Judgment-Grounded Expansion for Peer Review Generation

Anonymous ACL submission

Abstract

Automatic review generation is a promising direction for accelerating scientific progress. While most work adopts an end-to-end setup, its fully automated nature makes it less suitable for settings that demand accountability. To better balance automation and accountability, we formalize *judgment-grounded expansion*, a human-AI collaboration mode where a reviewer provides an evaluative claim and the system expands it into review comment candidate(s). We model it as a structured generate-check-refine process and conduct a user study to collect human-model interaction data. We study two practical challenges for judgment-grounded expansion: scalable evaluation and candidate set curation. We develop methods to simulate the process for large-scale evaluation, and show that conformal prediction is well suited to balancing candidate set size and target coverage. Our work establishes judgment-grounded expansion as a concrete task and provides empirical and methodological foundations for the design of future collaborative review generation systems.[1]

1 Introduction

Scientific progress relies on timely peer review. However, the academic publishing system is under increasing pressure from the rapid growth of submissions, creating a bottleneck that may slow scientific iteration (Schulz et al., 2022; Yuan et al., 2022; Biswas et al., 2023; Lin et al., 2023). There has been rising interest in automating the peer review process. In particular, automatic review generation has emerged as a promising direction with the potential to support faster scientific iteration (Kuznetsov et al., 2024; Liang et al., 2024; Chen et al., 2025a; Zhu et al., 2025; Zhuang et al., 2025).

Most existing automatic review generation systems adopt an end-to-end setup, autonomously generating a review with little or no human intervention during generation (Gao et al., 2025; Idahl and Ahmadi, 2025; Zhu et al., 2025). While full automation minimizes human effort, it limits human control during content creation and raises accountability concerns: studies show that fully automated reviews become homogenized across different papers (Lu et al., 2025; Baumann et al., 2026), exhibit verification failures (Dycke and Gurevych, 2026; You et al., 2026), and can be susceptible to manipulation (Ye et al., 2024).

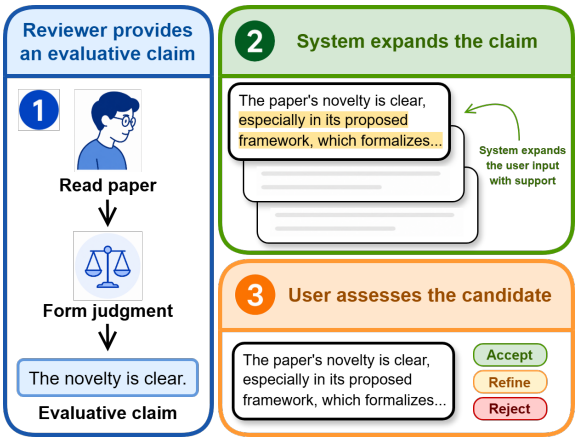


Figure 1: Judgment-grounded expansion encourages the reviewer's engagement and supports accountability.

Human-AI collaboration provides a compelling alternative (Yang, 2024; Dutta et al., 2024; Huang et al., 2025). Writing expansion is a major interaction mode in collaborative review generation: the reviewer provides a draft fragment, and the system expands it into a more complete review comment (Chen et al., 2025b). It is a promising direction for settings that demand accountability, as it maintains a degree of human control while delegating part of the work to the system. However, the interaction pipeline of writing expansion remains underspecified, especially in terms of human input and the quality check mechanism, which is crucial to the degree of human control during generation.

[1] Our code and data are available here.

We move beyond general writing expansion and formalize a more specific collaboration mode: *judgment-grounded expansion*. In this setup, the human reviewer provides an evaluative claim regarding the paper and the system expands it with support grounded in the submission, producing one or several review comment candidates for the reviewer to inspect (see Figure 1). It encourages the reviewer's active engagement: the reviewer initiates the process with an evaluative claim, remains responsible for the evaluative stance, and controls output quality through *accepting*, *rejecting*, or *refining* the candidate over multiple rounds. Anchoring generation in the reviewer's own evaluative claim is a key specification of judgment-grounded expansion, which better supports accountability in formal settings such as conference reviewing.

To study judgment-grounded expansion in practice, we conduct a user study and collect human-model interaction data. Our study characterizes reviewer input, quality check behavior, and user experience, which provides a resource for research on judgment-grounded expansion.

Moreover, we study two practical challenges for judgment-grounded expansion. The first is *scalable evaluation*: reproducible large-scale evaluation is essential for assessing design choices such as backbone model selection, yet judgment-grounded expansion requires human participation which is difficult to scale. We address this by developing methods to simulate judgment-grounded expansion. The second is *candidate set curation*: when the system presents multiple candidates to improve coverage of the reviewer's intended comment, it should keep the candidate set compact to avoid excessive reviewer burden. This size-coverage trade-off aligns naturally with conformal prediction (Angelopoulos and Bates, 2023), which constructs prediction sets carrying a user-specified coverage guarantee, keeping sets compact when the model is confident and enlarging them only when the model is uncertain. We accordingly cast candidate set curation as a conformal prediction problem.

Our contributions are summarized as follows. **(1) New task formulation.** We formalize and model judgment-grounded expansion as a generate-check-refine process. **(2) New interaction data.** We conduct a user study and collect interaction data, providing a resource for studying judgment-grounded expansion. **(3) New evaluation pipeline.** We develop a novel simulation method that enables reproducible evaluation at scale.

2 Related work

2.1 Review generation

End-to-end review generation is an active line of research in peer review automation. Existing approaches largely fall into three categories: (1) prompt-based (Du et al., 2024; Liang et al., 2024); (2) agent-based (Chamoun et al., 2024; Jin et al., 2024; Lu et al., 2024; Gao et al., 2025); (3) fine-tuned LLM (Faizullah et al., 2024; Tan et al., 2024; Idahl and Ahmadi, 2025; Zhu et al., 2025). With system prompts or pre-configured modules, they autonomously generate reviews with little or no human intervention. While it largely removes the need for human involvement, it places control over the review in the model during generation, which limits accountability. This necessitates human-AI collaborative generation, where reviewers remain involved in the generation process and can steer the system toward more accountable content.

Human-AI collaboration has gained growing attention in review generation (Sun et al., 2024a,b). Writing expansion, a major interaction mode in collaborative review generation (Chen et al., 2025b; Sadeghian et al., 2026), is particularly suitable for settings that demand accountability, as it keeps reviewers involved while delegating part of the work to the system. However, the key components of writing expansion, such as human input and quality check, remain under-specified. This gap motivates judgment-grounded expansion, a more specific collaboration mode aimed at better balancing automation and accountability.

2.2 Review evaluation

Evaluating human-AI collaboration ideally involves human participation. While being essential for capturing subjective experience (Lee et al., 2023; Fragiadakis et al., 2024), it limits reproducibility (Belz et al., 2023a,b) and scalability, especially for review writing, a costly task where participants must be domain experts and engage in the full reviewing process. Reproducible large-scale evaluation is essential for understanding how different task and model factors affect generation. We enable this by developing methods to simulate judgment-grounded expansion.

