GOODPOINT: Learning Constructive Scientific Paper Feedback from Author Responses

Jimin Mun [*1] **Chani Jung** [*2] **Xuhui Zhou** [1] **Hyunwoo Kim** [†3] **Maarten Sap** [†1]

Abstract

Rather than replacing scientific judgment, we study how LLMs can support researchers through *constructive feedback generation*: producing targeted, actionable feedback that helps improve scientific papers. We operationalize feedback quality using two author-centric signals derived from author responses: *validity*, whether authors acknowledge feedback as correct, and *actionability*, whether they commit to follow-up actions. We introduce GOODPOINT-ICLR, a dataset of 19K ICLR papers and review discussions annotated along these dimensions, and GOODPOINT, a training recipe that fine-tunes on successful feedback and further aligns models using real and synthetic preference pairs. Across 1.2K ICLR papers, a GOODPOINT-trained Qwen3-8B improves the predicted feedback success rate by 83.7% over its base model and achieves the strongest performance among size-comparable open models in matching human consensus feedback, surpassing Gemini-3-flash in precision. Human evaluation with paper authors further shows improvements in validity, actionability, specificity, and helpfulness.

1. Introduction

The growing use of LLMs across scientific research (Si et al., 2025; Huang et al., 2025; Chen et al., 2025; Ifargan et al., 2025; Schmidgall et al., 2025) has spurred proposals to automate core research processes, including peer review (Zhu et al., 2025; Zhou et al., 2024a; Liang et al., 2024). Yet unchecked automation risks eroding researchers' critical judgment and degrading the quality of science (Baumann et al., 2026; Wilder, 2025).

We argue that AI should instead empower researchers—particularly those who are traditionally disadvantaged (junior and/or non-native English speaking researchers; Liao et al., 2024)—rather than replace human scientific judgment. A high-value opportunity lies in using LLMs to deliver *constructive feedback* (Jefferson et al., 2002; Smith, 2006). This is especially consequential for authors who lack access to timely, expert feedback (Liao et al., 2024). However, generating such feedback for scientific papers is non-trivial: it demands deep domain knowledge, which is challenging even for human experts (Alberts et al., 2008) and unsurprisingly difficult for LLMs as well (Liang et al., 2024; Steiss et al., 2024).

Existing LLM-generated reviews reflect this difficulty, exhibiting weak specificity, limited actionability (Liang et al., 2024), and poor alignment with human reviewers in feedback prioritization (Steiss et al., 2024). Addressing these gaps requires confronting two fundamental challenges: (1) rigorously defining what constitutes constructive, effective feedback for improving scientific research, and (2) building training and evaluation frameworks that operationalize this definition. In this work, we tackle both through an author-centric lens, leveraging author responses in author–reviewer discussion as a natural signal to define, train, and evaluate feedback quality.

To address the first challenge, we introduce two author-centric axes of feedback quality: validity and author action (§3.1). As shown in Figure 1, validity captures whether authors agree with a raised concern, as invalid feedback tends to be rebutted. Author action captures whether authors propose concrete follow-up actions, noting that feedback can be valid yet not immediately actionable, still prompting meaningful discussion (e.g., suggestions deferred to future work).

Building on this definition, we operationalize constructiveness for both training and evaluation. We curate GOODPOINT-ICLR, which contains initial manuscript submissions and author–reviewer discussions for 19,534 ICLR papers (2020–2026), with each reviewer comment labeled for validity and author action (§3.2). We then introduce GOODPOINT, a training recipe for constructive review generation. We first fine-tune Qwen3-8B on successful

[*]Equal contribution. [†]Equal advising contribution. [1]Carnegie Mellon University [2]Independent Researcher [3]NVIDIA. Correspondence to: Jimin Mun <jmun@andrew.cmu.edu>, Chani Jung <1016chani@gmail.com>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

(i.e., valid and actionable) human reviews and then further align it via DPO on paired feedback data derived from both success labels and synthetic corruptions targeting key quality dimensions (§3). To evaluate LLM-generated feedback, we develop two complementary automatic evaluation frameworks—author response prediction and human consensus-based feedback evaluation—grounding on our definition of constructive feedback (§4.1, §4.2).

When evaluated on a held-out test set of 1,198 papers, GOODPOINT-trained Qwen3-8B achieves 83.7% improvement in predicted success rate compared to the base model. In feedback matching against multi-reviewer consensus feedback, GOODPOINT-SFT improves F1 by 58.8% over the base model and even exceeds Gemini-3-flash and GPT-5.2 in precision, suggesting that our model produces more selective critiques despite its substantially smaller scale. In an expert human study, GOODPOINT-DPO outperforms Qwen3-8B across validity, actionability, specificity, and helpfulness, meaningfully reducing the gap to Gemini-3-flash. This result demonstrates that our automated evaluation results translate into practical usefulness perceived by real-world authors. Together, these findings highlight the promise of training and evaluating LLMs grounded in an author-centric definition of constructive feedback, and underscore the potential of LLMs to augment—rather than replace—human researchers.

2. Related Works

LLMs are increasingly integrated into scientific workflows (Chen et al., 2025; Huang et al., 2025; Si et al., 2025), yet fully autonomous systems risks distorting scientific judgment (Baumann et al., 2026; Wilder, 2025). Instead, we advocate for human-AI collaboration. Although automated peer review has emerged to alleviate the burden of reviewers (Zhou et al., 2024b; Liang et al., 2024), current LLM-generated reviews often lack critical depth and tend to over-positivity (Zhou et al., 2024b; Idahl & Ahmadi, 2025). Existing approaches rely on specialized fine-tuning (Idahl & Ahmadi, 2025), review standardization (Yu et al., 2024), or literature retrieval (Zhu et al., 2025). However, unlike these systems that prioritize mimicking reviewer behavior, GOODPOINT focuses specifically on *constructive feedback*. By leveraging author responses as an explicit training signal, we prioritize feedback that is both valid and actionable from the researcher's perspective. See Appendix A for an extended discussion.

3. GOODPOINT: Constructive Feedback Learning Framework

We introduce GOODPOINT, a framework for generating author-centric paper feedback. Our approach builds on a dataset of ICLR 2020–2025 review discussions, from which we extract *successful* feedback items — those that authors both acknowledged as valid and committed to act upon — as the core training signal. We first fine-tune an LLM on this curated set via SFT to ground the model in constructive feedback patterns. We then sharpen both content and style through two DPO preference sets: one contrasting valid versus invalid feedback, and another using targeted corruptions along five dimensions — specificity, clarity, accuracy, prioritization, and supportive tone — as rejected samples.

3.1. GOODPOINT Definition

To operationalize our goal of generating constructive feedback, we use author responses to define a measurable signal of feedback success. Concretely, given a paper x and a set of self-contained feedback items $F = \{f_1, f_2, \ldots, f_n\}$ with corresponding author responses $R = \{r_1, r_2, \ldots, r_n\}$, we define a feedback item f_i as **successful**, $\mathbb{1}_{\text{success}}(f_i, r_i) = 1$, if the author acknowledges it as both *valid*, $\mathcal{V}(f_i, r_i) = 1$, and *actionable*, $\mathcal{A}(f_i, r_i) = 1$ (examples in Figure 1).

3.2. GOODPOINT-ICLR

We construct GOODPOINT-ICLR, a dataset of 18.9K ICLR papers and reviewer discussions spanning 2020–2026. Reviewer comments are decomposed into self-contained feedback units paired with corresponding author responses, from which we derive two labels: *validity*, whether authors acknowledge the feedback as correct, and *actionability*, whether authors indicate concrete follow-up actions. We automatically parse review discussions into feedback-response pairs using GPT-4.1 and verify the parsing quality through human annotation, achieving high accuracy for both validity and actionability extraction. Full dataset construction details are provided in Appendix B.

Feedback Corruption for Preference Pair Construction To construct informative preference pairs for DPO, we apply targeted quality corruptions (Li et al., 2025; Geng et al., 2025) to successful reviewer-written feedback. Following prior work on effective feedback (Steiss et al., 2024), we generate synthetic degradations targeting specificity, clarity/actionability, accuracy, essentialness, and supportive tone, while preserving other dimensions. We validate corruption quality using an LLM-as-a-judge framework and retain only high-quality preference pairs. Details of corruption generation and verification are provided in Appendix C.

