CSPaper Review: Fast, Rubric-Faithful Conference Feedback

Lele Cao[1], **Lei You**[1,2], **Kai Xie**[1], **Weiping Ding**[1], **Yong Du**[1], **Sven Salmonsson**[1], **Yumin Zhou**[1], **Vilhelm von Ehrenheim**[1]

[1]CSPaper @ Scholar7, [2]Technical University of Denmark

CSPaper Review (CSPR) is a free, AI-powered tool for rapid, conference-specific peer review in Computer Science (CS). Addressing the bottlenecks of slow, inconsistent, and generic feedback in existing solutions, CSPR leverages Large Language Models (LLMs) agents and tailored workflows to deliver realistic and actionable reviews within one minute. In merely four weeks, it served more than 7,000 unique users from 80 countries and processed over 15,000 reviews, highlighting a strong demand from the CS community. We present our architecture, design choices, benchmarks, user analytics and future road maps.

Correspondence: lele@scholar7.com
Date: October 2025

1 Why We Built It

Two pressing challenges have emerged in the fast-growing landscape of Computer Science (CS) research conferences, especially in AI and Machine Learning (ML). First, novice researchers often lack timely, targeted feedback tailored to their chosen conferences, with useful input arriving too late (typically after rejection) to guide meaningful revision. Second, the surge in submissions to top venues like ICML and NeurIPS has overwhelmed the traditional peer review system, leading to delays, inconsistent assessments, and declining review quality Kim et al. (2025); Guo et al. (2023); Naddaf (2025). As a result, reviewer capacity is stretched thin, compromising the depth and consistency of evaluations.

While Large Language Models (LLMs) are already quietly assisting with peer reviews – Liang et al. 2024a estimate that 6.5%∼16.9% of reviews at top AI conferences were ghostwritten or substantially revised by GPT-4 or alikes – the existing AI review tools fails to address the needs of paper authors representing a broader CS community.

CS stands out from other scientific disciplines in *three key ways* that makes it particularly suitable for AI-assisted reviewing. First, CS has evolved into a vast and fast-moving field where **conference publications dominate over journals** due to their strict timelines and rapid dissemination cycles. CS researchers therefore have a much stronger demand for **early feedback** to improve and iterate quickly. Second, CS conferences typically publish **well-defined and standardized review rubrics**, offering a natural scaffolding for aligning LLM-generated reviews with human expectations. This structured evaluation format is rare in other disciplines, making CS an ideal testbed for AI feedback. Third, the CS community is highly **active, decentralized, and open**, with an unmatched culture of preprints, open-source projects, and community-driven innovation. Fourth, some top-tier AI conference officially starts introducing AI-assisted review as a supplement to human reviewers AAAI (2026). This **strong communal foundation** is essential for "Human+AI" review systems.

However, existing tools such as Rigorous, WBS, GroundedAI, PaperWizard, and Hum fail to meet these CS-specific needs: they target journal workflows (not conference-style reviewing), take days to respond, lack rubric-aligned ratings, are prohibitively expensive and often tuned for non-CS domains like biology. To address this gap, we introduce CSPaper Review (CSPR), a **free** (up to 20 reviews per day) LLM-powered paper review system built from the ground up for CS researchers, with conference-specific evaluation criteria, fast turnaround, and integration into a researcher's early feedback loop.

2 How It Works

CSPR accepts either arXiv IDs/URLs or directly uploaded PDFs. Within **60 seconds**, the platform generates conference-specific reviews comprising three sections: *desk rejection assessment*, *expected review outcome*, and *critical reviewer ratings*.

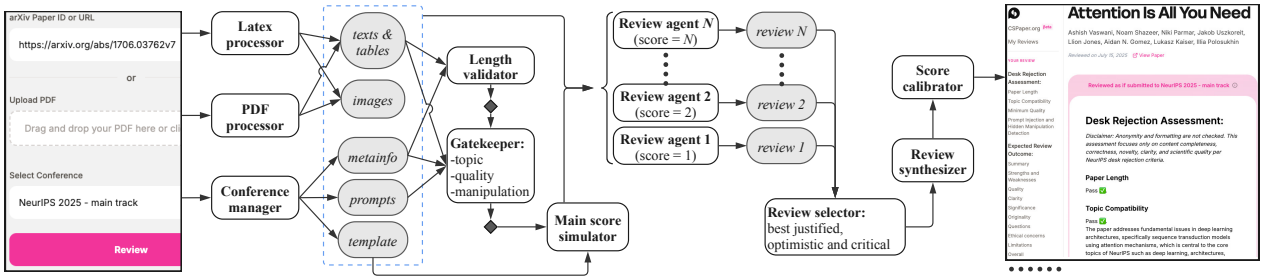

Figure 1: Input, output and workflow of CSPaper Review (CSPR). White boxes represent components or agents; gray boxes represent intermediate artifacts. Small diamonds indicate points where a failure in the preceding step terminates the entire workflow.