For evaluating generated review content, prior studies consider several desiderata, and some are operationalized with metrics (Ramachandran et al., 2017; Yuan et al., 2022; Sizo et al., 2025). Many studies measure alignment with human-written re-

views using text similarity metrics. Beyond surface similarity, human evaluation and LLM-as-a-judge (Chiang and Lee, 2023; Zheng et al., 2023) provide more nuanced assessments for qualities that are hard to quantify, such as constructiveness. Recent work further emphasizes two additional concerns: hallucination (Bang et al., 2023; McKenna et al., 2023; Zhang et al., 2024) and adversarial robustness (Ye et al., 2024; Zhu et al., 2025). In our evaluation, we adopt these commonly used methods (summarized in Table 3) and assess hallucination and adversarial robustness.

2.3 Conformal prediction

Conformal prediction provides a principled way to construct variable-size prediction sets with a target coverage guarantee (Angelopoulos and Bates, 2023; Angelopoulos et al., 2024). It has been increasingly applied to natural language processing tasks (Campos et al., 2024; Huang et al., 2024; Quach et al., 2024; Sheng et al., 2025; Vishwakarma et al., 2025). Conformal prediction is a natural fit for judgment-grounded expansion: set prediction is a useful strategy to increase the likelihood of covering the reviewer-intended review comment; however, the candidate set must remain compact to avoid excessive reviewer burden. We investigate conformal prediction as a promising solution to this size and coverage trade-off.

3 Judgment-grounded expansion

We model judgment-grounded expansion as a generate-check-refine process (see Figure 2).

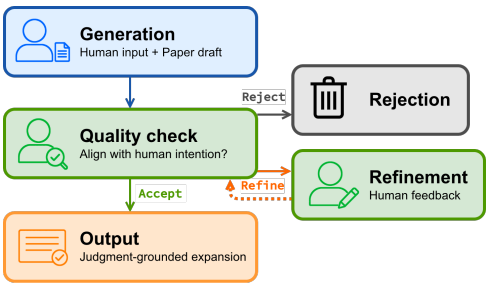


Figure 2: Our proposed framework.

3.1 Human input

The input should steer the generation by the reviewer's own evaluative claim. We therefore define the human input as, at minimum, a key point (e.g., *Novelty*) and its corresponding judgment (e.g., *strength*). This aligns with Hua et al. (2019)'s view of review as arguments each consisting of a key point, a judgment, and support.

3.2 Generation

The review generation system takes the human input and the paper as inputs, and outputs either a single review comment or a set of candidates. Compared to single-point prediction, set prediction is potentially more practical, as a single generated comment may not always match the reviewer's intended comment. Set prediction better supports judgment-grounded expansion by exploiting the generative capacity of models to explore multiple plausible expansions, increasing the likelihood that the candidate set covers the user's intention.

3.3 Quality check

After receiving the generated comment(s), the human reviewer performs quality check by assessing whether it aligns with their intention. If it passes the check, the workflow proceeds directly to *Output*. Otherwise, the human reviewer initiates *Refinement* or stops the interaction by rejecting the generation.

We define two high-level categories of quality check dimensions, **Intention** and **Grounding**, which reflect the key requirements of writing expansion: the system must interpret the reviewer's intent, and provide support to the reviewer's input. The quality check is carried out progressively, which isolates issues early and allows the reviewer to intervene in a systematic manner.

Passing the quality checks does not mean the generation is immediately ready for use. It marks the completion of judgment-grounded expansion. Additional revisions, such as tone adjustment, are handled outside the scope of this framework.

3.3.1 Intention

This dimension assesses whether the system correctly interprets the human input and captures their intention. It involves **Relevance**, whether the generation addresses the intended key point and judgment, and **Specification**, whether the generation captures the intended specificity of the input.

3.3.2 Grounding

This dimension concerns whether the generated review provides sufficient support for the human input. It includes **Evidence**, which evaluates whether the generated evidence supports the human input. The *Evidence* check inspects: hallucination, whether the generated evidence is hallucinated; match, whether the generated evidence matches the evidence the human reviewer would use (a mismatch does not necessarily mean an error);

sufficiency, when the generated evidence differs from what a human reviewer would use, whether it is still sufficient to support the human input. **Reasoning** assesses whether the reasoning is sufficiently developed to support the human input. A generation may fail this criterion if it lacks a clear logical connection to the human input.

3.4 Refinement

If the human reviewer chooses to refine the generation, they may take different actions depending on the specific issue identified. If the generation fails **Relevance**, the human reviewer may reiterate the original input to re-generate. If it fails **Specification**, the reviewer can clarify by providing a more detailed input. For **Evidence**, if hallucination is detected, the reviewer can request the system to generate alternative evidence; if sufficiency is the issue, the reviewer may ask the system to generate stronger or additional evidence. If the generation fails **Reasoning**, the human reviewer can prompt the system to elaborate the argumentation.

4 User study

Based on our proposed framework, we prototype an interface to support judgment-grounded expansion and conduct a user study. We recruited 6 participants, and collected 81 human-model interaction instances for 14 papers (see Appendix B for more details). We observe that:

(1) Concise input supports successful expansion. User inputs average 17.5 words, with 88.9% containing fewer than 30 words. 45.7% contain only a key point and evaluative stance (e.g., "The novelty of this paper is good."), 34.5% additionally reference specific paper elements (e.g., a section or table), and 19.8% include suggestions or recommendations. 72.8% of interactions result in immediate acceptance, suggesting that the system can often operate with concise user input.

(2) Failure cases reveal input requirements. 16.1% of interactions result in rejection. We observe 3 recurring user input patterns in these cases: (1) 33.3% are too brief (less than 10 words) or generic; (2) 25.0% are hedged inputs that express uncertainty or pose questions; (3) 16.7% reference external papers, which our interface cannot access.[2] 11.1% of interactions involve refinement before acceptance. In these cases, users mainly requested additional content to be incorporated, or clarified the intended focus. These patterns suggest that effective judgment-grounded expansion requires user input to be clear, definitive, and specific.

(3) User feedback highlights value and trade-offs. We collect participants' subjective experience through a post-study questionnaire. On 7-point Likert-scale questions, participants reported moderate effort in correcting generated reviews before submission (Mean=4.15) and satisfaction with the final reviews (Mean=5.69). Participants also indicated a preference for judgment-grounded expansion over writing reviews from scratch (Mean=5.92). Participants reported in open-ended responses that judgment-grounded expansion was more efficient than writing from scratch. One participant noted that it helped validate straightforward judgments quickly and reallocate effort to more uncertain cases. They felt responsible for the final review because the claim came from their own judgment. Participants also found multiple candidates helpful, but they introduced some additional cognitive burden.

5 Simulation and evaluation

We simulate the judgment-grounded expansion process to enable large-scale, reproducible evaluation, which is valuable for understanding the effect of task and model factors, but impractical to conduct with human reviewers.

5.1 Simulation

Human input. We design two methods to approximate human input: **(1) Exemplar-based proxy.** We use real human inputs collected from our user study as in-context exemplars to generate simulated human inputs (see Table 20 for the prompt). This produces inputs that reflect the characteristics of real reviewer inputs as summarized in §4. This proxy is suitable for simulation that mimics specific human input styles, but requires a sufficient pool of exemplars. **(2) Aspect-based proxy.** We use aspect and the *strength* and *weakness* labels in the review as a proxy of reviewer judgment. Aspects are characteristics of a paper that a reviewer makes judgment on. We identify aspects of human-written reviews using the tagger from Lu et al. (2025), which provides both COARSE and FINE aspects in terms of granularity (see Table 4). We begin with COARSE aspects, and use the corresponding FINE aspects when a generation fails the *Specifi-*

[2] We intentionally disable web access to prevent the system from retrieving publicly available reviews.

cation check. This proxy controls for variables such as stylistic variation. It limits leakage of reference wording while retaining the intended topic and stance signal. It is a low-information proxy as aspects are less expressive than real human input.