3.3. Training

To train a constructive feedback generation model, we use both supervised fine-tuning (SFT) and direct preference optimization (DPO). We first ground the model's feedback generation through SFT on *successful* (valid and actionable)

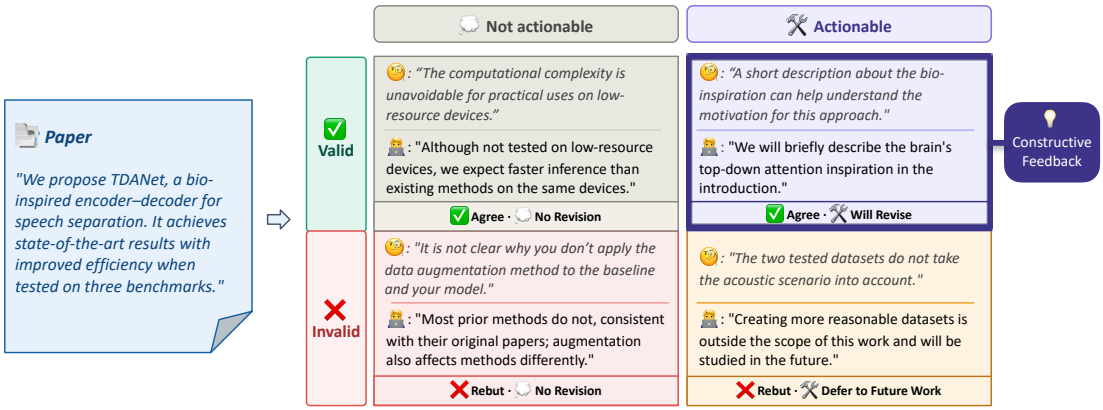

Figure 1. Overview of the two feedback quality axes and our definition of *constructive feedback*. Reviewer feedback is evaluated along two dimensions: validity—whether the raised concern is agreed upon by the authors—and actionability—whether it prompts concrete follow-up actions. Feedback that is both valid and actionable is considered successful; we build training and evaluation frameworks that operationalize this definition.

feedback. Then we use DPO to further align model to generate effective feedback in both content and style. We use Qwen3-8B as our base model.

GOODPOINT-SFT We fine-tune a base LLM on the valid and actionable feedback units from GOODPOINT-ICLR (6K papers), formatted as (paper, aggregated feedback) pairs. We aggregate all feedback given by a single reviewer per paper for the model to learn a consistent and diverse style. This grounds the model in constructive, author-acknowledged feedback patterns before preference optimization.

GOODPOINT-DPO We further align the SFT model using two preference datasets. The first contrasts valid versus invalid feedback items drawn directly from GOODPOINT-ICLR (9K papers). The second uses our corruption pipeline to form pairs of original versus corrupted feedback, targeting the five quality dimensions described above. We construct DPO samples such that the acceptance ratio is higher in the chosen set compared to rejected at least by 2 feedback items. In this phase of training, we aggregate all the reviews for one paper for comprehensive coverage of feedback and deduplicate items across reviewers using embedding similarity. Training details and hyperparameters are provided in Appendix D.

4. Experiment Setup & Results

Datasets Across experiments, we use a test set of 1,198 papers where the number of accepted and rejected papers are balanced. To build the test set, we combine 600 papers from ICLR 2020-2025 and 598 from ICLR 2026, to address possible contamination. In human consensus-based feedback evaluation, we use 943 papers from the test set, excluding the papers without any human consensus feedback. For human evaluation, we use author-provided manuscripts.

Baselines We compare our models with four baseline models: Llama 3.1-8b-Instruct, Qwen3-8B, GPT-5.2, and Gemini-3-flash. We additionally evaluate Qwen3-32B and DeepReviewer-7B-Fast on the held-out ICLR 2026 partition (DeepReviewer overlaps with our main test set); results are in Appendix E.2.5. In human evaluation, we compare one proprietary model, Gemini-3-flash, and the strongest size-comparable open model, Qwen3-8B, to GOODPOINT-DPO. The prompts and settings for generation are in Appendix F.

Subsampling for Length Normalization To mitigate potential biases arising from the number of feedback units generated by different LLMs, we randomly subsample 5 feedback units from each model's total output for a given paper before evaluating their quality. In automatic evaluations, we report the mean and 95% confidence intervals calculated over $B = 1,000$ bootstrap iterations. See Appendix G for a full analysis of results and discussion.

4.1. Author Response Prediction

To assess author-centric feedback quality, we use *author response prediction* as a proxy for feedback success. Given a paper and generated feedback, a separate LLM-fine-tuned evaluator predicts validity and author action. To improve robustness against model-specific failure modes (e.g., hallucinated or weakly grounded critiques), we additionally apply quality filtering based on accuracy, paper-specific grounding, constructive tone, and prioritization (Steiss et al., 2024). See Appendix E.1 for implementation details.

Table 1 reports combined success rates (validity and action prediction with quality score filtering). GPT-5.2 achieves the highest success rate ($45.8\% \pm 1.0\%$), followed by Gemini-

Table 1. Pass rates ($B = 1000$, 5 feedback/paper) across three scoring configurations: *Combined Success* uses both validity and author action. CI half-widths are from 95% confidence intervals; Δ for GOODPOINT models is relative to Qwen3-8b (Base).

Model	Combined Success (%)
Gemini-3-flash	37.9±0.9
GPT-5.2	**45.8**±1.0
Llama3.1-8b-Inst.	1.8±0.3
Qwen3-8b (Base)	8.0±0.6
GOODPOINT-DPO	14.7(+6.7)±0.5
GOODPOINT-SFT	9.2(+1.2)±0.5

Table 2. Model performances in matching human consensus feedback ($B = 1000$, 5 feedback/paper). CI half-widths computed from 95% confidence intervals for all values are less than 0.001. Among the size-comparable models (7B/8B), GOODPOINT-SFT achieves the best performance, followed by GOODPOINT-DPO, across all metrics. Notably, GOODPOINT-SFT achieves higher precision than the larger proprietary models.

Model	Precision	Recall	F1
Gemini-3-flash	0.128	**0.169**	**0.131**
GPT-5.2	0.130	0.165	0.130
Llama3.1-8b-Inst.	0.047	0.053	0.044
Qwen3-8b (Base)	0.069	0.084	0.068
GOODPOINT-DPO	0.093(+0.024)	0.107 (+0.023)	0.087(+0.019)
GOODPOINT-SFT	**0.138**(+0.069)	0.112(+0.028)	0.108(+0.040)

3-flash ($37.9\% \pm 0.9\%$). GOODPOINT-DPO and GOODPOINT-SFT outperform Qwen3-8B ($8.0\% \pm 0.6\%$) by 6.7% and 1.2%. These results indicate that GOODPOINT training meaningfully improves feedback quality.

4.2. Alignment with Human Consensus Feedback

We evaluate whether models generate *important* feedback using *human consensus feedback* as a proxy. We define human consensus feedback as successful feedback (i.e., valid and actionable) independently raised by multiple human reviewers. We then measure how well model-generated feedback aligns with this consensus set using precision, recall, and F1, capturing important critiques. Feedback matching is automatically assessed, with implementation details provided in Appendix E.2.1.

Table 2 presents the performance of GOODPOINT models and baselines in matching human consensus feedback. Notably, GOODPOINT-SFT achieves higher precision than the larger proprietary models, Gemini-3-flash (0.128) and GPT-5.2 (0.130). This suggests that GOODPOINT produces more selective and high-fidelity critiques with fewer low-quality comments, despite its substantially smaller scale. GOODPOINT-DPO follows closely, likewise outperforming all similarly sized baselines across every metric. These findings remain consistent across the two test partitions, including the temporally held-out ICLR 2026 set (Appendix E.2.5).

Table 3. Human evaluation at the feedback level (n per model shown). Superscripts denote significant pairwise differences: † vs. GOODPOINT-DPO, ‡ vs. Qwen3-8B. Ordinal metrics (Specificity, Helpfulness) use Kruskal–Wallis with Mann–Whitney U; binary metrics (Validity, Actionability) use χ^2 with Fisher's exact pairwise tests.

Model	Valid. %	Action. %	Specificity (SD)	Helpfulness (SD)	n
Gemini-3-flash	$72.3^{\dagger\ddagger}$	$56.9^{\ddagger}$	$4.42_{(0.86)}{}^{\dagger\ddagger}$	$3.40_{(1.40)}{}^{\dagger\ddagger}$	65
GOODPOINT-DPO	58.1	40.3	$3.50_{(1.16)}{}^{\ddagger}$	$2.77_{(1.29)}{}^{\ddagger}$	62
Qwen3-8B	41.5	32.3	$2.89_{(1.28)}$	$2.25_{(1.24)}$	65

4.3. Human Evaluation

We conduct a user study with $N = 13$ authors evaluating feedback generated on their own paper submissions. Authors assess feedback at the item level across four dimensions: validity, actionability, specificity, and helpfulness, following prior work (Sadallah et al., 2025; Liang et al., 2024). Additional study details are provided in Appendix E.3.

