PaperMentor: A Human-Centered Multi-Agent Writing Tutor for AI Research Papers on Overleaf

Jiarui Liu[1,2], Terry Jingchen Zhang[2,3], Ryan Faulkner[2], X. Angelo Huang[3,4]
Vilém Zouhar[4], Dominik Glandorf[5], Isabel Dahlgren[2,3,4], Van Q. Truong[2]
Rishit Dagli[2], Yuen Chen[6], Felix Leeb[7], Punya Syon Pandey[2], Yves Bicker[2,3]
Suvajit Majumder[2], Wenyuan Jiang[4], Zeju Qiu[7], Sankalan Pal Chowdhury[4]
Bernhard Schölkopf[4,7†], Mona Diab[1†], Zhijing Jin[2,3,7†]

[1]CMU [2]Jinesis Lab, University of Toronto & Vector Institute [3]EuroSafeAI [4]ETHZ
[5]EPFL [6]UIUC [7]Max Planck Institute for Intelligent Systems, Tübingen, Germany

Abstract

Expert writing feedback from experienced researchers is critical for early-career scholars to improve their manuscripts, yet high-quality feedback often remains scarce because reviewing research papers is labor-intensive. Emerging AI-powered writing assistants largely focus on grammar fixes or simulating peer review with final scores, yet they fall short of providing concrete, actionable suggestions that help students improve their papers during drafting. We present PaperMentor, a human-centered writing assistant system that delivers actionable suggestions as Overleaf-native inline comments while leaving the actual writing entirely to human authors. PaperMentor integrates an expert skill library carefully curated from established researchers' writing advice with 12 specialized agents covering different aspects of paper writing, such as formatting compliance, phrasing accuracy, and terminology consistency. In a user study ($n = 14$), 90.6% of the generated comments were rated actionable and 67.5% were rated valid, significantly outperforming a GPT-5.2 baseline uswithout the skill library. We release PaperMentor as open source for public use.[1]

1 Introduction

Scientific writing is a core research skill, but many junior AI researchers learn it through trial and error rather than structured mentorship. At top NLP/AI venues such as ACL and NeurIPS, reviewers evaluate clarity, narrative, organization, and adherence to conventions alongside technical merit (Rogers and Augenstein, 2020; Shah, 2022). For authors without experienced mentors, weak presentation can obscure otherwise strong ideas and affect a manuscript's final acceptance (Widom, 2006; Peyton Jones, 2014; Jin, 2024).

Current AI writing tools do not fill this mentoring gap. Grammar assistants such as Grammarly (Grammarly, 2026) and Writefull (Writefull, 2026) focus mainly on sentence-level edits, while AI-powered reviewing tools (Liang et al., 2024; Liu and Shah, 2023; Zhou et al., 2024) simulate peer review and judge paper quality. Neither class of system provides drafting-stage, text-anchored feedback on narrative, organization, and technical presentation, the kind of guidance that student authors need before submission.

We introduce PaperMentor, a human-centered multi-agent writing assistant for AI scientific writing. PaperMentor delivers expert-level feedback as native inline comments on Overleaf, so authors can review suggestions within their existing collaborative workflow while retaining full control over revisions. The system combines a curated library of over 40 expert skill files with 12 specialized agents that review different aspects of a paper, including methods, results, writing style, formatting, and terminology. Each agent is guided by the relevant skills, paper type, venue expectations, and user-provided context.

We evaluate PaperMentor through a user study with 14 AI researchers who annotated comments on 80 papers from ICLR 2026 submissions and internal student drafts. Compared with direct prompting using the same LLM without the skill library, PaperMentor improves validity by 6.5 percentage points and actionability by 4.1 percentage points. We release PaperMentor as open source, providing both an Overleaf-native writing tutor and evidence that expert skill files can improve LLM feedback quality without taking revision control away from authors.

[1]Our code is publicly available under the AGPL-3.0 license at https://github.com/jiarui-liu/overleaf. A live demo can be accessed at https://overleafmentor.ai.toronto.edu/, and our demonstration video at https://youtu.be/BD4caEJtGR0.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pages 397–407
July 2-7, 2026

Figure 1: The three-phase pipeline of PaperMentor. In Phase 1, the system merges the uploaded LaTeX project, collects user input about the target venue and role model paper, extracts structural elements, identifies the paper type, and assigns sections to the appropriate review domains. In Phase 2, specialized review agents analyze their assigned tasks using domain-specific expertise, paper type guidelines, venue expectations, and the selected role model paper to generate structured feedback. In Phase 3, agent comments are deduplicated, consolidated, and mapped back to the original LaTeX source files for display in the Overleaf interface.