Quality check. For the *Relevance* and *Specification* checks, we use the aspect tagger from Lu et al. (2025) to label the aspects in the generated reviews. These checks serve as an intent-alignment proxy: a generated review passes these checks only if it covers all aspects identified for the corresponding human-written review. For *Evidence* and *Reasoning*, we use OpenAI o3-mini to simulate the checks, and the prompt is shown in Table 21.

5.2 Evaluation

We first perform a sanity check of our simulation (§5.2.2). Then, building on our framework and simulation methods, we conduct reproducible large-scale evaluation that compares judgment-grounded expansion with an end-to-end baseline to quantify the effect of reviewer judgment (§5.2.3). We then introduce *judgment-utilization score* as a diagnostic measure of how much a system benefits from reviewer judgment (§5.2.4).

5.2.1 Experimental setup

Dataset. We use reviews from ARR and ICLR. Both datasets contain reviews along with the paper draft and metadata such as strength/weakness tags and review scores. We remove low-quality reviews (e.g., extremely short ones). We randomly select 100 reviews from each dataset. We segment each review by key points: when reviews contain explicit structure, such as bullet points or numbered lists, we use these cues for segmentation; otherwise, we manually segment them based on content.

Review generation. We experiment with a range of models, including vision LLMs. We also test DeepReview, an end-to-end review generation system that is shown to outperform other existing systems (Zhu et al., 2025). We include it to examine how such a system performs with reviewer judgment. The prompts for end-to-end generation and judgment-grounded expansion are kept identical except that the latter includes reviewer judgment (see Tables 17 and 18). Since DeepReview is a pre-configured system, we could not modify its generation pipeline. We therefore append the judgment-grounded expansion prompt to the paper text as input. See Appendix D for more details.

5.2.2 Sanity check of the simulation

We assess the sanity of two key components in our simulation: human input and quality check. For human input, we compare the generated exemplar-based proxies with the real user inputs we collected in terms of input patterns. The proxies largely follow the patterns reported in §4: they are concise, with an average length of 7.2 words, and often contain only a key point and evaluative stance (68.5%). Another 18.1% reference specific paper elements, and 13.5% include suggestions or recommendations, compared to 34.5% and 19.8% in real user inputs. This suggests that the exemplar-based proxy captures broad patterns of real user input, but also shows a potential bias toward the most common input style. For the aspect-based proxy, we use the aspect tagger from Lu et al. (2025), which reports an F1 of 0.77 on its labeling task, indicating that the resulting aspect signal is sufficient to serve as a low-information proxy for human judgment. While these checks cannot directly validate the proxies against real human input, it supports their use as reasonable approximations.

We conduct human evaluation to inspect the alignment between simulated quality checks and human assessments. Table 5 shows the agreement for the quality check, including both human-model and human-human agreement. The model achieves reasonably strong agreement with human evaluators on *Relevance*, *Evidence*, and *Reasoning*, and the two human evaluators also exhibit good agreement on these dimensions. *Specification* shows lower human-model and human-human agreement, indicating that this dimension may be intrinsically harder to judge consistently.

We also inspect the consistency of the quality check simulation across different LLMs. We show in Figure 5 the simulated checks are not idiosyncratic to a specific model. Larger models ($\geq$ 32B) generally yield consistent quality check outcomes.

5.2.3 The effect of reviewer judgment

We evaluate two dimensions to measure the effect of reviewer judgment. **Reference alignment** reflects how well a system follows the provided reviewer judgment by aligning with the reference review. We measure it through text similarity and recommendation accuracy for both content-level and overall stance alignment. **Practical utility** measures the practical value of judgment-grounded expansion. We assess it using LLM-as-a-judge, comparing generated reviews on several criteria

model	BLEU% ↑			ROUGE-L% ↑			BERTScore% ↑		
	e2e	jge	Δ%	e2e	jge	Δ%	e2e	jge	Δ%
G-O-20B	2.20	**3.00**	+36.36	15.27	**16.71**	+9.43	63.92	**65.47**	+2.42
G-O-120B	2.61	**2.87**	+9.96	15.61	**17.08**	+9.42	63.93	**65.68**	+2.74
Q3-1.7B	**4.70**	3.68	-21.70	**18.98**	18.86	-0.63	**67.63**	67.39	-0.35
Q3-8B	**5.52**	4.75	-13.95	18.90	**19.05**	+0.79	67.76	**67.94**	+0.27
Q3-14B	**4.69**	4.43	-5.54	18.45	**18.60**	+0.81	67.70	**67.72**	+0.03
Q3-32B	4.21	**4.69**	+11.40	17.82	**18.87**	+5.89	67.15	**67.95**	+1.19
Q3-V-8B	3.30	**3.68**	+11.52	17.31	**17.78**	+2.72	66.45	**66.84**	+0.59
Q3-V-32B	**3.32**	3.21	-3.31	16.62	**16.69**	+0.42	66.22	**66.66**	+0.66
R1-Q-14B	4.60	**4.91**	+6.74	17.33	**18.84**	+8.71	66.84	**67.85**	+1.51
R1-Q-32B	4.18	**4.85**	+16.03	17.31	**18.82**	+8.72	66.79	**67.89**	+1.65
R1-L-8B	**3.16**	3.00	-5.06	16.56	**17.93**	+8.27	65.92	**66.74**	+1.24
R1-L-70B	3.94	**5.00**	+26.90	17.59	**18.87**	+7.28	67.15	**68.12**	+1.44
DeepR	1.86	**1.98**	+6.45	11.09	**11.10**	+0.09	66.41	**66.50**	+0.14

(a) Text similarity

model	Soundness				Overall assessment			
	≤0.5% ↑		MAE% ↓		≤0.5% ↑		MAE% ↓	
	e2e	jge	e2e	jge	e2e	jge	e2e	jge
G-O-20B	**76.47**	74.47	53.53	**48.94**	64.71	**69.15**	63.53	**56.91**
G-O-120B	70.84	**76.77**	56.77	**51.52**	66.67	**76.76**	60.42	**47.98**
Q3-1.7B	**75.26**	64.40	**49.48**	63.56	**64.95**	64.40	59.28	**57.63**
Q3-8B	**59.79**	53.57	**67.01**	69.64	54.64	**55.95**	**65.46**	68.45
Q3-14B	**82.47**	76.19	**47.42**	51.19	**81.44**	78.57	**47.94**	57.19
Q3-32B	61.86	**65.22**	61.86	**59.24**	48.45	**51.09**	74.23	**69.02**
Q3-V-8B	72.00	**73.33**	55.00	**54.00**	57.00	**70.66**	63.50	**58.00**
Q3-V-32B	65.00	**70.00**	61.00	**54.00**	52.00	**58.00**	70.00	**64.50**
R1-Q-14B	68.05	**72.22**	56.19	**55.56**	50.52	**61.11**	76.80	**74.31**
R1-Q-32B	63.44	**65.17**	**60.22**	60.67	67.74	**70.79**	57.53	**55.06**
R1-L-8B	**55.55**	41.47	**76.77**	95.12	39.39	**41.47**	**95.96**	97.56
R1-L-70B	**69.39**	69.32	56.12	**51.14**	51.02	**59.09**	69.90	**64.20**

