

Further, some authors mentioned that meta-reviewers play a vital role in gatekeeping; they assess the quality and correctness of reviews and verify that reviewers are experts in the corresponding area. Authors pointed out that MetaWriter can ease the process, but may hinder meta-reviewers ability to filter out low-quality reviews and/or reviewers.

Among our three key AI scaffolding features, paper authors were 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."(A6). Paper authors were less concerned about the highlighted tags and extracted sentences since they think these features can mainly enhance reading and sense-making. Paper authors also suggested that such a system could also be useful to authors for the sake of writing rebuttals or reflecting on the reviews.

6 DISCUSSION

We conducted a mixed-method study that includes a within-subjects experiment with potential users (inexperienced and experienced meta-reviews) and an interview study with authors to explore the potential and perils of using AI in the meta-review process. In the within-subjects experiment, we found that participants not only preferred MetaWriter, they also wrote longer and more comprehensive meta-reviews in a shorter amount of time than when using a baseline editor. Meta-reviews written in the MetaWriter condition were also less diverse and more similar than the Baseline ones.

Both experienced and inexperienced meta-reviewers benefited from the scaffolding in terms of quality measures. We found that scaffolding not only improved the performance of novice meta-reviewers, it also helped experienced meta-reviewers, but for different reasons. Experienced participants appreciated the support they got for summarizing the content and being able to see a generated draft that they could edit. Inexperienced participants saw benefits in terms of time, but also for getting a sense of the important arguments and aligning these within a specific structure for meta-reviews. One open question this study raises is whether the generated draft works better to scaffold the meta-reviewing task compared with just showing an example from another paper. The literature on knowledge transfer and analogical reasoning would suggest that learners benefit from a close mapping between an example and a target domain [25]. A generated draft certainly reduces this gap more than a canned example, but it potentially introduces other problems. We found that experienced meta-reviewers seem to trust themselves to not be negatively impacted by the generated draft knowing they can make edits, while the novices were a bit more hesitant about what impact the draft might have on them. Authors also raised concerns that the generated draft could be misused by meta-reviewers and that this could potentially erode trust in the whole peer review system. Interestingly, the distrust seems to be placed more on the humans using the system, rather than on the algorithms' efficacy.

From the experiment, we observed everyone either modified the draft or did not use it at all. No one copied the draft without edits. Most participants, especially experienced meta-reviewers, perceived the generated draft as being helpful. While the system was designed to maintain users' agency informed by prior insights [18], many participants felt they lacked some control. One participant pointed out that the risks involved with leveraging a machine-generated draft seem to be different than the other features. 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 based on experience. Experienced meta-reviewers felt confident in their ability to assert control over the machine intelligence, and they put more emphasis on efficiency gains. Inexperienced meta-reviewers seemed to place less trust in the generated draft, and thus focused more on the structural and knowledge benefits. Building on prior research [74], maybe future tools like MetaWriter can bake in more control over generated drafts. For example, the draft and extracted sentences could be updated based on how users toggle their ultimate decision (accept or reject) or based on other variables like tone or length.

6.1 Creating AI scaffolding that supports users but prevents uniformity

Prior work on human-AI systems has also explored the potential trade-offs between efficiency and creativity [18, 27, 59]. In the meta-review writing scenario, we found that meta-reviewers not only perform the task faster, they also demonstrate higher quality by learning and adapting the structure of the generated draft. Future studies can explore the longitudinal effect of AI scaffolding in peer-review writing context. One goal of scaffolding is that users can eventually be able to perform the task without the support of the tool [19]. Then, researchers can discern from the novelty effect versus the diminishing need for expert scaffolding as novices gain more expertise on the task.

We also found that participants wrote more similar meta-reviews to each other (and more similar to the original meta-reviews for the seed papers) compared to participants with no scaffolding. On one hand, this could indicate that the MetaWriter will lead to homogeneous work across papers, but on the other hand, it could be an indication that users are more consistently following standards set by the research community and 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 creative expression.

6.2 Understanding how AI systems impact trust for individuals versus communities of use

Previous research showed that writing with an opinionated language model can affect participants' attitudes on social topics [39]. One potential explanation is that LLM suggestions may affect participants' thought processes and drive them to spend time evaluating the suggested content [11]. 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. Our participants expressed discomfort with how the whole community might appropriate the technology. Their concerns partly come from the limits of machine intelligence, but also the potential for over-reliance and for bias to creep in. Peer review is a complex process that involves multiple stakeholders, including authors, reviewers, and meta-reviewers [72]. 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 [98]. Future work should explore how to build trust not only for individual users but also within a community of users.