Latex/PDF processor: As depicted in Fig 1, dedicated processors extract text, tables, equations, and images from both LaTeX source packages and PDF files. For LaTeX inputs, the system performs downloading, main-tex resolution, consolidation of scattered tex files, and content cleaning. PDF inputs undergo OCR parsing, generating structured JSON content composed of markdown and images. Images are normalized to JPEG and downsampled (when excessively large) to ensure sensible LLM token consumption while keeping visual clarity.

Conference/track manager stores and retrieves conference-specific meta-information (name, track, year, call for papers, deadlines, etc.), review templates, and curated review prompts (with examples) tailored to individual conferences and tracks. It ensures generated reviews adhere strictly to the standards and expectations of selected venues.

Pre-review checks: Extracted artifacts are sequentially evaluated through a paper length validator and a set of gatekeepers verifying topic relevance, overall quality, and risk of prompt manipulation. Any failure at this stage immediately terminates the review process.

Review agents: For each valid rating/score level defined by the target conference (e.g., *1-strong reject* to *5-strong accept*), we **force** a dedicated agent to (concurrently) generate reviews that strictly justify the assigned score/rating. A review selector identifies three most realistic reviews: best justified, more optimistic, and more critical. They are synthesized into a coherent output primarily based on the best-justified review but selectively incorporating insights from the other two versions. Finally, a calibration step ensures coherence between overall and sub-dimensional scores (e.g., novelty, clarity), ensuring a well-aligned and balanced final review.

3 What We Found

LLM choice: We constructed a benchmark dataset of 100 papers by manually collecting reviews from OpenReview, official conference websites, and social media. We evaluated five LLMs on this benchmark. Mean Absolute Error (MAE) is calculated using the ground-truth overall scores as labels. The model with the lowest MAE (cf. Table 1 in Appendix) was selected as the serving LLM.

PDF parser: We qualitatively compared 4 PDF parsers (MinerU, Rigorous, Mistral, and LandingAI) on five CS papers with varied layouts. Mistral stood out with clean, structured JSON and highly accurate transcription of text, tables, equations, algorithms and images, while MinerU and Rigorous produced frequent, review-impacting errors. LandingAI showed similar quality to Mistral but is less viable due to pricing and speed.

Step-by-step vs. all-in-one prompting is a key question in LLM research; while step-by-step approaches are thought to enhance reasoning Yu (2024), explicit decomposition can sometimes harm performance Liu et al. (2024b). In our experiments, splitting each review agent into specialized sub-agents did not improve MAE, but increased token usage fivefold and latency over tenfold.

User analytics: Most traffic came from referral (44%), followed by direct (36%) and organic search (28%). Referral users were the most engaged, with a 3-minute average session and 48% of total page views. The number of arXiv and PDF review requests is largely equal. Among the 162 users who participated in our survey, 64% were undergraduate, graduate, or postgraduate students, consistent with the findings of Liang et al. (2024b). We identified three notable usage patterns: 1) the same paper reviewed across multiple conferences/tracks, likely to determine the most suitable submission venue; 2) different versions of a paper reviewed within the same conference/-track, suggesting iterative improvement of writing; and 3) one-time PDF review requests where filenames include real conference submission IDs, potentially indicating use of the tool for self-assessment during

the review process. Please refer to the Appendix for additional results and ethical discussions.

4 What's Next

CSPR has demonstrated real-world value in streamlining CS paper reviews, and our goal is to evolve it into both a practical tool and a research testbed for advancing human-AI collaboration in peer review. We aim to broaden coverage, enhance agent capabilities, develop interactive interfaces, and implement safeguards for trustworthy AI-assisted reviewing. Ultimately, we seek to benefit CS researchers while advancing the theory and practice of computational research assessment.