Table 3 summarizes the human evaluation results. Gemini-3-flash achieves the strongest performance in all four dimensions, and GOODPOINT-DPO ranks second overall, outperforming Qwen3-8B in all dimensions: validity (58.1% vs. 41.5%), actionability (40.3% vs. 32.3%), specificity (3.50 vs. 2.89), and helpfulness (2.77 vs. 2.25). These results indicate that GOODPOINT-DPO meaningfully improves author-perceived feedback quality, narrowing the gap to the strongest proprietary baseline and confirming that automatic evaluation gains translate into practical usefulness.

5. Conclusion

We present GOODPOINT, a framework for generating constructive, author-centric feedback by leveraging signals from real-world author responses. By defining feedback quality along the axes of validity and actionability, we developed both a large-scale dataset and a training strategy that aligns LLM outputs with feedback that authors acknowledge and act upon. Experiment results on both automatic metrics and human evaluation show that our approach delivers high-quality, precise feedback, even enabling small models to rival much larger ones. Our findings highlight the importance of grounding LLMs in human-centric signals. See Appendix H for limitations.

References

Alberts, B., Hanson, B., and Kelner, K. L. Reviewing peer review, 2008.

Baumann, J., Pei, J., Koyejo, S., and Hovy, D. Stop automating peer review without rigorous evaluation. In *Post-AGI Science and Society Workshop*, 2026.

Byrt, T., Bishop, J., and Carlin, J. B. Bias, prevalence and kappa. *Journal of clinical epidemiology*, 46(5):423–429, 1993.

Chen, Q., Yang, M., Qin, L., Liu, J., Yan, Z., Guan, J., Peng, D., Ji, Y., Li, H., Hu, M., et al. Ai4research: A survey of artificial intelligence for scientific research. *arXiv preprint arXiv:2507.01903*, 2025.

Geng, S., Ivison, H., Li, C.-L., Sap, M., Li, J., Krishna, R., and Koh, P. W. The delta learning hypothesis: Preference tuning on weak data can yield strong gains. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=9rwtezthwo.

Hattie, J. and Timperley, H. The power of feedback. *Review of educational research*, 77(1):81–112, 2007.

Hu, J., Wu, X., Zhu, Z., Xianyu, Wang, W., Zhang, D., and Cao, Y. Openrlhf: An easy-to-use, scalable and high-performance rlhf framework. *arXiv preprint arXiv:2405.11143*, 2024.

Huang, Y., Chen, Y., Zhang, H., Li, K., Zhou, H., Fang, M., Yang, L., Li, X., Shang, L., Xu, S., et al. Deep research agents: A systematic examination and roadmap. *arXiv preprint arXiv:2506.18096*, 2025.

Idahl, M. and Ahmadi, Z. OpenReviewer: A specialized large language model for generating critical scientific paper reviews. In Dziri, N., Ren, S. X., and Diao, S. (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pp. 550–562, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-191-9. doi: 10.18653/v1/2025.naacl-demo.44. URL https://aclanthology.org/2025.naacl-demo.44/.

Ifargan, T., Hafner, L., Kern, M., Alcalay, O., and Kishony, R. Autonomous llm-driven research—from data to human-verifiable research papers. *NEJM AI*, 2(1): AIoa2400555, 2025.

Jefferson, T., Alderson, P., Wager, E., and Davidoff, F. Effects of editorial peer review: a systematic review. *Jama*, 287(21):2784–2786, 2002.

Li, S. S., Mun, J., Brahman, F., Hosseini, P., Thomas, B. G., Sin, J. M., Ren, B., Ilgen, J. S., Tsvetkov, Y., and Sap, M. ALFA: Aligning LLMs to ask good questions a case study in clinical reasoning. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=12u7diwku0.

Liang, W., Zhang, Y., Cao, H., Wang, B., Ding, D. Y., Yang, X., Vodrahalli, K., He, S., Smith, D. S., Yin, Y., et al. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *NEJM AI*, 1(8):AIoa2400196, 2024.

Liao, Z., Antoniak, M., Cheong, I., Cheng, E. Y.-Y., Lee, A.-H., Lo, K., Chang, J. C., and Zhang, A. X. Llms as research tools: A large scale survey of researchers' usage and perceptions. *arXiv preprint arXiv:2411.05025*, 2024.

Lu, S., Kuznetsov, I., and Gurevych, I. Identifying aspects in peer reviews. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 6145–6167, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.326. URL https://aclanthology.org/2025.findings-emnlp.326/.

Sadallah, A., Baumgärtner, T., Gurevych, I., and Briscoe, T. The good, the bad and the constructive: Automatically measuring peer review's utility for authors. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 28991–29021, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1476. URL https://aclanthology.org/2025.emnlp-main.1476/.

Schmidgall, S., Su, Y., Wang, Z., Sun, X., Wu, J., Yu, X., Liu, J., Liu, Z., and Barsoum, E. Agent laboratory: Using llm agents as research assistants. *arXiv preprint arXiv:2501.04227*, 2025.

Si, C., Hashimoto, T., and Yang, D. The ideation-execution gap: Execution outcomes of llm-generated versus human research ideas. *arXiv preprint arXiv:2506.20803*, 2025.

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

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., and Olson, C. B. Comparing the quality of human and chatgpt feedback of students' writing. *Learning and Instruction*, 91:101894, 2024.

Wilder, B. Equilibrium effects of llm reviewing, May 2025. URL https://bryanwilder.github.io/files/llmreviews.html.

Yu, J., Ding, Z., Tan, J., Luo, K., Weng, Z., Gong, C., Zeng, L., Cui, R., Han, C., Sun, Q., Wu, Z., Lan, Y., and Li, X. Automated peer reviewing in paper SEA: Standardization, evaluation, and analysis. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10164–10184, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.595. URL https://aclanthology.org/2024.findings-emnlp.595/.

Zhang, D., Bao, Z., Du, S., Zhao, Z., Zhang, K., Bao, D., and Yang, Y. Re ^{2}: A consistency-ensured dataset for full-stage peer review and multi-turn rebuttal discussions. *arXiv preprint arXiv:2505.07920*, 2025.

Zhou, R., Chen, L., and Yu, K. Is LLM a reliable reviewer? a comprehensive evaluation of LLM on automatic paper reviewing tasks. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N. (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 9340–9351, Torino, Italia, May 2024a. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.816/.

Zhou, R., Chen, L., and Yu, K. 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)*, pp. 9340–9351, 2024b.

Zhu, M., Weng, Y., Yang, L., and Zhang, Y. Deepreview: Improving llm-based paper review with human-like deep thinking process. *arXiv preprint arXiv:2503.08569*, 2025.

A. Extended Related Works

LLMs in Scientific Research LLMs are increasingly used in scientific workflows, including literature synthesis, hypothesis generation, and manuscript writing (Chen et al., 2025; Huang et al., 2025; Si et al., 2025). However, the growing push for fully autonomous research systems (Ifargan et al., 2025; Schmidgall et al., 2025) may distort scientific judgment and create harmful incentive shifts in the research ecosystem (Baumann et al., 2026; Wilder, 2025). We advocate for an augmentation approach, where LLMs collaborate with researchers, enhancing their work without replacing their critical judgment.

Automated Peer Review Automated peer review has gained interest as a solution to reviewer burden and inconsistency (Zhou et al., 2024b; Liang et al., 2024; Idahl & Ahmadi, 2025). However, recent studies reveal limitations of LLM-generated peer reviews, including their struggle with score prediction and providing human-level critical feedback (Zhou et al., 2024b) and being overly positive (Idahl & Ahmadi, 2025). These works typically evaluate quality using reviewer-centric metrics like aspect coverage and score prediction. Existing peer review generation methods include fine-tuning language models on expert reviews (Idahl & Ahmadi, 2025), standardizing multiple reviews for supervision (Yu et al., 2024), and using structured analysis and literature retrieval for evidence-based reviewing (Zhu et al., 2025). Our focus is on generating constructive feedback to improve papers, rather than evaluating them for acceptance. Unlike prior systems that rely on fixed criteria or full human reviews, GOODPOINT uses authors' feedback acknowledgments as a training signal for author-centric feedback generation.

Table 4. Inter-annotator agreement for validity and author action labels ($n = 111$ units). PABAK is preferred over Cohen's κ due to high label skew (Byrt et al., 1993).