2 Related Work

LLM-Based Automated Paper Review Recent research has explored the use of LLMs for automated peer review, but results show that LLMs emphasize surface-level summaries over deeper methodological weaknesses and exhibit limited correlation with human scoring (Liang et al., 2024; Liu and Shah, 2023; Zhou et al., 2024; Yuan et al., 2022; Thakkar et al., 2025; Zhuang et al., 2025; Bougie and Watanabe, 2025; Gao et al., 2025; Cao et al., 2025). AAAI-2026 introduces AI-generated supplementary reviews alongside human reviews (AAAI, 2026). Prior work has also explored multi-agent decompositions (D'Arcy et al., 2024; Chamoun et al., 2024), multimodal input (Taechoyotin et al., 2024; Jin et al., 2024), and retrieval (Zhu et al., 2025). PaperReview.ai reports near-human scoring correlation (Stanford, 2026). However, all of these systems target *review-level* judgments, such as methodological soundness, novelty, and accept/reject reasoning, whereas PaperMentor generates *writing-level* suggestions: concrete, text-anchored comments on writing and structure that authors need during drafting.

Human-AI Collaborative Writing Commercial writing tools such as Writefull (Writefull, 2026) and Grammarly (Grammarly, 2026) provide grammar and vocabulary corrections. Writefull also powers Overleaf's built-in AI assistant, offering context-dependent LaTeX writing suggestions. Prism (OpenAI, 2026) provides an alternative full LaTeX writing workspace with inline AI editing. These tools address surface-level language quality, but they do not focus on structural and organizational feedback, which is especially important in scientific writing, particularly for junior researchers. In contrast to existing commercial tools, research on human-AI writing collaboration offers design principles relevant to our work (Lee et al., 2024), showing that feedback-based assistance (commenting rather than rewriting) better preserves authorial agency (Dhillon et al., 2024; Han et al., 2024). PaperMentor follows this approach by generating text-anchored comments on Overleaf, similar to the feedback a senior researcher would provide. It delivers AI domain-specific and venue-aware guidance through specialized agents informed by an expert skill library, while preserving the author's role as the person who ultimately makes the revisions.

3 Task Definition

Given a LaTeX project, PaperMentor generates a collection of review comments. Each comment includes four pieces of information: the source file it refers to, the character span of the highlighted text, the comment itself, and a severity label. The source file identifies which file in the project the comment

belongs to. The character span marks the beginning and end positions of the highlighted passage. The comment text contains the actual feedback. The severity label indicates the importance of the issue and is one of CRITICAL, WARNING, or SUGGESTION.

4 System Design

Figure 1 illustrates the overall architecture of PaperMentor.

4.1 The Skill Library

The skill library is a curated collection of expert guidance on writing strong AI research papers. It draws from two sources: internal feedback gathered from AI/ML/NLP faculty, and external publicly available writing guides by senior researchers (Jin, 2021; Eisner, 2010; Peyton Jones, 2014; Widom, 2006; Wilson; Rocktäschel and Foerster, 2022; Maddison; Black; Huang, 2023; ACL, 2021; Boyd-Graber; Parikh; Forbes, 2021).

Markup Taxonomy The sources were synthesized into a coherent skill structure by AI research experts with extensive publication experience, yielding six top-level categories: *setup*, *venues*, *paper types*, *sections*, *figures and tables*, and *writing style*. Topical markdown files are defined within each category according to separable sub-skills that address independent aspects of the skill topic, such as paper sections, paper types, and figure elements. Details on agent-skill assignment appear in Section A.

Skill Authoring We curated a collection of publicly available writing guidance, 32 high quality published examples, and 350 reviews from 2025 conferences (NeurIPS, ICLR, and COLM) as the source material for our writing skills. We then used Claude Opus 4.5 (Anthropic, 2025) to restructure and standardize this material into a consistent skill markup format. All generated markup was subsequently human reviewed and refined to ensure consistency, correctness, clarity, and conciseness throughout. The resulting library comprises over 40 skill files totaling more than 16,000 words of expert knowledge, covering paper types, target venues, individual paper sections, writing suggestions, and strategies for learning from role model papers.

4.2 Input Processing

The user uploads a LaTeX project and may optionally (1) select a target venue for submission and (2) provide a role model paper reflecting the style or standards they wish the system to emulate. We begin by resolving nested files to consolidate the project into a single LaTeX source file. We then extract the abstract and all sectioning headers, including sections and subsections up to the appendix. Using the merged source and the extracted structural information, the system identifies the paper type and assigns content to the appropriate review domains.