6.3 Making AI systems transparent without adding cognitive complexity

Increasing transparency around how and why the machine outputs certain information can potentially improve users' trust in automation[23, 62]. Prior research provides explanations for AI decisions or visualizes the origin of words/sentences in text suggestions [23, 70, 85], but this potentially bogs down the user trying to complete a task. For example, to explain how the meta-review draft is generated, MetaWriter could, in theory, include more visualizations on the source of each sentence, as well as explain how the model training works and show the model performance. This potentially leads to extra work for meta-reviewers to read and understand. The added 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 Exploring human-machine hybrid collaboration for the peer review ecosystem

What role should machine intelligence play in this meta-review process? Our study of MetaWriter indicates that participants prefer to use multiple hybrid methods as inspiration in the process of conducting their meta-review. We consider their interaction with machine intelligence as a sort of partnership. 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 [83]. 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 some aspects of cognition 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 scaffold the independent reviewers evaluate papers and write high-quality paper reviews [84, 96]. 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 scaffold more of the academic peer review ecosystem [80]. In addition, while the current study was conducted in the context of a machine learning conference, we hope to explore the value of AI scaffolding for a broader range of academic communities.

6.5 Ethical considerations of AI scaffolding for academic review

Generative AI and LLMs introduce numerous opportunities but also raise ethical considerations in the design of human-AI collaboration systems. In the context of academic meta-review, a primary concern involves the potential violation of academic integrity when directly using automatically generated content in writing artifacts. Another concern revolves around the tendency of LLMs to create inaccurate information or to mislead people [39]. We try to mitigate this in the MetaWriter by staging the process, taking out the decisive sentences in the draft, and preventing direct copying and pasting of the entire draft. By unpacking the concerns of participants and authors, we emphasize that meta-reviewers should not over-rely on machine-generated written artifacts, instead, they should deliver their own judgment and decisions.

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. 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 reduce the complexity by only presenting the original reviews to participants. Notably, the reviewers of the two papers selected for this study did not change their ratings after the rebuttal. However, both the rebuttal phase and the discussion among reviewers can change the paper's final decision. Future systems could explore embedding the discussion text and training the model such that it takes this extra content into consideration [45].

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 do on this task. In a real scenario, our interviewees suggest that meta-reviewers take at least one hour to read and write meta-review on full papers. In our study, for simplicity, we used shorter papers, only provided the original reviews, and eliminated 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 cover all situations, such as when all reviewers indicate 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. Our limited scenarios may put limits to what extent we can generalize the results. In future studies, we could explore machine-generated features for all different types of review situations and different research domains.

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 ChatGPT or GPT4 [5] could be used and fine-tuned to our data. Moreover, for the meta-review draft feature, we used an 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 [94]. However, in the co-creation process with generative LLMs, a primary concern involves from the author's concerns when meta-reviewers directly using automatically generated content or misused the generated artifacts. In addition, there is a 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 degrade. Also reviewing culture is different for each peer review community and therefore, using fixed models could inhibit cultural or norm changes.

8 CONCLUSION

In this paper, we built a novel platform, MetaWriter, that used machine intelligence to generate highlights, summaries, and drafts to support information synthesis and writing for academic meta-reviews. We conducted a mixed-method study to empirically understand how users perceive the value of the different machine supports within an interactive tool. A within-subjects experiment with 32 participants evaluated how MetaWriter affected the time and quality of the meta-reviewing process and the perceptions of participants. We found that MetaWriter significantly shortened the time spent on meta-reviewing and helped participants write longer and better meta-reviews. However, our analysis reveals several risks of using LLMs in peer-review writing including the loss of agency and over-reliance, from participants as well as paper authors.

REFERENCES

[1] 2009. Grammarly. https://grammarly.com/.
[2] 2018. Notion. https://www.notion.so/.
[3] 2020. sudowrite. https://www.sudowrite.com/.
[4] 2020. wordtune. https://www.wordtune.com.
[5] 2022. ChatGPT. https://chat.openai.com/.
[6] 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.
[7] 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.
[8] 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.

[9] 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.

[10] 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].

[11] 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.

[12] 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.

[13] 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.

[14] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[15] 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.

[16] 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.

[17] 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.

[18] 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).

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

[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. Metaphoria: An algorithmic companion for metaphor creation. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.

[27] 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].

[28] 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).

[29] 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.

[30] 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.

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

[32] 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.

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

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

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

[36] 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.

[37] 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.

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

[39] 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.

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

[41] 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).

[42] 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.

[43] 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).

[44] 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.

[45] 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).

[46] Jeongyeon Kim, Sangho Suh, Lydia B Chilton, and Haijun Xia. 2023. Metaphorian: Leveraging Large Language Models to Support Extended Metaphor Creation for Science Writing. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. 115–135.