Label	Observed Agr.	PABAK	Cohen's κ	n
Validity	0.874	0.748	0.000	111
Author Action	0.919	0.838	0.271	111

B. GOODPOINT-ICLR Dataset Construction

B.1. Data Collection and Splits

We construct GOODPOINT-ICLR from ICLR submissions spanning 2020–2026. Papers are sourced from three outlets: Re^2 (Zhang et al., 2025) for 2020–2023, arXiv via title matching for 2024–2025, and direct PDF submissions to OpenReview for 2026. We convert all papers into markdown format using Marker.[1]

Each submission is paired with its corresponding OpenReview discussion thread. We partition the data so that ICLR 2020–2025 serves as the training, development, and test split, while ICLR 2026 is held out as a temporally separate test set, minimizing potential contamination, since all baseline models have knowledge cutoffs prior to August 2025. The dataset contains 18,936 papers from ICLR 2020–2025 (14,517 rejected, 4,419 accepted). An additional 598 ICLR 2026 papers are held out as a temporally separate test set, bringing the total to 19,534 papers spanning 2020–2026.

B.2. Review Parsing and Annotation

To parse human reviews into self-contained feedback units paired with author responses, we use GPT-4.1 (`gpt-4.1-2025-04-14`) with temperature 0.7 and a maximum generation length of 10,240 tokens. We annotate validity, author action, feedback dimensions (Hattie & Timperley, 2007), and key aspects following Liang et al. (2024). The user prompt below summarizes the extracted fields and label options.

System Prompt

```
You are a helpful assistant, an expert at analyzing academic peer review responses
and extracting feedback.
```

[1] https://github.com/datalab-to/marker

User Prompt

You will receive a conversation between *reviewers* and *authors* for an academic paper. Your task is to **parse each reviewer feedback item** into a **self-contained unit** and evaluate the **corresponding author response**. Ignore purely positive feedback with no actionable critique, question, or suggestion. For **each feedback unit**, extract and annotate the following fields:

1. **feedback_text** --- Rewrite the reviewer feedback as a clear, standalone unit, preserving the original text as closely as possible.
2. **author_response_text** --- Rewrite the author response as a clear, standalone unit, preserving tone and first-person pronouns.
3. **validity** --- Did the authors agree the feedback is valid? agreed_by_authors | rebutted_by_authors | unclear
4. **author_action** --- What did the authors commit to do? will_revise | defer_future_work | point_to_existing_content | no_revision_accept | no_revision_contest | no_action_other | unclear_or_no_response
5. **dimensions** --- Break down the feedback into three dimensions if present: *Feed Up* (goal), *Feed Back* (gap), *Feed Forward* (next steps).
6. **aspects** --- Identify emphasized aspects (e.g., Clarity and Presentation, Reproducibility, Novelty, Theoretical Soundness, *etc.*)

Return output strictly as JSON following the prescribed schema. Full decision rules, label definitions, and output schema are omitted here for brevity.

Input: {conversation_text}

C. Feedback Corruption & Verification

Feedback corruption We generate corrupted feedback to construct informative preference pairs for DPO. Using GPT-5 mini (`gpt-5-mini-2025-08-07`; medium reasoning, medium verbosity, maximum generation length of 4096 tokens), we rewrite successful reviewer feedback by degrading exactly one quality dimension at a time while preserving the others.

System Prompt

The model acts as a text transformation assistant that rewrites reviewer feedback by deliberately corrupting exactly one quality dimension at a time, while preserving all others.

User Prompt

Inputs: Paper title, abstract, and original reviewer feedback.

Given these inputs, the model produces five rewrites of the feedback, each degrading a single dimension:

Corruption	Description
Generic	Remove all paper-specific details so the feedback reads as boilerplate applicable to any paper in the field.
Vague	Retain paper-specific references but strip concrete questions, examples, and actionable guidance.
Inaccurate	Introduce plausible but factually wrong claims about the paper's content, methods, or results.
Non-essential	Shift focus from core fixable issues to peripheral, out-of-scope, or stylistic concerns.
Unsupportive	Replace constructive, hedged language with blunt, dismissive, or commanding phrasing.

Output is a JSON object with five keys (generic, vague, inaccurate, nonessential, unsupportive), each containing the full rewritten feedback as a string.

Table 5. Corruption quality verification results by dimension. Accuracy measures whether the verifier correctly identifies the targeted dimension. Degradation and collateral preservation scores are on a 1–3 scale.

Dimension	Accuracy	Target Degradation	Collateral Preservation
Generic	0.625	1.83	2.92
Vague	0.945	1.86	2.98
Inaccurate	0.992	2.98	2.48
Non-essential	0.969	3.00	2.56
Unsupportive	0.982	2.68	2.88

Automatic verification of corrupted feedback In addition to feedback corruption, to ensure the quality of corrupted samples, we used LLM-as-a-judge for verification. We remove any samples with erroneous corrupted dimension prediction and corruption and collateral preservation score below 2. Verification results, prior to filtering, are shown in Table 5.

System Prompt

```
The model acts as a judge of AI-generated text transformations in academic peer
review, identifying which quality dimension has been corrupted and evaluating the
quality of the corruption.
```

User Prompt

Inputs: Paper title, abstract, original reviewer feedback, and a randomized list of rewritten feedback variants (corruption dimensions unknown).

For each rewrite, the model predicts the targeted corruption dimension and scores it on two criteria:

Score	Description
Target Degradation (1--3)	How clearly the intended dimension is degraded (3 = clearly degraded, 1 = barely degraded).
Collateral Preservation (1--3)	How well all other dimensions are preserved (3 = fully preserved, 1 = multiple dimensions affected).

Output is a JSON object with a results array ordered by rewrite index, each entry containing: rewrite_index, predicted_dimension, target_degradation_score, collateral_preservation_score, and reasoning.

Human verification of corrupted feedback To further validate the quality of our automated corruption pipeline, we conducted human annotation on a post-filtered sample of $N = 297$ corrupted feedback items across 20 papers. Two authors independently assessed whether each item correctly instantiated the intended corruption dimension, given the original feedback and the source paper for reference. Overall correctness reached 93.9%, with four of five dimensions achieving $\geq$87.9% and three achieving 100%. The lower rate for *vague* (71.4%) reflects inherent annotation difficulty at this dimension boundary, resolved by post-discussion consensus. PABAK scores confirm strong inter-annotator reliability overall. Results are shown in Table 6.

Table 6. Human verification of corrupted feedback items (post-automated filtering, $N = 297$ across 20 papers). PABAK is omitted for *vague* due to full post-discussion agreement.

Dimension	N	% Correct	Agreement	PABAK
Generic	33	87.9%	0.879	0.758
Vague	49	71.4%	—	—
Inaccurate	70	100.0%	1.000	1.000
Non-essential	71	100.0%	1.000	1.000
Unsupportive	74	100.0%	0.986	0.973
Overall	**297**	**93.9%**	**0.929**	**0.859**

D. Training Setup

GOODPOINT-SFT is fine-tuned from `Qwen3-8B` and GOODPOINT-DPO is subsequently initialized from GOODPOINT-SFT, both trained using OpenRLHF (Hu et al., 2024) with the following configurations:

Table 7. SFT and DPO training hyperparameters.

Hyperparameter	SFT	DPO
Base model	Qwen3-8B	GOODPOINT-SFT
Max sequence length	30,000	30,000
Train batch size	128	128
Micro batch size	8	4
Learning rate	5×10^{-6}	5×10^{-6}
Epochs	1	1
Precision	BF16	BF16
β (KL penalty)	—	0.1
NLL loss coefficient	—	0.2

For GPU optimization, we use ZeRO stage 3 with flash-attention-2 with ring attention size 8 with head stride 2. We also early stop DPO at step 50 to prevent overfitting.

For GOODPOINT-DPO, we also use cosine similarities of embeddings using OpenAI's `text-embedding-3-small` to deduplicate multiple overlapping reviewer feedback. We randomly select among the matched instances if the similarity is greater than 0.5.

E. Evaluation Details

E.1. Author Response Prediction Training & Results

Table 8. Author response prediction results on a held-out test set of 1,000 samples (864 valid, 136 invalid).

Task	Precision	Recall	F1	Accuracy
Validity	0.815	0.717	0.754	0.717
Action	0.779	0.706	0.741	0.650

E.1.1. AUTHOR RESPONSE PREDICTOR TRAINING

Author response predictor is trained on 3K papers and their corresponding feedback items, author validity, action, and response. We balance valid (agreed) and invalid (rebutted) feedback items to address label imbalance in the dataset. Due to the lack of invalid feedback samples, we supplement our 3K paper set with filtered out unused invalid feedback samples from SFT data. We also used OpenRLHF (Hu et al., 2024) for training with the following hyperparameters: a maximum sequence length of 30,000 tokens, a train batch size of 32 (micro batch size of 1), and a learning rate of 5×10^{-6} for 1 epoch in BF16 precision. Distributed training uses ZeRO stage 3 with FlashAttention-2, ring attention of size 8 (head stride 2), sample packing, gradient checkpointing, and Adam offload. Table 8 shows full performance metric over 1000 test samples with 864 valid and 136 invalid samples.