Paper type identification Different categories of AI research papers follow distinct writing conventions. For example, a dataset paper is expected to describe data collection procedures, annotation guidelines, and evaluation details (Bender and Friedman, 2018; Gebru et al., 2021), whereas a methods paper should clearly present the motivation, formal definition, and comparisons against relevant baselines. Our system recognizes several paper types for which we maintain dedicated expert guidance: *analysis*, *dataset*, *method*, *engineering*, *interdisciplinary*, and *position paper*. Using descriptions of these categories drawn from our skill library, an LLM identifies the most appropriate type. If the paper does not fit any supported category, the type is left unspecified.

Review domain assignment Following standard scientific writing conventions, we define a set of section-level review domains: abstract, introduction, related work, methods (encompassing methodology, datasets, task formulation, and preliminaries), results (encompassing experiments, findings, and discussion), conclusion (encompassing limitations, ethical considerations, and acknowledgements), and appendix. Expert guidance in the skill library is organized around these domains. Given the full draft, an LLM maps each lowest-level section header to one or more review domains. We additionally define global review domains that are not tied to any specific section, such as writing style, mathematical formatting, and table and figure captions.

4.3 Multi-Agent Review

Because the skill library is highly modular and scientific papers are strongly structured, the review task decomposes naturally across multiple

Domain	Severity	Highlighted Text	Comment
Introduction	CRITICAL	*"In this paper, we propose a novel method that outperforms all baselines."*	The contribution statement is too vague. Specify *what* the method does differently and *which* baselines are outperformed. A concrete summary such as "We propose X, which achieves Y% improvement over Z on benchmark W" gives reviewers a clear reason to read further.
Methods	WARNING	*"We follow standard practice for hyperparameter tuning."*	This lacks the detail needed for reproducibility. Specify the hyperparameter search space, the selection criterion (e.g., validation loss), and the number of configurations tried. Reviewers at ACL expect full reproducibility details.
Results	WARNING	*"Table 2 shows the results."*	Start the results interpretation with the key finding, not a pointer to the table. For example: "Our method achieves 42.1 F1, outperforming the strongest baseline by 3.2 points (Table 2)." Then discuss why this improvement matters.
LaTeX	SUGGESTION	*"as shown in table 2"*	Use `\cref{tab:results}` for consistent cross-referencing. This automatically capitalizes and formats the reference (e.g., "Table 2") and updates if table numbering changes.

Table 1: Example comments generated by PaperMentor on a sample paper, illustrating the range of feedback across agents and severity levels. Each comment is anchored to a specific text span and provides a concrete suggestion for improvement.

specialized agents. PaperMentor runs twelve review agents concurrently. Seven section agents each target one review domain (abstract, introduction, related work, methods, results, conclusion, and appendix); three global agents review the whole document for writing style, LaTeX and mathematical formatting, and figures and captions; and two dynamic agents are instantiated per run from the identified paper type and the selected target venue. Each agent receives the relevant LaTeX source, domain-specific skill files, paper-type-specific guidance, venue-specific expectations, and any provided role model paper; Table 3 in Section A gives the full agent skill assignment.

Skills from the library are assigned to agents according to their specialization. Section-specific agents receive only the text relevant to their assigned sections, supplemented by the abstract and introduction for context, so that their inputs remain tightly aligned with their focus. Global agents, such as those handling writing style or formatting, receive the full merged source. Each agent generates comments in the format defined in Section 3. When the input assigned to an agent exceeds a predefined length threshold, the task is further decomposed into smaller subtasks handled by lower-level subagents.

4.4 Comment Aggregation

Comments are aggregated, deduplicated, and mapped back to the original LaTeX files. Even with specialized skills, different agents may occasionally produce overlapping feedback on the same passage. We therefore remove near-duplicate comments whose highlighted spans overlap substantially and whose comment text is lexically similar. When two comments are merged, we keep the one with the higher severity, preferring section-specific agents over global agents. Finally, using the character spans produced by the agents, we map each comment back to its corresponding source file and render it in the Overleaf interface.

5 System Demonstration

PaperMentor is built on the open-source Overleaf Community Edition[2]. This choice preserves the familiar Overleaf writing environment that researchers already use, requiring no change to their existing workflow. AI-generated comments are injected via Overleaf's native ShareJS operational transformation protocol, so they appear in the review panel exactly as human reviewer comments would. Table 1 presents example comments generated by different agents on a sample paper.