(b) Recommendation accuracy

model	TA ↑		CV ↑		AD ↑	
	e2e	jge	e2e	jge	e2e	jge
G-O-20B	**64.82**	25.93	**72.23**	20.99	**68.52**	24.69
G-O-120B	**53.08**	36.30	**62.78**	30.32	**54.26**	37.76
Q3-1.7B	47.46	**48.31**	**55.94**	41.53	44.92	**54.24**
Q3-8B	**46.43**	45.84	**54.17**	41.08	42.86	**52.98**
Q3-14B	44.26	**46.15**	**58.10**	36.50	44.94	**47.86**
Q3-32B	**49.46**	42.39	**63.04**	34.24	**51.09**	45.11
Q3-V-8B	37.84	**50.00**	**59.46**	32.44	**48.65**	45.95
Q3-V-32B	42.28	**51.66**	**52.46**	44.52	43.91	**52.56**
R1-Q-14B	29.17	**66.67**	31.95	**65.28**	25.70	**71.53**
R1-Q-32B	19.10	**77.53**	26.97	**71.91**	19.67	**76.97**
R1-L-8B	**54.46**	44.30	**63.21**	36.80	**53.08**	46.93
R1-L-70B	19.54	**75.86**	22.99	**74.71**	16.09	**82.19**
DeepR	42.90	**47.29**	45.66	**45.93**	**47.06**	45.23
AVERAGE	41.70	**51.11**	**50.07**	45.65	42.35	**53.38**

(c) LLM-as-a-judge win rates

model	Soundness				Overall assessment			
	$\Delta_{\leq 0.5}$%		Δ_{MAE}%		$\Delta_{\leq 0.5}$%		Δ_{MAE}%	
	e2e	jge	e2e	jge	e2e	jge	e2e	jge
G-O-20B	76.85	48.25	147.13	104.33	87.13	56.31	150.87	101.34
G-O-120B	39.30	34.87	72.63	51.40	46.00	41.38	67.16	70.90
Q3-1.7B	48.79	27.07	105.25	34.69	29.44	22.36	29.15	27.50
Q3-8B	46.55	2.31	60.01	10.20	49.07	14.80	78.75	14.67
Q3-14B	33.87	12.50	49.11	13.95	54.11	19.38	117.02	15.85
Q3-32B	48.87	10.00	81.44	10.09	68.42	14.90	89.02	15.75
R1-Q-14B	14.28	16.19	29.77	14.87	15.46	18.18	28.85	16.86
R1-Q-32B	22.87	18.77	36.02	16.35	48.17	26.88	69.20	27.13
R1-L-8B	22.59	21.77	19.19	15.08	8.61	34.84	9.94	27.43
R1-L-70B	28.69	14.08	47.88	31.83	37.36	25.84	61.50	26.01
AVERAGE	38.27	**20.58**	64.84	**30.28**	44.38	**27.49**	70.15	**34.34**

(d) Adversarial robustness

Table 1: (a) Text similarity scores between end-to-end generations (**e2e**), judgment-grounded expansions (**jge**), and human-written reviews. (b) The percentage of score predictions within 0.5 points of the human-assigned scores (**≤0.5**) and the mean absolute error (**MAE**). (c) LLM-as-a-judge win rates, TA = Technical Accuracy, CV = Constructive Value, AD = Analytical Depth. (d) Relative score changes under adversarial and non-adversarial settings. G-O = GPT-OSS, Q = Qwen, L = LLaMA, V = VL, R1 = Deepseek-R1, DeepR = DeepReview.

(Zhu et al., 2025), and adversarial robustness (Ye et al., 2024). We use the aspect-based proxy and generate one review comment per input to reduce confounds from variations in human input and the prediction paradigm. See E.2 for more details.

For **reference alignment**, generated reviews under judgment-grounded expansion generally align better with the reference review. The relative gain over end-to-end generation in Table 1a reflects how much additional text alignment is obtained when reviewer judgment is provided. We observe that larger models tend to have larger gains. Models trained with the DeepSeek-R1 distillation objective appear to benefit more from reviewer judgment, as shown by the comparison between base Qwen3 and R1-Distill-Qwen models. DeepReview also shows gains, indicating that a complex review generation system originally designed for end-to-end generation can still benefit from reviewer judgment. Table 1b shows judgment-grounded expansion generally achieves higher **≤0.5** and lower **MAE** scores, especially on *Overall assessment*. This suggests that reviewer judgment helps anchor the generation to the reviewer's overall impression of a paper.

Judgment-grounded expansion further shows **practical utility**. It achieves generally higher LLM-as-a-judge win rates on technical accuracy and analytical depth (see Table 1c), possibly because reviewer judgment constrains the development of technical points and the scope of the generation, which reduces technically inaccurate expansion and makes the comment more focused. Table 1d shows both $\Delta_{\leq 0.5}$% and Δ_{MAE}% change substantially for end-to-end generation, while judgment-grounded expansion shows much smaller changes under the adversarial setting. Anchoring generation in reviewer judgment makes the system less affected by prompt injection, which is important for maintaining accountability.

We do not claim that judgment-grounded expansion universally outperforms end-to-end generation. Our results show that the reviewer judgment does not always lead to positive gains. This further highlights the need for reproducible large-scale eval-

uation to inspect how factors such as backbone models affect judgment-grounded expansion.

We report several complementary analyses in Appendix E. We study the effect of quality check refinement and observe further gains in some cases (E.2.1). We manually analyze cases where judgment-grounded expansion underperforms the end-to-end baseline to characterize when reviewer judgment fails to help (E.2.2). We report results on ICLR and compare the exemplar-based and aspect-based proxies (E.2.3).

5.2.4 Judgment-utilization ability

Building on our simulation and evaluation pipeline, we introduce *judgment-utilization ability* as a diagnostic measure of the gains a system obtains from the reviewer judgment. This measure is not intended to be a setting-independent capability score. It is conditioned on the task, domain, metric, and model pool. For each model m and condition l (task, domain, metric), we denote by $p^j_{m,l}$ and $p^e_{m,l}$ its judgment-grounded expansion and end-to-end performance. We compute the normalized gain[3] and then aggregate it to obtain a single **JUScore**:

$$\tilde{p}^s_{m,l} = \frac{p^s_{m,l}}{\max_{m' \in \mathcal{M}}(p^j_{m',l}, p^e_{m',l})}, s \in \{j, e\};$$
$$JUScore(m) = \frac{1}{|\mathcal{L}|}\sum_{l}^{\mathcal{L}}(\tilde{p}^j_{m,l} - \tilde{p}^e_{m,l}), \quad (1)$$

where $\mathcal{M}$ is the set of backbone models and $\mathcal{L}$ is the set of conditions. The JUScore provides a task-, domain-, and metric-specific perspective of a model's ability to utilize reviewer judgment.

We compute JUScore on the review generation task for ARR papers using text similarity, recommendation accuracy, and adversarial robustness. Figure 3 visualizes the resulting JUScores.