E.1.2. QUALITY DIMENSION SCORING SETUP

We obtain quality scores for feedback units using GPT-5 mini (`gpt-5-mini-2025-08-07`), with medium verbosity, medium reasoning effort, and a maximum generation length of 4096 tokens. We use the following prompts (truncated and summarized for space):

System Prompt

```
The model acts as an expert meta-reviewer evaluating peer-review feedback
impartially and evidence-based, citing specific phrases to justify every score.
```

User Prompt

```
Given a paper excerpt, venue, and reviewer feedback, the model scores feedback
across five dimensions (1--5 each):
```

Dimension	Description
Accuracy	Are claims about the paper's content, methods, and results factually correct?
Prioritisation	Does feedback focus on high-impact scientific issues (validity, novelty, reproducibility) over cosmetic ones?
Constructive Tone	Is feedback collegial, respectful, and improvement-oriented?
Paper-Specific Grounding	Is feedback anchored to named components, results, and claims of this paper vs. generic boilerplate?
Actionability	Are revision directions specific, feasible, and tied to concrete sections or analyses?

```
Output is a structured JSON object containing per-dimension scores, justifications,
and supporting evidence, plus an overall summary with average score, strengths,
weaknesses, and holistic judgment.
```

Figure 2 shows the distribution of quality scores across models and human reviewers. We observe that LLM-generated feedback tends to receive lower scores in paper-specific grounding, prioritization, and, to a lesser extent, accuracy. To mitigate these model-specific failure modes, we apply quality filtering before author response prediction.

To determine filtering thresholds, we score 2K human-reviewer feedback examples balanced across valid and invalid instances. For each quality dimension, we use the average human-reviewer score as the acceptance threshold for model-generated feedback. In the main experiments, we apply thresholds for accuracy, paper-specific grounding, constructive tone, and prioritization; actionability is excluded because it is already captured by author response prediction. The resulting human-reviewer averages are 4.37 for accuracy, 4.44 for prioritisation, 4.62 for constructive tone, 4.36 for paper-specific grounding, and 3.49 for actionability.

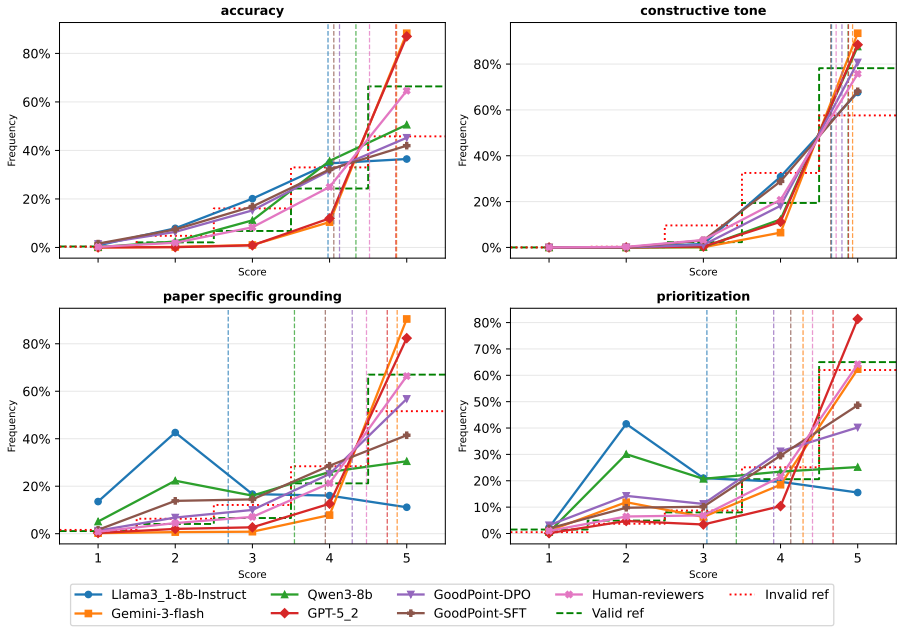

Figure 2. LLM-judged feedback quality score distributions for baseline models, GOODPOINT models, and human reviewers on test set. Valid ref and Invalid ref indicate scores on 1000 valid and invalid feedback units respectively. All dimensions are scored on a Likert scale from 1 to 5. Dotted vertical lines indicate mean scores.

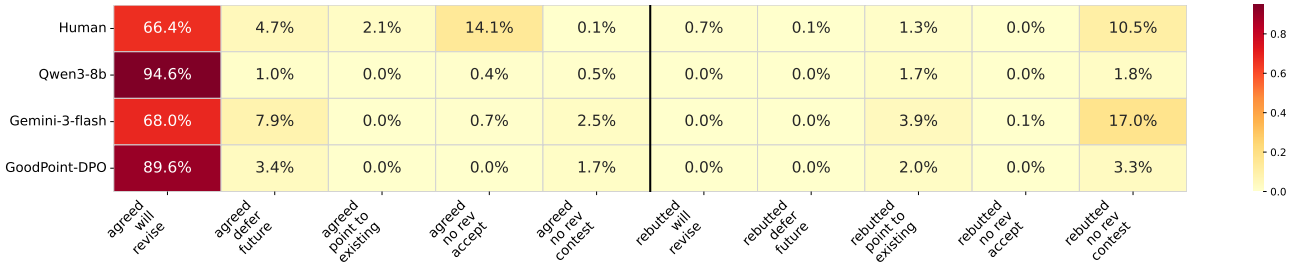

Figure 3. Human reviewer feedback author response distribution from entire GOODPOINT-ICLR. The heatmap shows percentage taken over all feedback. We see the highest concentration of author agreed with committed revisions.

Human verification of quality-dimension scoring To validate the reliability of LLM-assigned quality scores, two authors independently verified whether scores assigned by the LLM matched their own judgments on $N = 408$ samples (102 per dimension), given the paper and feedback for reference. All four dimensions achieved a both-correct rate above 91%, with *constructive tone* reaching perfect agreement. The overall both-correct rate of 94.2% (PABAK = 0.828) confirms that LLM scoring aligns closely with human judgment. Results are shown in Table 9.

Table 9. Human verification of LLM quality-dimension scoring ($N = 408$, 102 per dimension).

Dimension	N	**% Correct**	**Agreement**	**PABAK**
Accuracy	102	91.7%	0.892	0.784
Constructive Tone	102	100.0%	1.000	1.000
Paper-Specific Grounding	102	94.1%	0.902	0.804
Prioritisation	102	91.2%	0.863	0.725
Overall	**408**	**94.2%**	**0.914**	**0.828**

E.1.3. AUTHOR RESPONSE PREDICTION RESULTS

Figure 3 shows validity and author action distribution over test set across different models and human reviewers. We observe that Gemini-3-flash has most similar distribution as the human reviewers. We also observe that Qwen3-8b and GOODPOINT-DPO show relatively high validity and actionability possibly due to spurious acceptance of generic feedback units.

E.2. Human Consensus-based Feedback Evaluation

E.2.1. DEFINITION AND EXAMPLE OF FEEDBACK MATCH

In human consensus-based feedback evaluation, we consider a pair of feedback as a match if the two items include at least one concrete point that shares a targeted part of the paper, a deficiency pointed out, the quality dimension addressed, and the action requested.

Figure 4 shows example feedback matches for feedback sets provided by human reviewers and LLM. Human consensus feedback is constructed by labeling matches between human feedback. Then matches between each human consensus feedback and LLM-generated feedback are labeled to evaluate LLM's ability to generate essential human feedback.

E.2.2. CONSENSUS FEEDBACK METRICS

Table 10 shows the formula of each consensus feedback metric used in the human consensus-based feedback evaluation.