Frontend The frontend extends the standard Overleaf interface with a new panel in the editor's sidebar rail, implemented as a React component in TypeScript (Figure 2). The panel exposes four controls: (1) a model selection dropdown for the backbone LLM, (2) an optional target venue field, (3) an optional role model paper upload, and (4) a "Run Full Review" button. Once triggered, a progress

[2] https://github.com/overleaf/overleaf

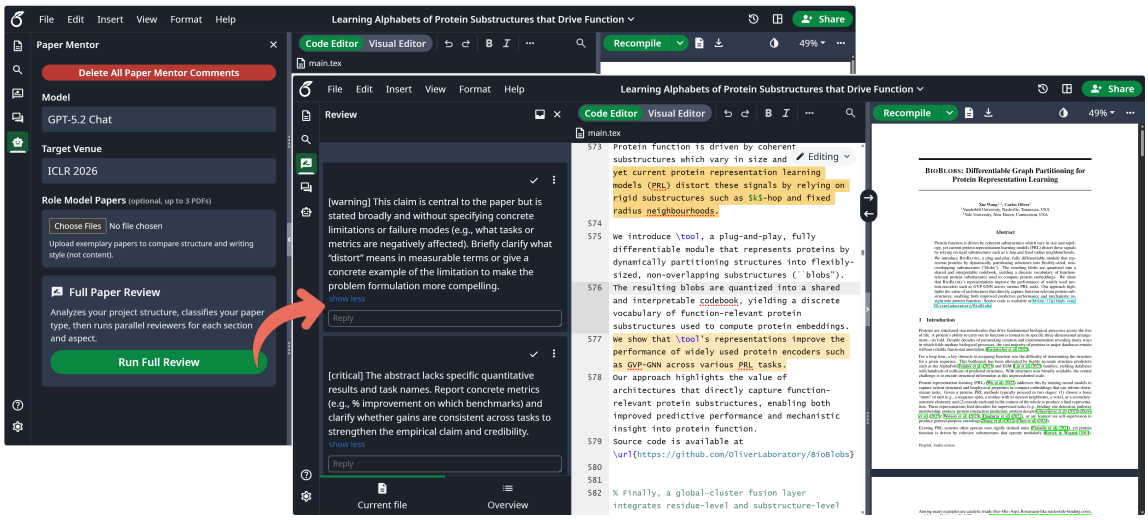

Figure 2: The PaperMentor panel within the Overleaf editor acts as a plugin that appears in the Overleaf sidebar. Left: After selecting the underlying agent model, the intended target venue for submission, and optionally one or more role model papers for reference, the user clicks "Run Full Review" and waits one to two minutes. Right: Once the comments are generated, the user navigates to the review panel to view all feedback produced by PaperMentor. We show this example view using the Wang et al. (2025) paper.

indicator is displayed until the review completes. The results appear as a collapsible summary showing the detected paper type alongside a file-by-file comment list with severity indicators. Comments are simultaneously applied to Overleaf's native review panel, where they appear with highlighted spans anchored to the corresponding locations in the LaTeX source.

Backend The backend consists of new Express.js route handlers and the review orchestration engine, implemented as ES modules within the existing Overleaf web service. When the user clicks "Run Full Review," the frontend issues a POST request to the `/ai-tutor-review` endpoint carrying the project ID and selected model. The backend then retrieves all project documents, produces the merged TEX file, executes the three-phase pipeline, and returns results organized by source file.

6 User Study for Evaluation

Because our core contribution lies in the expert-guided skill library, we evaluate whether our system, powered by this skill library, outperforms state-of-the-art LLM baselines in providing comments and writing suggestions.

6.1 Experimental Setup

Systems Compared For the baseline, we use the same LLM to directly generate comments on a paper without access to the skill library, while keeping all other prompt components identical. This ensures that any performance differences can be attributed to the incorporation of the skill library rather than other variations. We use GPT-5.2 (OpenAI, 2025) for both PaperMentor and the baseline.

Dataset We collect a total of 80 papers with compilable LaTeX sources: 10 from prior internal student submissions and 70 randomly sampled from ICLR 2026 submissions that include arXiv links with downloadable LaTeX source files. We intentionally sample from all submissions rather than only accepted papers to ensure a broad spectrum of paper quality.

Annotation Criteria We evaluate comment quality along three dimensions: *validity*, *actionability*, and *conciseness*. *Validity* asks whether the feedback is factually correct and relevant to the highlighted text. *Actionability* asks whether the feedback clearly suggests what the author should change. *Conciseness* asks whether the feedback is brief and to the point, without unnecessary detail or repetition.

6.2 User Study Design

Participants We recruit 14 researchers in AI with academic backgrounds ranging from undergraduate to PhD students. Each participant logs into an assigned account on our hosted Overleaf platform and annotates four papers. For each paper, participants evaluate 60 comments: 30 generated