[47] 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.

[48] 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.

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

[50] 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*. 1–19.

[51] 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).

[52] 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].

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

[54] 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.

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

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

[57] 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.

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

[59] 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].

[60] 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.

[61] 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).

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

[63] 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.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] 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

[69] 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.

[70] 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.

[71] 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.

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

[73] 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).

[74] Chenhui Shen, Liying Cheng, Ran Zhou, Lidong Bing, Yang You, and Luo Si. 2021. Mred: A meta-review dataset for structure-controllable text generation. *arXiv preprint arXiv:2110.07474* (2021).

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

[76] 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.

[77] 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.

[78] 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.

[79] 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.

[80] Lu Sun, Aaron Chan, Yun Seo Chang, and Steven P. Dow. 2024. ReviewFlow: Intelligent Scaffolding to Support Academic Peer Reviewing. In *29th International Conference on Intelligent User Interfaces*. https://doi.org/10.1145/3640543.3645159

[81] Jakko Van der Pol, Wilfried Admiraal, and P Robert-Jan Simons. 2006. The affordance of anchored discussion for the collaborative processing of academic texts. *International Journal of Computer-Supported Collaborative Learning* 1 (2006), 339–357.

[82] 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.

[83] 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.

[84] 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].

[85] 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.

[86] 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).

[87] 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.

[88] 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.

[89] 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.

[90] 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

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

[92] Kevin Yang, Dan Klein, Nanyun Peng, and Yuandong Tian. 2022. Doc: Improving long story coherence with detailed outline control. *arXiv preprint arXiv:2212.10077* (2022).

[93] Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 7378–7385.

[94] 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.

[95] 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.

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

[97] Zheng Zhang, Jie Gao, Ranjodh Singh Dhaliwal, and Toby Jia-Jun Li. 2023. VISAR: A Human-AI Argumentative Writing Assistant with Visual Programming and Rapid Draft Prototyping. *arXiv preprint arXiv:2304.07810* (2023).

[98] 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.

[99] 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.

9 APPENDIX

9.1 ICLR dataset

Table 4 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 meta-reviews and 34,219 independent reviews. Table 4 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 increases 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 [6] on the dataset, we found that most reviews are framed in a negative sentiment [30]. In the independent reviews, 81.7% of sentences were 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	100.0
2019	1419	502	917	470.4	139.1
2020	2213	687	1526	478.4	126.2
2021	2616	860	1756	567.8	180.0
2022	2645	1094	1551	596.8	187.0
Total	9,803	3,478	6,325	508.9	146.5

Table 4. 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 an averages where the lengths from each year are weighted equally.

9.2 Algorithm details

9.2.1 Tags that highlight multi-aspects.

MetaWriter directly adopts a fine-tuned tagger that is trained on a similar conference peer review dataset to color code reviews [96]. 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 [43, 84, 96]. Note that the original model predicts the sentiment of the tag as well but we opted not to include the sentiment in MetaWriter as this could lead to reduced authorial agency amongst users. Among all independent reviews in our ICLR dataset, 30.4% of words received an aspect tag and the rest of the words do not reach the probability of belonging to any of the tags. The research team downloaded the review ID from the training dataset of the tagging model to filter out them in the MetaWriter training data Table 5 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.

9.2.2 Extractive summarization algorithm.

Extractive summarization techniques have shown evidence of being able to select strong candidate sentences for a summary [56]. Here, we utilized the extractive summarization technique and fine-tuned a pre-trained model to select key sentences from each independent review to shorten participants' time reading the long reviews. 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 [16, 56, 91]. The

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

Tagged aspect	Example	Freq (%)
Summary	This paper presents a neural 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 5. Tagged aspect category, examples sentences appear in the independent reviews and frequency. Among all tagged aspects, summary appears most frequently.

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 [91] 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 [53]. ROUGE-n scores count the number of overlapping n-grams and word sequences between the summary to be evaluated and the human written summary.

We primarily use ROUGE scores to compare our fine-tuned models with those 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 [54, 74]. 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 on our dataset of reviews and their beam-searched labels [56]. 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.

9.2.3 Abstractive summarization algorithm.

To generate a realistic and natural draft, we fine-tuned an abstractive summarization model on our dataset using the BART model [51]. 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 used the real meta-review as the output target. In order to separate different reviewers' inputs, we add a special token "<SEP>" between them. We used our dataset to fine-tune the bart-large-cnn model, one variant of the BART model [51]. More specifically, we use the PyTorch implementation in the open-source library Fairseq [61]. All experiments are conducted on 4 V-100 GPUs, using a batch size of 4. During fine-tuning, we set the maximum number of tokens as 1024 where the transformer model truncates the source length to 1024 tokens by default. For the rest of the hyperparameters, we use the pre-trained model's default values, where we set the learning rate as 3e05, update frequency as 16 and maximum epoch as 30. 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. After generating the draft, we removed the sentence that specifically mentioned paper's decision using heuristics.