E.2.3. EMBEDDING SIMILARITY-BASED FEEDBACK PAIR FILTERING

Before automatically matching human–human and human–LLM feedback pairs using GPT-5.2, we filter the pairs based on the cosine similarity of their text embeddings. We determined the thresholds for this filtering based on human annotation

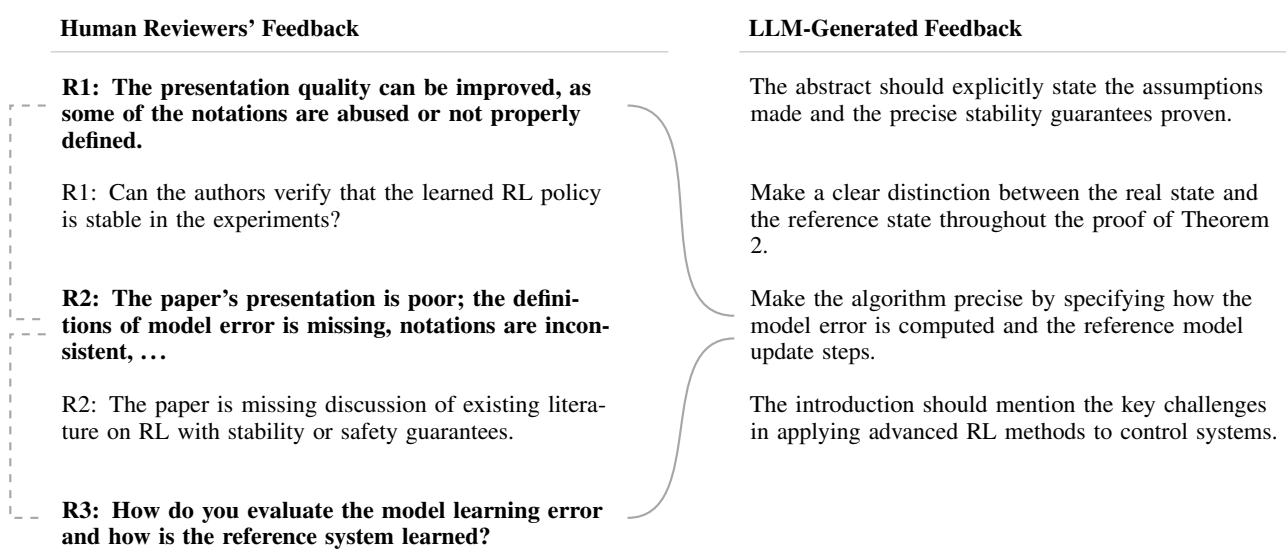

Figure 4. Illustration of human consensus-based feedback evaluation with example feedback sets. Matches between human feedback (dashed lines) are labeled to construct human consensus feedback (bold), defined as the human feedback agreed upon by multiple reviewers (R1, R2, and R3). Then each human consensus feedback is matched by LLM-generated feedback (curved lines) to estimate the LLM's ability to generate important human feedback.

Table 10. Formula of consensus feedback metrics. $\mathcal{M}$: LLM feedback units matched by human consensus feedback; $\mathcal{F}_{\text{LLM}}$: all LLM feedback; $\mathcal{G}$: human consensus feedback; $\mathcal{C}$: human consensus feedback units matched by LLM feedback.

Metric	Formula
Precision	$\frac{\lvert\mathcal{M}\rvert}{\lvert\mathcal{F}_{\text{LLM}}\rvert}$
Recall	$\frac{\lvert\mathcal{G}\cap\mathcal{C}\rvert}{\lvert\mathcal{G}\rvert}$
F1	$\frac{2\cdot\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}}$

results from a sample of feedback pairs. Specifically, to compare human-annotated match rates across different levels of similarity, we stratified the feedback pairs from 40 sample papers into seven cosine similarity ranges: <0.15, 0.15–0.25, 0.25–0.35, 0.35–0.45, 0.45–0.55, 0.55–0.65, and >0.65. From each stratum, we randomly sampled 20 human–human pairs and 20 human–LLM pairs, resulting in a total of 140 pairs for each category.

Each feedback pair was independently annotated by three annotators with more than three years of research experience. The final match label for each pair was determined by a majority vote. The annotation guidelines provided to them were as follows:

Definition of a match between two feedback units:

The two feedback units include at least one concrete point that shares all of the following aspects:

- **Targeted part of paper**
  - *Detailed Guideline for Reference Resolution:* When a feedback unit references a specific element without an explicit section (e.g., "Proposition 2"), infer the likely section based on academic conventions and context (e.g., Proposition 2 $\rightarrow$ Method/Theory). If the inferred section aligns with the section referenced in the other feedback unit, treat them as the same targeted part.

Table 11. Annotated match rate across cosine similarity strata for human–human and human–LLM feedback pairs.

	Annotated match rate	
Stratum	Human–human	Human–LLM
< 0.15	0.00	0.00
0.15–0.25	0.00	0.00
0.25–0.35	0.00	0.00
0.35–0.45	0.00	0.05
0.45–0.55	0.05	0.10
0.55–0.65	0.65	0.40
≥ 0.65	0.75	0.65

- **Deficiency pointed out**
- **Quality dimension addressed**
- **Action requested**

The annotated match rate for each stratum is presented in Table 11. The inter-annotator agreement was moderate (Krippendorff's $\alpha = 0.642$ for human–human pairs and 0.655 for human–LLM pairs), reflecting the intrinsic ambiguity of match judgments due to differences in granularity and subjective decisions at category boundaries. Based on these results, we excluded the embedding similarity ranges with a match rate below 0.1 by setting the final thresholds at 0.55 for human–human pairs and 0.45 for human–LLM pairs.

E.2.4. AUTOMATED FEEDBACK MATCH

We automatically match feedback pairs that raise a common concrete point using GPT-5.2. We use the following prompt.

User Prompt

```
You are given an abstract of a paper and a pair of feedback items about the paper.
Your task: decide whether the two feedback items match.

Definition of a match: The two feedback include at least one concrete point that
shares the following aspects.

  • Targeted part of paper
  • Deficiency pointed out
  • Quality dimension addressed
  • Action requested

Do NOT mark as a match if:

  • They only refer to the same broad section/topic but raise different issues.
  • One is praise/positive framing while the other is criticism/negative framing (or
    they otherwise differ in stance).
  • They use similar keywords but the underlying substance or requested change is
    different.

Paper abstract: {abstract}
Feedback 1: {feedback1}
Feedback 2: {feedback2}

Return a JSON object:

  • "match": "1" if they match, otherwise "0"
```

Table 12. Performance of LLM-based feedback match assessment across cosine similarity strata. Distribution weights reflect the proportion of feedback pairs in each stratum in the full test set. Accuracy, precision, recall, and F1 are reported only for strata retained after embedding-similarity filtering. The final row in each subtable reports distribution-weighted metrics, computed as weighted averages of stratum-level metrics using the test-set stratum proportions.

Stratum	Distribution weight	Accuracy	Precision	Recall	F1
0.55–0.65	0.757	0.85	1.00	0.77	0.87
≥ 0.65	0.243	0.80	0.92	0.80	0.86
Distribution-weighted	1.000	0.838	0.981	0.777	0.868

(a) Human–human feedback pairs

Stratum	Distribution weight	Accuracy	Precision	Recall	F1
0.45–0.55	0.753	1.00	1.00	1.00	1.00
0.55–0.65	0.210	0.70	0.67	0.50	0.57
≥ 0.65	0.038	0.85	0.92	0.85	0.88
Distribution-weighted	1.000	0.932	0.929	0.890	0.906

(b) Human–LLM feedback pairs

```
  • "explanation":  a brief justification focusing on the specific overlap (or lack
    of it)

Output format:

[
  {
    "match": "0/1",
    "explanation": "<brief explanation>"
  }
]
```

We use the same human-annotated sample described in Appendix E.2.3 to evaluate the performance of LLM-based feedback match assessment. Sampling feedback pairs uniformly across cosine similarity strata ensures a sufficient number of positive examples in the evaluation set by oversampling the higher-similarity strata relative to their prevalence in the full dataset.

For this evaluation, we retain only pairs whose embedding similarity falls at or above the filtering threshold determined separately for each pair type in the Appendix E.2.3. We then compare the LLM-generated match labels with the human majority-vote labels and compute accuracy, precision, recall, and F1 score for each stratum (Table 12).

Because the evaluation sample is uniformly stratified rather than distributed according to the true similarity distribution of the test set, we additionally report a distribution-weighted F1 score. Specifically, we computed the metrics within each retained stratum and then took their weighted average, using as weight the actual number of feedback pairs in each stratum according to the cosine similarity histogram of the test set of GOODPOINT-ICLR. This weighting corrects for the sampling imbalance introduced by uniform stratified sampling and yields a summary measure that better reflects performance on the real test set distribution.