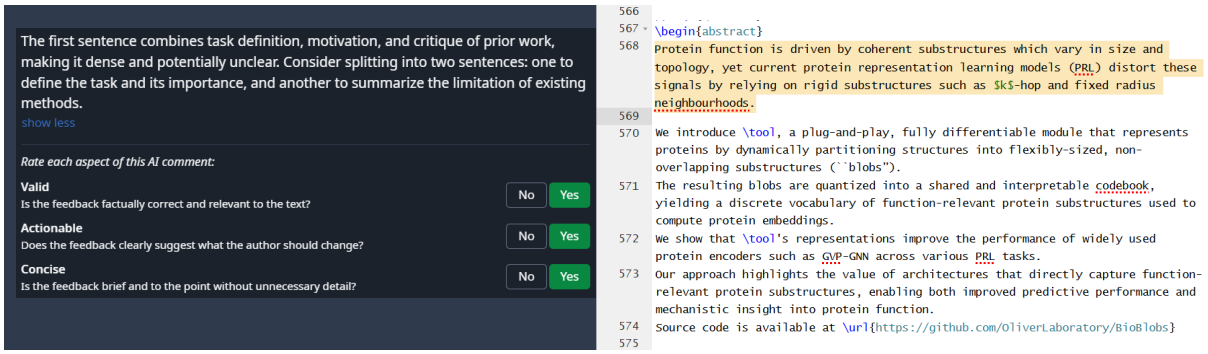

Figure 3: An example showing the annotation of a PaperMentor generated comment on our interface. The paper shown is written by Wang et al. (2025).

by PaperMentor and 30 by the baseline. On the frontend, all these comments look exactly the same without layout distinctions, avoiding potential bias in the annotators' ratings.

Procedure Annotators are provided with a detailed guideline document outlining the evaluation criteria. For each comment, they assess three dimensions: validity, actionability, and conciseness, selecting a binary judgment (Yes or No) for each. An example of the annotation interface is shown in Figure 3.

IRB We follow the research ethics guidelines at ETH Zurich.[3] This study is exempt from ethics approval as it constitutes a survey that (1) focuses exclusively on expert knowledge (the expert acts as an informant and is not the object of the research itself); (2) offers no financial compensation; and (3) includes no experimental features such as deception, incomplete information about the study, interventions, or stimuli. We ensure (a) data protection in accordance with GDPR, (b) informed consent obtained from each expert annotator[4], (c) strictly voluntary participation, and (d) that all collected data contain no personally identifying information.

6.3 Results

Table 2 presents the annotation results for PaperMentor and the baseline. PaperMentor significantly outperforms the direct prompting baseline in both validity and actionability. In contrast, baseline comments achieve higher conciseness on average. Overall, incorporating the skill library enables PaperMentor to generate feedback that is more accurate and more actionable.

System	Validity	Actionability	Conciseness
PaperMentor	0.675 ± 0.023	0.906 ± 0.014	0.900 ± 0.015
BASELINE	0.610 ± 0.023	0.865 ± 0.016	0.973 ± 0.008
Δ	$+0.065^*$	$+0.041^*$	-0.073^*

Table 2: Mean human annotation ratings for PaperMentor and the baseline across three binary metrics: validity, actionability, and conciseness ($\pm$ 95% CI). $^*p < 0.001$ (Mann–Whitney U test).

Although the prompt provides the same instructions, incorporating the skill library increases comment length. This suggests a trade-off between conciseness and improvements in validity and actionability when adhering to structured writing guidelines.

Approximately 40% of comments focus on the Methods and Results sections (see Section B). When accounting for section length, PaperMentor allocates relatively more attention to high-impact sections such as the Abstract and Methods, while placing less emphasis on appendices (see Section C). Annotation scores remain consistent across major sections, indicating that comment quality generalizes well across different parts of the paper.

After completing the annotations, we qualitatively collected annotators' feedback on the comments they reviewed.[5] Overall, respondents viewed the AI feedback positively, with most agreeing that it mimicked a professor's tone, was easy to understand, useful for improving their paper, and generally balanced in its level of critique. The system was seen as particularly effective for clarity, depth of analysis, and grammar, and it led to moderate improvements in thesis clarity, supporting evidence,

[3] https://ethz.ch/en/research/ethics-and-animal-welfare/research-ethics.html

[4] https://docs.google.com/forms/d/e/1FAIpQLSd9R7c-gltvVZz9z7njYZVs9gHGDY01Nbh0k3Jm4QGyPm8Rqg/viewform?usp=header

[5] https://docs.google.com/forms/d/e/1FAIpQLSe76XgkNmVhftxiV3zisP3O0f98tvDvs9TdWmbmH2d1m0vPXQ/viewform?usp=preview

and academic rigor.

7 Skill Library Extensibility

The skill library is designed as a living resource that can evolve over time. Researchers can extend it by contributing new skills or refining existing ones through simple text-based edits. This design makes the system easier to adapt than a fixed prompt or monolithic reviewer, since venue expectations, paper types, and disciplinary writing norms can be updated independently as the community's standards change.