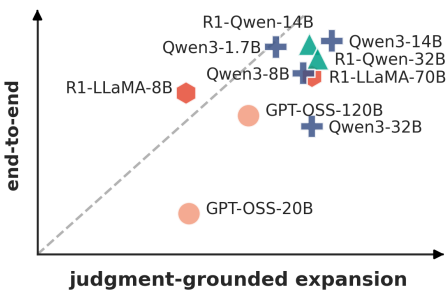


Figure 3: End-to-end versus judgment-grounded expansion performance. The horizontal offset to the dashed line ($y = x$) corresponds to a model's JUScore.

[3]Since metrics differ in direction, we negate lower-is-better metrics before normalization.

In our setting, Qwen3-32B shows the largest gains from reviewer judgment (the largest positive offsets from $y = x$). Higher end-to-end performance may not imply a high JUScore: the Pearson coefficient between end-to-end performance and JUScore is moderately negative but not statistically significant ($r = -0.5058$, $p = 0.0816$). JUScore is a diagnostic that shows which models in a given pool benefit most from reviewer input. Its value is in informing backbone selection in a specific setting, not in producing a universal ranking.

6 Balancing size and coverage

Compared with single-point prediction, set prediction is a more practical setting for judgment-grounded expansion. By generating multiple possible expansions of a reviewer judgment, set prediction may increase the likelihood of covering the reviewer's intended comment. Though our user study participants found multiple candidates helpful, they commented that having multiple candidates adds cognitive load (see §4). This leads to a key practical question: how to curate a compact set while preserving coverage of the reviewer-selected review comment. Our user study data provides a useful resource for this analysis, as it records both the generated candidate set and the user's selected candidate for each input.

A straight forward way to curate a compact candidate set is **score-based ranking**, assigning each candidate a score and keeping the top-k candidates. We evaluate 2 types of commonly used and interpretable scoring methods: (1) *uncertainty-based* scores, including negative log-likelihood (NLL), maximum sequence probability (MSP), mean token entropy (MTE) (Fomicheva et al., 2020), which are calculated using the candidate's generation probabilities; (2) *similarity-based* score, calculated as SBERT similarity between the reviewer input and the candidate (Reimers and Gurevych, 2019).

method	@1	@2	@3	avg. rank
random	26.21	52.49	78.59	2.47
NLL	33.82	58.82	80.88	2.28
MSP	**42.65**	64.71	82.35	2.13
MTE	41.18	63.24	82.35	2.16
Sim	35.29	**66.18**	85.29	2.16
MSP+Sim	41.18	64.71	**86.76**	**2.09**

Table 2: Score-based ranking results. We report the percentage of cases where the human-selected candidate ranks first (**@1**) or within the top two or three (**@2**, **@3**), and its average rank (**avg. rank**).

Table 2 shows the results of score-based ranking. Results are based on 68 interactions where users selected a candidate, with an average of 3.93 candidates per instance. Compared with the random baseline, both uncertainty-based and similarity-based scores capture meaningful signals related to user preference. We further combine the strongest **@1** and **@2** methods, i.e., MSP+Sim, by normalizing MSP and Sim separately within the candidate set and summing the normalized scores. The combined score improves **@3** and yields the best average rank, suggesting that different methods may provide complementary signals.

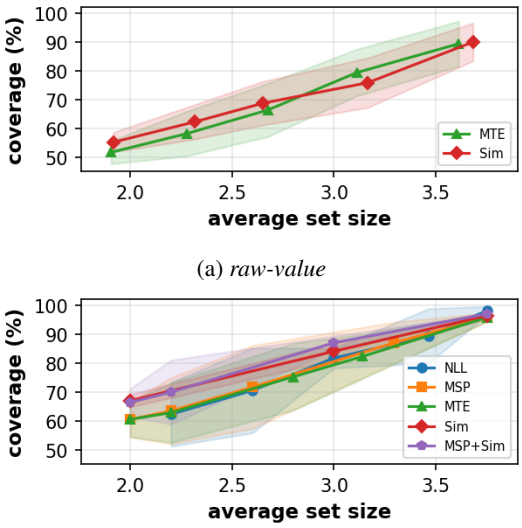
A limitation of score-based set curation is that it produces prediction sets of fixed size. In practice, however, the desired set size should adapt to instance-level "signals" such as uncertainty or confidence. Conformal prediction is therefore better suited to this setting: by calibrating a threshold on a held-out calibration set, it returns variable-size prediction sets with a formal guarantee of target coverage (Angelopoulos et al., 2024).

We split the 68 interactions where users selected a candidate evenly into calibration and test sets. To reduce variance from a single split, we report average results over 5 random splits. Each calibration instance consists of a user input x, a generated candidate set $\mathcal{Y}(x)$, and the human-selected candidate c^*. For a target coverage $1-\alpha$, conformal prediction estimates a threshold τ_α from the nonconformity scores of human-selected candidates in the calibration set. At test time, it returns all candidates $\mathcal{C}_\alpha(x)$ whose nonconformity scores are below this threshold:

$$\mathcal{C}_\alpha(x) = \{c : S(x,c) \leq \tau_\alpha\}. \quad (2)$$

Conformal prediction constructs an adaptive candidate set by applying this calibrated threshold at test time. We evaluate two types of nonconformity scores. For *raw-value* nonconformity, $S(x,c)$ is computed directly from the scoring function. For *rank-based* nonconformity, $S(x,c)$ is the candidate's rank within $\mathcal{Y}(x)$ according to the scoring function (Huang et al., 2024). We evaluate the same scoring functions as in score-based ranking. We exclude NLL and MSP from *raw-value* nonconformity, since their absolute values are not comparable across inference instances.

Figure 4 shows that *rank-based* nonconformity generally provides a better size and coverage trade-off than *raw-value* nonconformity, achieving higher coverage at similar average set sizes. Among *rank-based* methods, Sim and MSP+Sim perform particularly well. Compared with fixed score-based ranking, conformal prediction better supports judgment-grounded expansion: it allows the system to adjust candidate set size according to a nonconformity score, and by selecting different target coverage levels, practitioners can manually tune whether the system favors broader candidate sets with higher coverage or smaller sets that reduce reviewer cognitive load. This makes conformal prediction a promising technique for balancing target coverage against reviewer cognitive load.



(a) *raw-value*

(b) *rank-based*

Figure 4: Conformal prediction results. We test $\alpha \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$. Each point indicates the coverage rate of the human-selected candidate and the average candidate set size. Shaded regions indicate variation across random calibration/test splits. The target coverage is $1-\alpha$, but empirical coverage may not exactly match it due to finite calibration/test samples.

Though the scoring functions capture reviewer preference to some extent, there remains room for improvement. Future work could design nonconformity scores tailored to peer review by incorporating review-specific factors, such as actionability, verifiability, and helpfulness (Sadallah et al., 2025).

7 Conclusion

Judgment-grounded expansion is more than an alternative for producing reviews. It accommodates a genuine need that balances automation and accountability. Our work presents a systematic study of judgment-grounded expansion, establishes it as a concrete collaboration mode, and provides a foundation for future review generation systems.

Limitations

Reviewer expertise. Judgment-grounded expansion relies on the reviewer's ability to provide reasonable evaluative claims and assess generated comments. As a result, this setting is potentially less suitable for less experienced reviewers.