E.2.5. HUMAN CONSENSUS-BASED FEEDBACK EVALUATION ON TEST SET PARTITIONS

Table 13 presents the results of the human consensus-based feedback evaluation separately for ICLR 2020–2025 and ICLR 2026 ($n = 424$ and $n = 519$, respectively). This analysis helps distinguish performance on earlier papers, which may be more susceptible to contamination, from performance on the temporally held-out 2026 set. The overall trend is consistent with the aggregated results in the main results. Across both partitions, GOODPOINT models outperform all size-comparable open-weight baselines, indicating that the gains of our models are not explained solely by possible contamination in the earlier years.

The relative behavior of the two GOODPOINT variants shows some differences across the two partitions. Compared with ICLR 2020–2025, GOODPOINT-DPO is stronger on ICLR 2026, where it exceeds GOODPOINT-SFT in recall (0.126

Table 13. Performance of GOODPOINT-trained Qwen3-8B variants and baseline models in matching human consensus feedback on GOODPOINT-ICLR test sets. For ICLR 2020–2025 ($n = 424$), GOODPOINT-SFT achieves the best performance across all metrics among size-comparable models (7B/8B). For ICLR 2026 ($n = 519$), GOODPOINT-SFT achieves the best precision and F1 among size-comparable models, while GOODPOINT-DPO achieves the best recall. Qwen3-32B and DeepReviewer-7B-Fast are evaluated on ICLR 2026 only (see §E.2.5 for details). Confidence interval half-widths for all values are less than 0.001.

	ICLR 2020–2025			ICLR 2026		
Model	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Gemini-3-flash	0.10	0.15	0.11	**0.15**	**0.18**	**0.15**
GPT-5.2	0.12	**0.16**	**0.12**	0.14	0.17	0.14
DeepReviewer-7B-Fast	—	—	—	0.14	0.16	0.13
Qwen3-32B	—	—	—	0.10	0.12	0.10
Llama3.1-8B-Inst.	0.05	0.06	0.05	0.04	0.04	0.04
Qwen3-8B (Base)	0.08	0.10	0.08	0.06	0.07	0.06
GOODPOINT-DPO	0.07 (-0.004)	0.08 (-0.019)	0.07 (-0.010)	0.11 (+0.046)	0.13 (+0.058)	0.10 (+0.044)
GOODPOINT-SFT	**0.13** (+0.051)	0.11 (+0.003)	0.10 (+0.022)	0.15 (+0.083)	0.12 (+0.050)	0.11 (+0.055)

vs. 0.118) and narrows the F1 gap (0.102 vs. 0.113). This suggests that GOODPOINT-DPO may be relatively better at generalization in alignment with consensus feedback.

We also observe a change in the relative ranking of the proprietary models across the two partitions. GPT-5.2 performs better than Gemini-3-flash on ICLR 2020–2025, but Gemini-3-flash becomes the strongest on ICLR 2026. Despite this shift, GOODPOINT-SFT continues to compare favorably in precision: on ICLR 2026, it achieves 0.145 precision, higher than GPT-5.2 (0.141) and close to Gemini-3-flash (0.149). In general, these partitioned results reinforce the robustness of the main findings while highlighting a meaningful temporal variation in model behavior.

We additionally compare against Qwen3-32B and DeepReviewer-7B-Fast on the ICLR 2026 partition. DeepReviewer-7B-Fast was trained on ICLR 2025 data, which overlaps with our main test set; comparisons are therefore restricted to the temporally held-out 2026 partition to avoid contamination. Despite being four times smaller, GOODPOINT-SFT outperforms Qwen3-32B across all metrics and achieves higher precision than DeepReviewer-7B-Fast (0.145 vs. 0.139), while DeepReviewer achieves higher F1 (0.129 vs. 0.113). This difference is expected given that DeepReviewer is trained to generate full peer reviews optimizing reviewer overlap, whereas GOODPOINT is trained to generate targeted, author-centric feedback—a distinction reflected in the complementary author-response prediction results reported below.

Table 14. Author response prediction results on ICLR 2026 partition including supplementary baselines. Values are mean ± 95% CI over $B = 1{,}000$ bootstrap iterations.

Model	**Combined**	**Validity Only**	**Action Only**
GPT-5.2	0.435[0.420,0.450]	0.440[0.425,0.455]	0.435[0.420,0.450]
Gemini-3-flash	0.372[0.359,0.384]	0.388[0.375,0.401]	0.372[0.359,0.384]
Qwen3-32B	0.159[0.147,0.171]	0.161[0.149,0.173]	0.159[0.147,0.171]
GOODPOINT-DPO	0.155[0.148,0.163]	0.160[0.152,0.168]	0.155[0.148,0.163]
GOODPOINT-SFT	0.097[0.090,0.104]	0.103[0.096,0.110]	0.097[0.090,0.104]
Qwen3-8B	0.076[0.067,0.085]	0.076[0.067,0.086]	0.076[0.067,0.085]
DeepReviewer-7B-Fast	0.064[0.059,0.070]	0.065[0.060,0.071]	0.064[0.059,0.070]
Llama3.1-8B-Inst.	0.014[0.011,0.017]	0.014[0.011,0.017]	0.014[0.011,0.017]

E.3. Human Evaluation

We recruited a total of $N = 13$ authors for human annotations. Of these 13 authors, 35.3% had 2 years of experience in research and 23.5% had 6 years of experience with all participants having at least 2 years of research experience. Most participants had research experience in NLP (13), with 2 participants with experience in robotics and ML, and one in HCI. Of the 13 unique papers submitted by participants, 10 of them were submitted to *CL conferences with 2 submitted to ICML and 2 to ICLR.

We asked a total of 4 questions for each feedback unit: validity (binary), specificity (Likert scale), author action (select one out of six with an optional text field for details), and helpfulness (Likert scale). See Table 15 for exact wording of questions.

Table 15. Author annotation questions for evaluating LLM-generated feedback.

#	Dimension	Question & Response Options
1	Validity	Do you agree that this feedback is a valid issue/question/suggestion? *Agree* — The point is valid. *Disagree* — You disagree with the premise, or the reviewer is mistaken.
2	Specificity	Is the feedback anchored to specific parts of the paper? (1–5) 5 = Very specific 4 = Mostly specific 3 = Moderately specific 2 = Mostly vague 1 = Very vague
3	Action	What action are you willing to take? *Will revise* — Make a concrete change to the manuscript. *Defer to future work* — Acknowledge but defer as out of scope. *Point to existing content* — The paper already addresses this. *Accept with no revision* — Valid but no change or deferral. *Contest with no revision* — Dispute or reject with no change. *No action (other)* — No action for another reason.
4	Helpfulness	How useful is the feedback overall to the authors? (1–5) 5 = Very helpful 4 = Helpful 3 = Moderately helpful 2 = Slightly helpful 1 = Not helpful

We randomly sampled 5 feedback units for each model to control for the number of feedback units evaluated per model and to mitigate annotator fatigue. We randomly shuffled the order, and participants saw one feedback unit at a time.

F. Baseline Generation

We use the following standardized prompt for all baseline models.

User Prompt

```
You are an expert researcher.  Please review the following research paper and
generate a set of constructive feedback to improve it.

Paper content:  {paper_content}
```

Model Configurations and Infrastructure. We use specific API endpoints and snapshots for reproducibility of our results. The open-source models, Llama-3.1-8B-Instruct and Qwen3-8B, are accessed via the Together AI API. For proprietary models, we use GPT-5.2 (`gpt-5.2-2025-12-11`) through the OpenAI API and Gemini-3-Flash (`gemini-3-flash-preview`) through the Google Gemini API .

Decoding and Length Settings. We set the decoding temperature to 0.7 across all models to balance quality and consistency. For Llama-3.1 and Qwen3-8B, the maximum number of new tokens is set to 4096. In the case of GPT-5.2 and Gemini-3-Flash, where internal reasoning tokens also count toward the total generation limit, we use larger values to accommodate full feedback generation: `max_completion_tokens` is set to 8192 for GPT-5.2, and `max_output_tokens` is set to its maximum permissible value for Gemini-3-Flash. For Qwen3-8B, we follow the provider's recommended best practices by fixing the sampling parameters to top-p=0.8, top-k=20, and `min_p`=0, while other models utilize their respective default settings. In cases where the input prompt exceeds a model's context window, the manuscript is truncated from the end to fit the maximum input token limit.

G. Analysis & Discussion

To better understand the role of LLMs in feedback generation, we analyze both their limitations relative to human and their potential to complement human feedback. Specifically, we first present our analysis of LLM failure modes through author response distribution in human evaluation and quality scoring. We then turn to a complementary question: whether LLMs can provide novel feedback beyond what human reviewers typically offer.