We envision a community-driven development model in which writing advice from senior AI researchers across diverse subfields such as HCI, NLP, and computer vision is systematically encoded into the library. Such contributions can either enhance existing skills or be incorporated as additional skill modules. Over time, this process could turn PaperMentor from a single writing assistant into shared infrastructure for collecting, maintaining, and operationalizing practical paper-writing knowledge.

8 Conclusion

PaperMentor introduces a human-centered, multi-agent writing assistant that delivers expert-guided, actionable feedback directly within the Overleaf drafting workflow. By grounding specialized review agents in a curated skill library distilled from senior researchers' guidance, the system significantly improves the validity and actionability of comments over a direct prompting baseline. More broadly, our results suggest that AI writing support for research papers should move beyond generic rewriting toward structured, mentor-like feedback that helps authors revise their own work while preserving authorship and judgment.

Limitations and Future Work

PaperMentor currently operates primarily over LaTeX source and may therefore miss issues that depend on rendered PDF output, visual figure quality, or numerical verification. Our evaluation includes 80 papers and 14 annotators, which is sufficient to demonstrate statistically significant improvements over the baseline, but does not capture the full diversity of writing styles, venues, disciplines, and researcher backgrounds. In addition, the system depends on both the coverage of the skill library and the reliability of the underlying LLM. Consequently, its feedback should be viewed as drafting assistance rather than authoritative review judgments.

Several directions remain for future work. First, our evaluation focuses on an ablation study that isolates the contribution of the skill library by comparing PaperMentor against the same LLM without access to expert writing skills. While this design allows us to measure the effect of the skill library, it does not directly compare system-generated feedback against comments written by experienced researchers. Collecting and benchmarking against expert authored Overleaf comments would provide a stronger reference point for evaluating the overall quality and usefulness of the system.

Second, our results reveal a tradeoff between specialization and global document awareness. To improve efficiency, section-specific agents operate on limited portions of the manuscript rather than the entire paper. As a result, some validity errors occur when agents identify terms, definitions, or experimental details as missing even though they are introduced elsewhere in the document. Providing every agent with the full paper could mitigate these errors but would substantially increase computational cost and API usage. A promising direction is therefore to develop lightweight mechanisms for document-wide grounding, such as shared summaries, global definitions, or structured representations of paper content that can be efficiently accessed by all review agents.

We are actively improving PaperMentor to address these limitations and enhance the quality of its feedback. We also welcome community contributions to extend the skill library and refine the system over time, enabling it to evolve alongside the writing practices and standards of the AI research community.

Ethical Considerations

All external papers used in our evaluation are publicly available preprints sourced from arXiv, downloaded solely for non-commercial research purposes in accordance with their respective licenses. Our user study follows the research ethics guidelines at ETH Zurich and is exempt from formal ethics review. All collected annotation data were stored securely and used exclusively for the evaluation reported in this paper.

Beyond study design, we acknowledge broader

ethical considerations in deploying AI-powered writing assistance. PaperMentor is intended to support junior researchers who lack access to experienced mentors, with the goal of reducing inequalities in scientific writing guidance across institutions and geographic regions. However, we caution that over-reliance on AI feedback could inadvertently homogenize writing styles or suppress diverse rhetorical voices in scientific communication. The system generates suggestions rather than rewrites, deliberately preserving authorial agency. We also recognize that the skill library, though distilled from expert guidance, reflects the norms and conventions of predominantly English-language, Western AI venues, and may not generalize equitably to researchers writing from different cultural or disciplinary backgrounds. We encourage ongoing community contributions to the skill library to mitigate these biases over time.

Acknowledgments

This material is based in part upon work supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039B; by the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645; and by the Canadian AI Safety Institute Research Program at CIFAR.

References

AAAI. 2026. AAAI launches AI-powered peer review assessment system. https://aaai.org/aaai-launches-ai-powered-peer-review-assessment-system/. Accessed 2026-02-27.

ACL. 2021. Ethics FAQ: How to write ethical considerations. Online guide.

Anthropic. 2025. Introducing claude opus 4.5. https://www.anthropic.com/news/claude-opus-4-5.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. In *Transactions of the Association for Computational Linguistics*, volume 6, pages 587–604.

Michael Black. Writing is laying out your logical thoughts. Twitter thread. Max Planck Institute Tuebingen.

Nicolas Bougie and Narimawa Watanabe. 2025. Generative reviewer agents: Scalable simulacra of peer review. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 98–116, Suzhou (China). Association for Computational Linguistics.

Jordan Boyd-Graber. Style. Online guide. University of Maryland.

Lele Cao, Lei You, and R&d Team. 2025. CSPaper review: Fast, rubric-faithful conference feedback. In *Proceedings of the 18th International Natural Language Generation Conference: System Demonstrations*, pages 3–7, Hanoi, Vietnam. Association for Computational Linguistics.