9.3 Examples of MetaWriter's highlighted tags, extracted sentences and generated draft

Parts of reviewer 1 comment with highlighted tags:

– summary – comparison – substance – originality – clarity – soundness

Parts of R1:

This paper explores the problem of deep heteroscedastic multivariate regression where the goal is to regress over symmetric positive definite matrices; that is, the deep learning model should take as input data points, and produce a conditional covariance matrix as the output. The key challenge in this setting is how to ensure the predicted matrix is positive definite (and thus follows the non-linear geometry of these matrices), how the neural network can be trained for this task, and what loss function can be used for effective training. The paper proposes a neural network with bilinear layers in this regard, and uses the von Neumann divergence as the loss function to regress the predicted covariance against a ground truth SPD matrix. The gradients of the von Neumann divergence are provided for learning via backpropagation. Experiments on several synthetic datasets and small scale datasets are provided, showcasing some benefits.

Pros:

1. The use of von Neumann divergence as a loss for this task is perhaps novel.
2. The use of alpha-derivatives, while computationally demanding, is perhaps novel in this context as well.

Cons:

1. I do not think the problem setting or the proposed framework is entirely new or is the best choice of its ingredients. Specifically, the idea of using second-order neural networks have been attempted in several prior papers, including the ones the paper cite (such as Ionescu et al. ,2015).

In comparison to these methods, it is not clear how the proposed setup is novel, or in what way is method better. There are no comparisons to these methods, and thus it is difficult to judge the benefits even empirically.

[...]

4. The experiments are not compelling, there are no comparisons to alternative models and the datasets used are small scale. Thus, it is unclear if the design choices in the paper have any strong bearing in the empirical performances.

Overall, the paper makes an attempt at designing neural networks for learning SPD matrices. While, there are some components in the model that are perhaps new, the paper lacks any justifications for their choices, and as such these choices seem inferior to alternatives that have been proposed earlier. Also, the experimental results are not convincing against prior works.

Selected sentences from Reviewer 1:

✓ The use of von Neumann divergence as a loss for this task is perhaps novel.

✓ The paper should also include and perhaps compare to their datasets.

✓ I do not think the use of von Neumann divergence as a loss is the best choice one could have, esp. for a deep neural network learning setting. This divergence includes the matrix logarithm, which is perhaps computationally expensive.

✓ The experiments are not compelling, there are no comparisons to alternative models and the datasets used are small scale. Thus, it is unclear if the design choices in the paper have any strong bearing in the empirical performances.

✓ Overall, the paper makes an attempt at designing neural networks for learning SPD matrices. While, there are some components in the model that are perhaps new, the paper lacks any justifications for their choices, and as such these choices seem inferior to alternatives that have been proposed earlier.

Generated meta-review from all reviewers' comments:

This paper presents a novel approach to designing neural networks for learning SPD matrices, a case where a semidefinite positive matrix is learned at the output. The paper presents theoretical derivations that look sound, and validating experiments on synthetic and real data show that the proposed approach outperforms existing methods on E_quad and other datasets. While the approach is novel, there are some concerns about the empirical evaluation, which are raised by the reviewers:
1) The empirical evaluation is not compelling, as there are no comparisons to alternative models and the datasets used are small scale.
2) The use of von Neumann divergence as a loss for this task is perhaps novel. However, the gradients of the von Neumann divergence are provided for learning via backpropagation. This divergence includes the matrix logarithm, which is perhaps computationally expensive.
3) The heteroscedastic regression evaluation is problematic. The heteroscedastic regression experiments don't evaluate on one of the key issues, which is uncertainty estimation.
4) The network structure as a whole needs greater validation.
Overall, the paper makes an attempt at designing neural networks for learning SPD matrices, but the paper lacks any justifications for their choices.

Table 6. The example shows how MetaWriter highlights tags, extracts sentences and generates the meta-review draft.

9.4 Demographics of participants

Table below reported the demographics of participants and their review experience.

Gender	Female: 10 Non-binary or gender non-conforming: 1 Male: 21
ML paper writing experience	More than 5 year experience: 13 3-4 years of experience: 3 More than one year experience: 16
ML conferences review experience	Three or more times per year: 17 Once or twice per year: 3 Only once or twice before: 12
ML conferences meta-review experience	Three or more times per year: 5 Only once or twice before: 1

Table 7. Demographics of participants and review experience

Received January 2023; revised July 2023; accepted November 2023