G.1. LLM Generated Feedback Failure Modes

One of the main immediately noticeable differences in human and LLM generated feedback includes their length. The number of feedback generated vary where human reviewer generated 5.66 feedback units across GOODPOINT-ICLR while LLM generated feedback numbers varied with highest number generated by Qwen-3-8B (20.85), followed by GPT-5.2 (20.72), Llama3.1-8B-Instruct (11.96), and Gemini-3-flash (10.79), suggesting possible verbosity of model-generated feedback. On the other hand, GOODPOINT-SFT and GOODPOINT-DPO generated 4.15 and 5.71 average feedback units respectively, more closely following human distribution but could indicate lower coverage.

LLMs generate bland and general feedback without training. One noticeable failures of LLM-generated feedback is lack of specificity, especially for smaller (7-8B) open-sourced models. This is evident in both quality scoring in §**??** and human evaluation responses. As shown in Figure 5, both Qwen3-8B and Llama3.1-8B-Instruct show mean paper specific grounding and prioritization scores considerably below human scores and reference score obtained on 1000 invalid reviewer feedback items. This is corroborated by human annotation results both quantitatively (Table 3) and qualitatively by participant provided details on feedback generated by Qwen3-8B ("*not specifc[sic] enough to make revision or judge whether it is valid or not*", P7). Through training, GOODPOINT models achieve higher scores for specificity (+0.75; GOODPOINT-DPO) and prioritization (+0.7; GOODPOINT-SFT).

LLMs miss existing content or hallucinate more than human reviewers. As shown in Figure 6, the author response distribution shows that GOODPOINT-DPO outperforms base model in both validity (agreed) and actionability (will revise, defer to future work). Moreover, our results highlight some key failure modes of LLMs in missing already addressed details. For human reviewers, point to existing content action only occurs 2.1% when feedback is valid and 1.3% when invalid. However, in LLMs, this rate increases to 3.1-8.1% when valid and 4.6-24.2% when invalid. This could indicate that models struggle to handle and provide feedback for long-context input, highlighting a key improvement area. Hallucination was another key issue noted by participants during human evaluation (*"This would read as LLM-hallucinated feedback."*, P5 on Gemini-3-flash generated feedback).

G.2. Human Alignment vs. Novel Feedback Generation

While alignment with reviewer consensus is important, LLMs can also offer complementary feedback beyond what human reviewers typically mention. To study this, we decompose the success rate improvement of GOODPOINT over the base model into two components: *aligned feedback* (*Matched & Successful*) and *novel feedback* (*Non-matched & Successful*), based on our automatic evaluations (§4.1, §4.2).

DPO drives novelty, while SFT drives alignment. Both GOODPOINT-DPO and GOODPOINT-SFT show larger relative gains in aligned feedback rate (i.e., the proportion of aligned feedback among all generated feedback) than in novel feedback rate (i.e., the proportion of novel feedback among all generated feedback) (GOODPOINT-DPO: 0.93 vs. 0.88; GOODPOINT-SFT: 0.64 vs. 0.24). However, GOODPOINT-DPO is more balanced, likely because SFT directly imitates successful human feedback, whereas DPO optimizes preferences over broader quality dimensions. This matches our automatic evaluations: GOODPOINT-SFT outperforms GOODPOINT-DPO on consensus match (§4.2) but not on success rate (§4.1).

To better understand where the gains in novel feedback come from, we classify model-generated feedback using a predefined

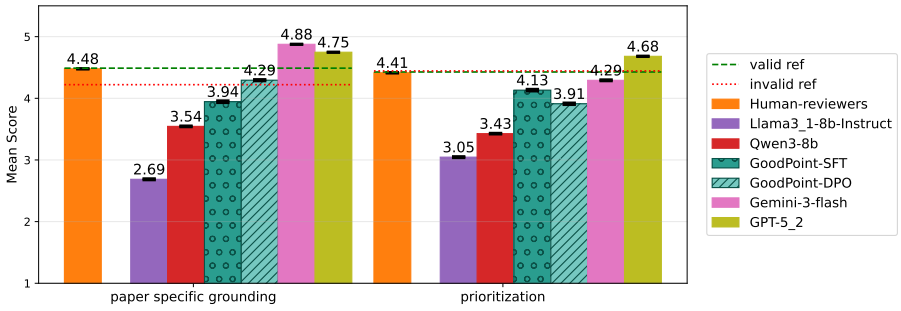

Figure 5. Mean scores for paper-specific grounding and prioritization across models with SEM. The dotted lines show reference scores of 1000 valid and invalid feedback items each.

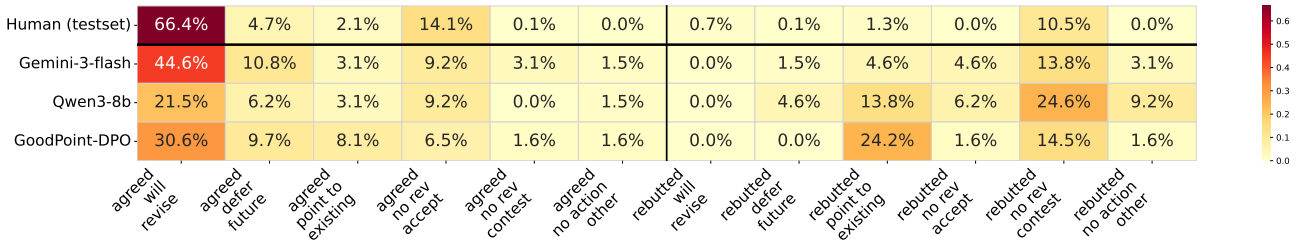

Figure 6. Author action and validity distributions from human evaluation. The top row shows results from human reviewer/author baselines on the test set. 'No rev' = no revision.

Table 16. Relative improvement over Qwen3-8B in novel and aligned feedback rates by aspect. Both GOODPOINT-SFT and GOODPOINT-DPO improve aligned feedback more than novel feedback overall, but GOODPOINT-DPO shows a more balanced pattern and broader gains across aspects. Bold indicates the larger improvement between novel and aligned feedback within each model. N/A indicates undefined due to zero feedback rate in Qwen3-8B.

Aspect	GOODPOINT-SFT		GOODPOINT-DPO	
	Novel	Aligned	Novel	Aligned
Add Experiments on More Datasets	**0.88**	0.04	**1.73**	0.56
Add Ablations Experiments	0.06	**0.38**	**1.08**	0.53
Algorithm Efficiency	-0.30	**-0.10**	**0.08**	0.05
Theoretical Soundness	-0.37	**0.16**	**0.14**	0.09
Implications of the Research	0.45	**0.81**	**2.48**	1.98
Ethical Aspects	1.58	N/A	1.49	N/A
Missing Citations	11.50	N/A	6.85	N/A
Novelty	-0.55	**-0.18**	-1.00	-1.00
Clarity and Presentation	0.31	**0.74**	0.88	**0.96**
Comparison to Previous Studies	**0.30**	0.23	0.66	**1.19**
Reproducibility	-0.20	**0.93**	0.39	**2.66**
Overall	0.24	**0.64**	0.88	**0.93**

taxonomy of 11 feedback aspects proposed by Liang et al. (2024). Each feedback item is annotated with one or more emphasized aspects, when applicable. We integrate this classification into the feedback parsing pipeline using gpt-4.1-mini (Appendix B.2), following recent work showing that LLM-based aspect classification for peer review text can provide useful and consistent annotations (Lu et al., 2025).

DPO yields broader aspect-level gains in novel feedback. Table 16 reports the relative improvement of GOODPOINT models over Qwen3-8B in both novel and aligned feedback rates by aspect. GOODPOINT-SFT's gains in novel feedback are concentrated in *Add Experiments on More Datasets*, whereas GOODPOINT-DPO shows gains across a broader range of aspects, including *Add Ablations Experiments*, *Algorithm Efficiency*, and *Theoretical Soundness*. This suggests that DPO promotes a more generalized capacity for generating complementary feedback beyond direct imitation of successful human reviews.

H. Limitations

Our approach uses author responses as a supervision signal, but these are an imperfect proxy for feedback quality, as replies can be shaped by time constraints, community norms, and rebuttal incentives; while we mitigate this with DPO and quality filtering, future work could explore expert audits, finer-grained actionability labels, or revision-based outcome signals. Additionally, our focus on ICLR limits generalizability, as feedback norms and criteria may differ across venues and disciplines, warranting future validation across additional fields.

We further acknowledge that author-response prediction, while useful as a training-alignment check, is not fully independent of the training signal. We address this limitation by triangulating across three evaluations of increasing independence: author-response prediction, human-consensus feedback matching, and a direct human study with authors. The latter two

evaluations are independent of the training signal and provide convergent evidence for the practical usefulness of the generated feedback.