Simulation. (1) Our simulation does not aim to capture human preferences. Our goal is therefore not to replace human evaluation, but to complement it by enabling reproducible large-scale evaluation for analyzing task and model factors. (2) For human input simulation, the exemplar-based proxy captures broad patterns of real user input but may be biased toward the most common input style. Our aspect-based proxy relies on the aspect tagger from Lu et al. (2025), which is not perfectly accurate. This is also a reason why we treat the aspect-based proxy as a low-information proxy: it may provide an inaccurate signal about the reviewer's intended key point. As a result, findings based on the aspect-based proxy, as well as quality check results involving aspect information, may be affected by tagging errors. Since more accurate aspect labels would yield stronger evaluative signals, this limitation likely makes our estimates conservative. (3) For quality check simulation, the *Specification* check shows lower human-model and human-human agreement, suggesting that simulating this dimension remains an open question for future work. (4) Since it is impractical to compare simulated results with corresponding human-involved results, we inspect the validity of our simulation through sanity checks on its two key components, i.e., human input and quality check. While these checks provide supporting evidence, they do not constitute a full validation. In particular, the human evaluation used to inspect the simulated quality check is limited in scale.

User study. The scale of our user study is limited, which may affect the generality of the findings based on this data in §4 and §6. However, scaling such studies is challenging as participants must read papers, form judgments, and interact with the system, which is time-consuming. Despite this limitation, our study provides an initial resource for studying judgment-grounded expansion.

Experimental setup. (1) We do not experiment with closed-source backbone models. Since our target is formal settings such as conference reviewing, we consider open-source models more appropriate, as they better support transparency and deployability. (2) Since DeepReview is a pre-configured system and we could not modify its generation pipeline, we append the judgment-grounded expansion prompt to the paper text as input. This is not a fair setup for DeepReview. Nevertheless, we still observe gains from reviewer judgment in this case. (3) We do not have access to pre-rebuttal review scores in both the ARR and ICLR data, which may introduce a version mismatch between the review scores and the review text. Thus, we report the percentage of score predictions within 0.5 points of the human-assigned scores in our evaluation to partially mitigate this issue.

Evaluation metrics. The evaluation methods we use are not direct measures of review quality, and each metric has its own shortcomings. We adopt them for two reasons. First, they follow common practice in prior work. Second, they capture different perspectives, which together provide a more well-rounded view of model performance.

Reproducibility. The reproducibility of quality check simulations using OpenAI o3-mini may be limited, as it is a closed-source model. We chose this model because quality check simulation is a key component of our pipeline that may benefit from strong reasoning ability. To mitigate this, we release all prompts and the simulation outputs.

Ethics statement

Judgment-grounded expansion mitigates some accountability concerns of fully automatic review generation by anchoring the content in the reviewer's own evaluative claim. However, it does not resolve questions regarding authorship, disclosure, and responsibility for the review. Our work explores how to better balance automation and accountability, rather than offering a complete solution to these questions.

Our experiments use publicly available peer review data licensed for research use. The user study was conducted with ethics approval.

We used an AI assistant to improve the writing and language of the manuscript.

Acknowledgments

References

Anastasios N. Angelopoulos and Stephen Bates. 2023. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591.

Anastasios Nikolas Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. 2024. Conformal risk control. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

Joachim Baumann, Jiaxin Pei, Sanmi Koyejo, and Dirk Hovy. 2026. Stop automating peer review without rigorous evaluation. *arXiv preprint arXiv:2605.03202*.

Anya Belz, Craig Thomson, Ehud Reiter, Gavin Abercrombie, Jose M. Alonso-Moral, Mohammad Arvan, Anouck Braggaar, Mark Cieliebak, Elizabeth Clark, Kees van Deemter, Tanvi Dinkar, Ondřej Dušek, Steffen Eger, Qixiang Fang, Mingqi Gao, Albert Gatt, Dimitra Gkatzia, Javier González-Corbelle, Dirk Hovy, and 23 others. 2023a. Missing Information, Unresponsive Authors, Experimental Flaws: The Impossibility of Assessing the Reproducibility of Previous Human Evaluations in NLP. In *Proceedings of the Fourth Workshop on Insights from Negative Results in NLP*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

Anya Belz, Craig Thomson, Ehud Reiter, and Simon Mille. 2023b. Non-Repeatable Experiments and Non-Reproducible Results: The Reproducibility Crisis in Human Evaluation in NLP. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3676–3687, Toronto, Canada. Association for Computational Linguistics.

Som Biswas, Dushyant Dobaria, and Harris L. Cohen. 2023. ChatGPT and the Future of Journal Reviews: A FeasibilityStudy. *The Yale Journal of Biology and Medicine*, 96(3):415–420.

Margarida Campos, António Farinhas, Chrysoula Zerva, Mário A. T. Figueiredo, and André F. T. Martins. 2024. Conformal prediction for natural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 12:1497–1516.

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, Bangkok, Thailand. Association for Computational Linguistics.

Qiguang Chen, Ming-Hsuan Yang, Libo Qin, Jinhao Liu, Zheng Yan, Jiannan Guan, Dengyun Peng, Yiyan Ji, Hanjing Li, Mengkang Hu, Yimeng Zhang, Yihao Liang, Yu Zhou, Jiaqi Wang, Zhi Chen, and Wanxiang Che. 2025a. Ai4research: A survey of artificial intelligence for scientific research. *CoRR*, abs/2507.01903.

Shiping Chen, Duncan Brumby, and Anna Cox. 2025b. Envisioning the Future of Peer Review: Investigating LLM-Assisted Reviewing Using ChatGPT as a Case Study. In *Proceedings of the 4th Annual Symposium on Human-Computer Interaction for Work*, CHIWORK '25, pages 1–18, New York, NY, USA. Association for Computing Machinery.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can Large Language Models Be an Alternative to Human Evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Peng Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Srinath, Haoran Ranran Zhang, Vipul Gupta, Yinghui Li, Tao Li, Fei Wang, Qin Liu, Tianlin Liu, Pengzhi Gao, Congying Xia, and 21 others. 2024. LLMs Assist NLP Researchers: Critique Paper (Meta-)Reviewing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5081–5099, Miami, Florida, USA. Association for Computational Linguistics.

Subhabrata Dutta, Timo Kaufmann, Goran Glavas, Ivan Habernal, Kristian Kersting, Frauke Kreuter, Mira Mezini, Iryna Gurevych, Eyke Hüllermeier, and Hinrich Schütze. 2024. Problem solving through human-ai preference-based cooperation. *CoRR*, abs/2408.07461.

Nils Dycke and Iryna Gurevych. 2026. Automatic reviewers fail to detect faulty reasoning in research papers: A new counterfactual evaluation framework. *Transactions of the Association for Computational Linguistics*, 14:465–488.

Abdur Rahman Bin Mohammed Faizullah, Ashok Urlana, and Rahul Mishra. 2024. LimGen: Probing the LLMs for Generating Suggestive Limitations of Research Papers. In *Machine Learning and Knowledge Discovery in Databases. Research Track*, pages 106–124, Cham. Springer Nature Switzerland.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

George Fragiadakis, Christos Diou, George Kousiouris, and Mara Nikolaidou. 2024. Evaluating human-ai

collaboration: A review and methodological framework. *CoRR*, abs/2407.19098.

Xian Gao, Jiacheng Ruan, Jingsheng Gao, Ting Liu, and Yuzhuo Fu. 2025. Reviewagents: Bridging the gap between human and ai-generated paper reviews. *CoRR*, abs/2503.08506.