References

AAAI. AAAI-26 Main Technical Track: Call for Papers, 2026. URL https://aaai.org/conference/aaai/aaai-26/main-technical-track-call/. Accessed: 2025-08-25.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The Llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407, 2024.

Yanzhu Guo, Guokan Shang, Virgile Rennard, Michalis Vazirgiannis, and Chloé Clavel. Automatic analysis of substantiation in scientific peer reviews. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10198–10216. Association for Computational Linguistics, 2023.

Jaeho Kim, Yunseok Lee, and Seulki Lee. Position: The AI conference peer review crisis demands author feedback and reviewer rewards. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, et al. Monitoring ai-modified content at scale: A case study on the impact of chatgpt on ai conference peer reviews. In *International Conference on Machine Learning*, pages 29575–29620. PMLR, 2024a.

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

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024a.

Ryan Liu, Jiayi Geng, Addison J Wu, Ilia Sucholutsky, Tania Lombrozo, and Thomas L Griffiths. Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. In *Forty-second International Conference on Machine Learning*, 2024b.

Miryam Naddaf. AI is transforming peer review — and many scientists are worried. *Nature*, 639(8056):852–854, 2025.

Yijiong Yu. Do LLMs really think step-by-step in implicit reasoning? *arXiv preprint arXiv:2411.15862*, 2024.

Appendix of

CSPaper Review: Fast, Rubric-Faithful Conference Feedback

A More Figures and Tables

Figure 2, adopted from Google Analytics Dashboards, illustrates the distribution of unique users by country.

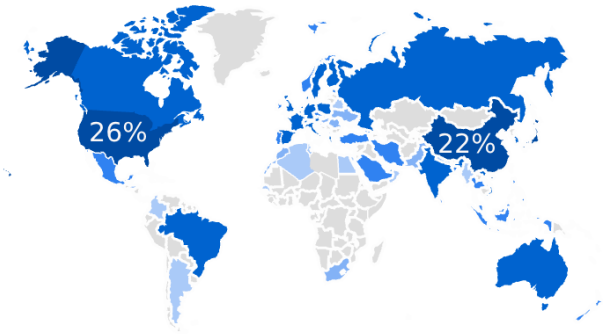

Figure 2: The geographical distribution of over 7,000 unique CSPR users from 80 countries.

Table 1 presents the mean absolute error (MAE) for five LLMs, GPT-4.1, GPT-o3, GPT-o4-mini, Deepseek-v3 Liu et al. (2024a), and Llama3-8b Dubey et al. (2024), evaluated across eight conferences, using a benchmarking dataset of 100 carefully selected research papers. We applied the following practices while constructing the dataset:

- For accepted papers, we did not randomly sample from all accepted works. Instead, we prioritized well-received papers such as spotlights, award-winning papers, or those that drove significant community discussion (e.g., on OpenReview, Alphaxiv and social media). These papers are generally considered exemplars in their respective venues and thus represent strong, trusted evaluation anchors.
- We acknowledge that rejected papers are generally harder to obtain, as reviews are often not made public. To address this, we relied on manual sourcing where possible, including data from conferences with open review processes (ICLR and partially NeurIPS) and from authors who are willing to share their rejected work and reviews. This ensured our negative examples came from verifiable, credible sources rather than arbitrary low-quality drafts.
- We explicitly identified cases where the final decision diverged from the average score or where reviewer opinions were highly polarized. In such cases, we asked established senior researchers (not involved in our team) to calibrate the scores, providing a more stable and reliable label for benchmarking. This step directly mitigates the concern that our benchmark might inherit inconsistencies from the review pool.
- Our benchmark dataset was deliberately balanced across multiple top-tier CS conferences and tracks to avoid bias toward a single venue's reviewing style or quality distribution. This diversity helps ensure the evaluation is not overfitted to one conference's reviewing idiosyncrasies.

Conference	GPT-4.1	o3	o4-mini	DS3	Llama3	GPT-5
AAAI	**0.044**	0.077	0.113	0.110	0.170	0.086
CVPR	**0.033**	0.100	0.100	0.082	0.150	0.067
EMNLP	**0.100**	0.160	0.180	0.170	0.210	0.120
ICLR	**0.120**	0.200	0.240	0.230	0.280	0.200
ICML	**0.092**	0.175	0.275	0.263	0.320	0.175
IJCAI	**0.100**	0.125	0.125	0.220	0.280	0.125
KDD	0.188	**0.125**	0.333	0.310	0.390	0.188
NeurIPS	**0.098**	0.131	0.348	0.333	0.395	0.131

Table 1: Benchmarking results to choose serving LLMs. "DS" denotes DeepSeek. Best results are highlighted in **bold**, and second-best results are underlined.