Eric Chamoun, Michael Schlichtkrull, and Andreas Vlachos. 2024. Automated focused feedback generation for scientific writing assistance. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9742–9763.

Mike D'Arcy, Tom Hope, Larry Birnbaum, and Doug Downey. 2024. Marg: Multi-agent review generation for scientific papers. *arXiv preprint arXiv:2401.04259*.

Paramveer S Dhillon, Somayeh Molaei, Jiaqi Li, Maximilian Golub, Shaochun Zheng, and Lionel Peter Robert. 2024. Shaping human-ai collaboration: Varied scaffolding levels in co-writing with language models. In *Proceedings of the 2024 CHI conference on human factors in computing systems*, pages 1–18.

Jason Eisner. 2010. How to write a paper? Online guide. Johns Hopkins University.

Maxwell Forbes. 2021. Figure creation tutorial: Making a figure 1. Online guide. University of Washington.

Xian Gao, Jiacheng Ruan, Zongyun Zhang, Jingsheng Gao, Ting Liu, and Yuzhuo Fu. 2025. Reviewagents: Bridging the gap between human and ai-generated paper reviews. *Preprint*, arXiv:2503.08506.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Grammarly. 2026. Grammarly. https://www.grammarly.com/. Accessed 2026-02-27.

Jieun Han, Haneul Yoo, Junho Myung, Minsun Kim, Hyunseung Lim, Yoonsu Kim, Tak Yeon Lee, Hwajung Hong, Juho Kim, So-Yeon Ahn, and 1 others. 2024. LLM-as-a-tutor in EFL writing education: Focusing on evaluation of student-llm interaction. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*, pages 284–293.

Jia-Bin Huang. 2023. How to write math in a paper? Twitter post. University of Maryland.

Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. 2024. Agentreview: Exploring peer review dynamics with llm agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1208–1226.

Zhijing Jin. 2021. Resources to help global equality for PhDs in NLP / AI. GitHub repository. Open resources and information for people to succeed in PhD in CS and career in AI/NLP, including writing suggestions from various professors.

Zhijing Jin. 2024. Nlp phd global equality: Writing suggestions from various professors. https://github.com/zhijing-jin/nlp-phd-global-equality. Max Planck Institute for Intelligent Systems.

Mina Lee, Katy Ilonka Gero, John Joon Young Chung, Simon Buckingham Shum, Vipul Raheja, Hua Shen, Subhashini Venugopalan, Thiemo Wambsganss, David Zhou, Emad A Alghamdi, and 1 others. 2024. A design space for intelligent and interactive writing assistants. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–35.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Yi Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, and 1 others. 2024. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *NEJM AI*, 1(8):AIoa2400196.

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

Chris Maddison. How to write an ML paper. Notion page. Step-by-step writing guide.

OpenAI. 2025. Introducing gpt-5.2.

OpenAI. 2026. Prism. https://prism.openai.com/. Accessed 2026-02-27.

Devi Parikh. Shortening papers to fit page limits. Medium blog post.

Simon Peyton Jones. 2014. How to write a great research paper. https://www.microsoft.com/en-us/research/video/how-to-write-a-great-research-paper-3/. Microsoft Research.

Simon Peyton Jones. 2014. How to write a great research paper: Seven simple suggestions. Slides and talk. Microsoft Research. Talk available at https://www.microsoft.com/en-us/research/video/how-to-write-a-great-research-paper-3/.

Tim Rocktäschel and Jakob Foerster. 2022. How to ML paper. Twitter post. UCL/DeepMind and University of Oxford.

Anna Rogers and Isabelle Augenstein. 2020. What can we do to improve peer review in NLP? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1256–1262. Association for Computational Linguistics.

Nihar B. Shah. 2022. Challenges, experiments, and computational solutions in peer review. *Communications of the ACM*, 65(6):76–87.

Stanford. 2026. Paperreview.ai. https://paperreview.ai/. Accessed 2026-02-27.

Pawin Taechoyotin, Guanchao Wang, Tong Zeng, Bradley Sides, and Daniel Acuna. 2024. Mamorx: Multi-agent multi-modal scientific review generation with external knowledge. In *Neurips 2024 Workshop Foundation Models for Science: Progress, Opportunities, and Challenges*.

Nitya Thakkar, Mert Yuksekgonul, Jake Silberg, Animesh Garg, Nanyun Peng, Fei Sha, Rose Yu, Carl Vondrick, and James Zou. 2025. Can llm feedback enhance review quality? a randomized study of 20k reviews at iclr 2025. *arXiv preprint arXiv:2504.09737*.

Xin Wang, Kaiwen Shi, and Carlos Oliver. 2025. Bioblobs: Unsupervised discovery of functional substructures for protein function prediction. *arXiv preprint arXiv:2510.01632*.