Zhaolin Gao, Kianté Brantley, and Thorsten Joachims. 2024. Reviewer2: Optimizing review generation through prompt generation. *CoRR*, abs/2402.10886.

Kilem Li Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.

Xinyu Hua, Mitko Nikolov, Nikhil Badugu, and Lu Wang. 2019. Argument Mining for Understanding Peer Reviews. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2131–2137, Minneapolis, Minnesota. Association for Computational Linguistics.

Chen Huang, Yang Deng, Wenqiang Lei, Jiancheng Lv, Tat-Seng Chua, and Jimmy Huang. 2025. How to enable effective cooperation between humans and NLP models: A survey of principles, formalizations, and beyond. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 466–488, Vienna, Austria. Association for Computational Linguistics.

Jianguo Huang, Huajun Xi, Linjun Zhang, Huaxiu Yao, Yue Qiu, and Hongxin Wei. 2024. Conformal prediction for deep classifier via label ranking. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pages 20331–20347. PMLR / OpenReview.net.

Maximilian Idahl and Zahra Ahmadi. 2025. OpenReviewer: A Specialized Large Language Model for Generating Critical Scientific Paper Reviews. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 550–562, Albuquerque, New Mexico. Association for Computational Linguistics.

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, Miami, Florida, USA. Association for Computational Linguistics.

Ilia Kuznetsov, Osama Mohammed Afzal, Koen Dercksen, Nils Dycke, Alexander Goldberg, Tom Hope, Dirk Hovy, Jonathan K. Kummerfeld, Anne Lauscher, Kevin Leyton-Brown, Sheng Lu, Mausam, Margot Mieskes, Aurélie Névéol, Danish Pruthi, Lizhen Qu, Roy Schwartz, Noah A. Smith, Thamar Solorio, and 5 others. 2024. What can natural language processing do for peer review? *CoRR*, abs/2405.06563.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, pages 611–626, New York, NY, USA. Association for Computing Machinery.

Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, Rose E. Wang, Minae Kwon, Joon Sung Park, Hancheng Cao, Tony Lee, Rishi Bommasani, Michael S. Bernstein, and Percy Liang. 2023. Evaluating human-language model interaction. *Transactions on Machine Learning Research*, 2023.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Yi Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, Daniel A. McFarland, and James Zou. 2024. Can Large Language Models Provide Useful Feedback on Research Papers? A Large-Scale Empirical Analysis. *NEJM AI*, 1(8):AIoa2400196. Publisher: Massachusetts Medical Society.

Jialiang Lin, Jiaxin Song, Zhangping Zhou, Yidong Chen, and Xiaodong Shi. 2023. Automated scholarly paper review: Concepts, technologies, and challenges. *Information Fusion*, 98:101830.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173. Place: Cambridge, MA Publisher: MIT Press.

Ryan Liu and Nihar B. Shah. 2023. Reviewergpt? an exploratory study on using large language models for paper reviewing. *CoRR*, abs/2306.00622.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob N. Foerster, Jeff Clune, and David Ha. 2024. The AI scientist: Towards fully automated open-ended scientific discovery. *CoRR*, abs/2408.06292.

Sheng Lu, Ilia Kuznetsov, and Iryna Gurevych. 2025. Identifying aspects in peer reviews. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6145–6167, Suzhou, China. Association for Computational Linguistics.

Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Hosseini, Mark Johnson, and Mark Steedman. 2023. Sources of Hallucination by Large Language Models on Inference Tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2758–2774, Singapore. Association for Computational Linguistics.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. *CoRR*, abs/2508.10925.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzilay. 2024. Conformal language modeling. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Lakshmi Ramachandran, Edward F. Gehringer, and Ravi K. Yadav. 2017. Automated Assessment of the Quality of Peer Reviews using Natural Language Processing Techniques. *International Journal of Artificial Intelligence in Education*, 27(3):534–581.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Abdelrahman Sadallah, Tim Baumgärtner, Iryna Gurevych, and Ted Briscoe. 2025. The good, the bad and the constructive: Automatically measuring peer review's utility for authors. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28991–29021, Suzhou, China. Association for Computational Linguistics.

Soroush Sadeghian, Alireza Daqiq, Radin Cheraghi, Sajad Ebrahimi, Negar Arabzadeh, and Ebrahim Bagheri. 2026. Peerprism: Peer evaluation expertise vs review-writing AI. *CoRR*, abs/2604.14513.

Robert Schulz, Adrian Barnett, René Bernard, Nicholas J. L. Brown, Jennifer A. Byrne, Peter Eckmann, Małgorzata A. Gazda, Halil Kilicoglu, Eric M. Prager, Maia Salholz-Hillel, Gerben ter Riet, Timothy Vines, Colby J. Vorland, Han Zhuang, Anita Bandrowski, and Tracey L. Weissgerber. 2022. Is the future of peer review automated? *BMC Research Notes*, 15(1):203.

Huanxin Sheng, Xinyi Liu, Hangfeng He, Jieyu Zhao, and Jian Kang. 2025. Analyzing uncertainty of LLM-as-a-judge: Interval evaluations with conformal prediction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11286–11328, Suzhou, China. Association for Computational Linguistics.

Amanda Sizo, Adriano Lino, Álvaro Rocha, and Luís Paulo Reis. 2025. Defining quality in peer review reports: a scoping review. *Knowledge and Information Systems*, 67(8):6413–6460.

Lu Sun, Aaron Chan, Yun Seo Chang, and Steven P. Dow. 2024a. ReviewFlow: Intelligent Scaffolding to Support Academic Peer Reviewing. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, IUI '24, pages 120–137, New York, NY, USA. Association for Computing Machinery.

Lu Sun, Stone Tao, Junjie Hu, and Steven P. Dow. 2024b. MetaWriter: Exploring the Potential and Perils of AI Writing Support in Scientific Peer Review. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW1):94:1–94:32.

Cheng Tan, Dongxin Lyu, Siyuan Li, Zhangyang Gao, Jingxuan Wei, Siqi Ma, Zicheng Liu, and Stan Z. Li. 2024. Peer review as A multi-turn and long-context dialogue with role-based interactions. *CoRR*, abs/2406.05688.

Llama Team. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Qwen Team. 2025. Qwen3 technical report. *CoRR*, abs/2505.09388.

Keith Tyser, Ben Segev, Gaston Longhitano, Xin-Yu Zhang, Zachary Meeks, Jason Lee, Uday Garg, Nicholas Belsten, Avi Shporer, Madeleine Udell, Dov Te'eni, and Iddo Drori. 2024. Ai-driven review systems: Evaluating llms in scalable and bias-aware academic reviews. *CoRR*, abs/2408.10365.

Harit Vishwakarma, Alan Mishler, Thomas Cook, Niccolò Dalmasso, Natraj Raman, and Sumitra Ganesh. 2025. Prune 'n predict: Optimizing LLM decision-making with conformal prediction. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large Language Models are not Fair Evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

Qingyun Wang, Qi Zeng, Lifu Huang, Kevin Knight, Heng Ji, and Nazneen Fatema Rajani. 2020. ReviewRobot: Explainable Paper Review Generation based on Knowledge Synthesis. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 384–397, Dublin, Ireland. Association for Computational Linguistics.

Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. 2025. Cycleresearcher: Improving automated research via automated review. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Diyi Yang. 2024. Human-AI Interaction in the Age of Large Language Models. *Proceedings of the AAAI Symposium Series*, 3(1):66–67.