Figure 3 visualizes the distribution of user profiles among the 162 respondents to our questionnaire.

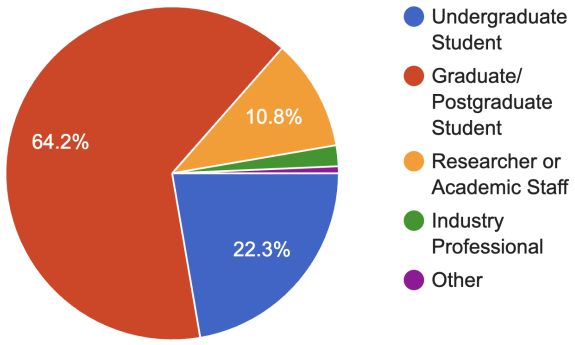

Figure 3: Percentage of user profiles from questionnaire.

Figure 4 presents the frequency of review activities reported by users in their daily work, as captured by the same questionnaire.

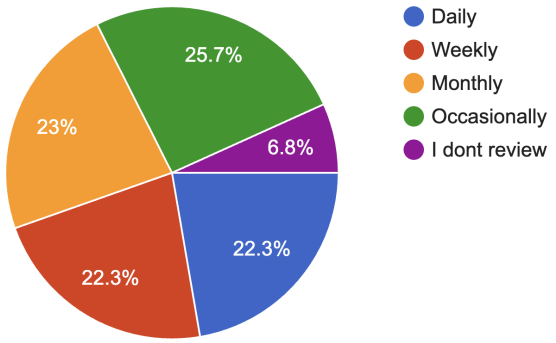

Figure 4: User review frequency from questionnaire.

B Data Handling and Ethical Considerations

CSPaper Review adheres to responsible data handling and transparency principles. All data collected and analyzed in this study (including uploaded PDF manuscript files, selected target conferences, and review preferences) were processed in accordance with our publicly available privacy policy.[1] Manuscript files are processed automatically using LLMs, under contractual agreements that explicitly prohibit the use of submitted content for model training or fine-tuning.

Uploaded files are temporarily stored on secure cloud infrastructure (e.g., Microsoft Azure) and are deleted either upon user request or after a defined expiration period. No user-submitted content is sold, shared, or publicly disclosed. On rare occasions, individual manuscripts may be reviewed internally for debugging and improvement, strictly under secure, privacy-preserving conditions.

User analytics presented in this paper (e.g., referral sources, usage patterns) are aggregated and fully anonymized. No personally identifiable information (PII)[2] is collected or disclosed. Observations such as filenames containing conference submission IDs (e.g., NeurIPS) were recorded passively and are not linked to individual users.

All users provide explicit consent to these practices when submitting their manuscripts. Only minimal cookies are used in a strict way.

C Acknowledgments

We are grateful to William Stoddart, Mathias Holst, and Sylvia Li for their support in the organizational, financial and operational matters.

We thank Orhan Uyaver, Wen Zhou, Filip Jasson and Rui Zhou for their early contributions in verifying the minimum viable product (MVP). We also appreciate Xiaolong Liu (Intel), Wenbing Huang (and his Lab in Renmin University), Yongfeng Zhang (and his Lab in Rutgers University), Heng Fang (KTH), Ye He (UCL), Sofiane Ennadir (KTH and Microsoft), Zineb Senane (Télécom Paris), Fangkai Yang (Microsoft), Valentin Buchner (University of Amsterdam), Tianze Wang (KTH and Microsoft) and Johannes F. Lutzeyer (Ecole Polytechique) for their valuable early evaluations in their respective research domains.

We are also grateful to Alexandra Stark and Tim Elgar from King (part of Microsoft) for reviewing this paper from communication and legal perspectives, respectively.

We thank the INLG 2025 conference reviewers for their constructive feedback, three detailed double-blind reviews and one meta-review, which significantly contributed to the improvement of this manuscript.

Finally, we sincerely acknowledge the feedback and encouragement from the broader CSPaper community, whose engagement has been invaluable to the development of this work.

[1] https://forum.cspaper.org/assets/uploads/review/privacy-policy.pdf

[2] https://en.wikipedia.org/wiki/Personal_data