Jennifer Widom. 2006. Tips for writing technical papers. Online guide. Stanford University.

Shomir Wilson. Guide for scholarly writing. Penn State University.

Writefull. 2026. Writefull. https://writefull.com/. Accessed 2026-02-27.

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

Ruiyang Zhou, Lu Chen, and Kai Yu. 2024. Is llm a reliable reviewer? a comprehensive evaluation of llm on automatic paper reviewing tasks. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)*, pages 9340–9351.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. 2025. DeepReview: Improving LLM-based paper review with human-like deep thinking process. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29330–29355, Vienna, Austria. Association for Computational Linguistics.

Zhenzhen Zhuang, Jiandong Chen, Hongfeng Xu, Yuwen Jiang, and Jialiang Lin. 2025. Large language models for automated scholarly paper review: A survey. *Information Fusion*, 124:103332.

A Agent Configuration and Skill Assignment

PaperMentor instantiates twelve review agents per run: ten with fixed scope and two configured dynamically. Table 3 lists each agent, the text it reviews, and the expertise it draws from the skill library. Section agents review only their assigned section, supplemented by the abstract and introduction for context; global agents review the full merged source; and the figures agent reviews the extracted figure and table environments. The paper-type agent is grounded in the conventions of the type identified in Phase 1, and the venue agent in the requirements of the user-selected venue.

B Distribution of Comments and Annotation Scores Across Review Domains

Figure 4 presents the distribution of comments generated by PaperMentor alongside the corresponding human annotation scores across different review domains.

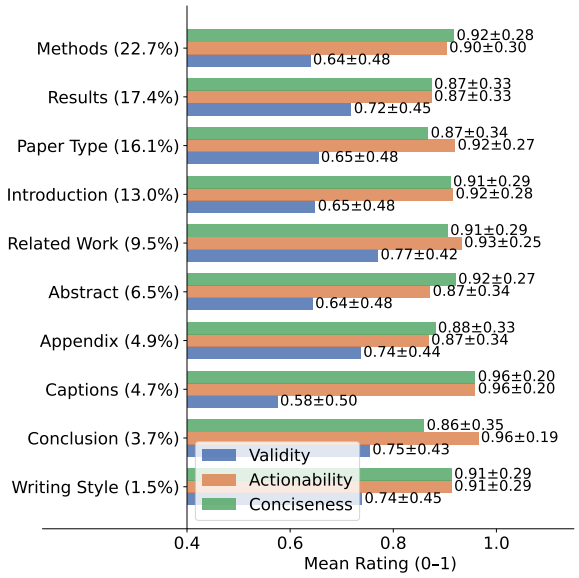

Figure 4: Distribution of generated comments across review domains, along with the mean human annotation scores for each domain on three evaluation metrics. Error bars indicate standard deviation.

C Section Level Comment Distribution vs. Text Length Distribution

Table 4 presents the percentage distribution of comments across the main sections, compared with the proportion of text length in each section. We observe that comments are generated more frequently in core sections of the paper, such as the Abstract and Methods. This indicates that PaperMentor prioritizes more important document content, while assigning relatively fewer comments to less critical sections such as the Appendix.

Agent	Class	Review Scope	Representative Skill Focus
Abstract	Section	Abstract	Abstract structure, specificity, and clarity
Introduction	Section	Introduction	Narrative framing, motivation, and contribution statements
Related Work	Section	Related work	Coverage, comparison and contrast, and citation conventions
Methods	Section	Methods, task formulation, preliminaries	Methodological clarity, task formulation, and mathematical notation
Results	Section	Experiments, findings, discussion	Results presentation and evidence support for claims
Conclusion	Section	Conclusion, limitations, ethics	Concise conclusions, limitations, and ethical considerations
Appendix	Section	Appendix	Organization of supplementary material
Writing Style	Global	Full document	Grammar, tone, formality, and terminology consistency
LaTeX & Formatting	Global	Full document	LaTeX conventions, cross-referencing, and table and equation formatting
Captions	Global	Figure and table environments	Caption quality
Paper Type	Dynamic	Full document	Writing conventions for the identified paper type
Venue	Dynamic	Full document	Formatting and content expectations of the target venue

Table 3: The twelve review agents in PaperMentor, their review scope, and a representative summary of the writing expertise each draws from the skill library. Ten agents have fixed scope; the two dynamic agents are configured at runtime from the identified paper type and selected target venue.

Category	Text (%)	Comments (%)
Abstract	2.5	8.4
Introduction	10.5	17.1
Related Work	5.4	12.3
Methods	11.3	24.5
Results	16.5	24.4
Conclusion	3.7	4.8
Appendix	49.9	8.4

Table 4: Percentage of total text length and percentage of total comments for each main section.