Rui Ye, Xianghe Pang, Jingyi Chai, Jiaao Chen, Zhenfei Yin, Zhen Xiang, Xiaowen Dong, Jing Shao, and Siheng Chen. 2024. Are we there yet? revealing the

risks of utilizing large language models in scholarly peer review. *CoRR*, abs/2412.01708.

Lei You, Lele Cao, and Iryna Gurevych. 2026. Preventing the collapse of peer review requires verification-first AI. *CoRR*, abs/2601.16909.

Weizhe Yuan and Pengfei Liu. 2022. KID-Review: Knowledge-Guided Scientific Review Generation with Oracle Pre-training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11639–11647.

Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2022. Can We Automate Scientific Reviewing? *Journal of Artificial Intelligence Research*, 75:171–212.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: transformers for longer sequences. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, pages 17283–17297, Red Hook, NY, USA. Curran Associates Inc.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2024. How language model hallucinations can snowball. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, pages 59670–59684, Vienna, Austria. JMLR.org.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pages 46595–46623, Red Hook, NY, USA. Curran Associates Inc.

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, Torino, Italia. ELRA and ICCL.

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(C).

A More on related work

Table 3 summarizes common evaluation metrics used in prior work for review evaluation.

method	related work
Text similarity	Chamoun et al. (2024); Faizullah et al. (2024); Gao et al. (2024); Tan et al. (2024); Zhou et al. (2024); Gao et al. (2025)
Recommendation	Yuan and Liu (2022); Lu et al. (2024); Zhou et al. (2024); Idahl and Ahmadi (2025); Zhu et al. (2025); Weng et al. (2025)
LLM-as-a-judge	Gao et al. (2024); Zhou et al. (2024); Idahl and Ahmadi (2025); Zhu et al. (2025)
Human evaluation	Wang et al. (2020); Yuan and Liu (2022); Liang et al. (2024); Liu and Shah (2023); Chamoun et al. (2024); Du et al. (2024); Tyser et al. (2024); Zhou et al. (2024)

Table 3: Common evaluation methods used in literature. Return to main text.

B More on user study

We implement a web-based interface for judgment-grounded expansion. The interface has two main components, a PDF reader on one side, and an interaction input panel on the other where users enter evaluative claims and inspect system outputs. Users initiate the process by typing an evaluative claim (e.g., "The novelty is clear") into the input panel. The system then generates a set of review comment candidates using a two-step pipeline: it first prompts an LLM to identify 2-5 plausible angles implied by the claim with respect to the paper, and then generates one candidate comment per angle in a separate call, conditioned on the paper, the claim, and the corresponding angle. We adopt this two-step generation pipeline so that candidates are meaningfully different along interpretable dimensions, rather than paraphrases of each other. We use GPT-OSS-120B as the backbone model. This choice was made before our experiments and was not based on the findings reported in the paper. We selected it as a strong large open-source model that we expected to perform well. For each inference, we record all the generated candidates and their `logprobs`. See here for the implementation of the interface.

We recruited 6 participants through an open call. All participants are PhD students with at least 2 years of research experience. English is their primary language for reading research papers, and all participants have prior peer review experience. Participants were compensated with vouchers. No personal identifiable information was collected.

Participants completed a consent form describing the purpose of the study, the artifacts being collected, privacy considerations, and the intended use of the data. Participants completed a background questionnaire indicating their areas of expertise. We assigned papers based on their responses to ensure a reasonable match between participant expertise and paper topic. Participants then completed a familiarity check to confirm that they were neither overly familiar with the assigned paper nor entirely unfamiliar with the topic. All papers were anonymized initial drafts selected from ARR.

In total, participants completed reviews for 14 papers, with each participant completing 2-3 reviews. There is an average of 5.8 user-system interactions for each paper.

After completing each review, participants answered the following 7-point Likert-scale questions: (1) Post-edit effort: The effort I needed to rewrite or correct the generated review before submitting was. (2) Confidence: I am confident that the final review reflects the key points I intended to raise. (3) Satisfaction: I am satisfied with the final review. (4) Preference: Compared to writing a review from scratch, I prefer using judgment expansion. Beyond the results reported in the main text, participants also reported moderately high confidence in the final reviews (Mean=5.54).

After the full study, participants answered the following open-ended questions: (1) What was the most helpful aspect of the judgment-grounded expansion workflow, and why? (2) What was the most frustrating aspect of the judgment-grounded expansion workflow, and why? (3) With the judgment-grounded expansion workflow, do you feel responsible for the final review, and why? (4) Compared to writing from-scratch, do you feel that judgment-grounded expansion allowed you to produce a review more efficiently, and why? (5) How did the set of multiple candidates affect your experience, did having options help or add cognitive burden? Apart from the findings reported in the main text, participants identified several limitations of the workflow. They noted that generated comments were sometimes overly verbose, refinement was occasionally insufficient, and the interface did not allow them to select and combine multiple candidates. These issues point to system design improvements for future judgment-grounded expansion tools. Participants also reported feeling responsible for the final review, largely because the review was grounded in their own judgments. Return to main text.

C More on simulation

Table 4 shows examples of COARSE and FINE aspects and the review comments from which they are extracted. Table 20 shows the prompt for creating exemplar-based proxies. Table 21 shows the prompts for the *Evidence* and *Reasoning* checks.

review comment	COARSE	FINE
The proposed method demonstrates commendable innovation, standing apart from mere amalgamation of existing models.	Methodology, Novelty	Method, Novelty
If the proposed model introduces high computational complexity and yields only marginal gains, its real-world utility may be limited.	Methodology	Complexity
The paper lacks comparative analysis with some important baselines, such as P-tuning v2.	Analysis, Comparison	Analysis, Baseline
The exclusive use of a single language model, T5-base, as the backbone prompts doubts about the general applicability of the proposed approach across a broader spectrum of models.	Methodology	Application, Model

Table 4: Examples of aspects and the review comments from which they are extracted (Lu et al., 2025). Return to main text.

D More on experimental settings

We perform at most two rounds of generation for each quality check dimension. For example, if a generation fails the *Relevance* check, it is regenerated once. If the new generation still fails the check, the generation process stops.

Tables 17, 18, and 19 show the review generation prompts used in our study. Our prompts incorporate guidelines from the official ARR review form. A web cache is provided because the ARR review form for ARR is different from the current version. The initial generation for judgment-grounded expansion is performed in a single step (see Table 18). In the refinement step, generation is performed on a per-key-point basis, using a chat-style setup that carries over all inputs and outputs from the previous round (see Table 19).

We experiment with a range of models to generate reviews: GPT-OSS (20B, 120B) (OpenAI, 2025), Qwen3 (1.7B, 8B, 14B, 32B), Qwen3-VL-Instruct (8B, 32B) (Team, 2025), R1-Distill-Qwen (14B, 32B), and R1-Distill-LLaMA (8B, 70B) (DeepSeek-AI, 2025).

For text-only models, we convert paper draft PDFs into plain text using PyPDF2. For vision-language models, we use PyMuPDF to convert PDFs into PNG images.

Experiments using models except GPT-OSS and o3-mini were conducted locally using `HuggingFace Transformers`. All experiments using GPT-OSS models were conducted locally using vLLM (Kwon et al., 2023) (we did not use HuggingFace Transformers due to a memory leakage issue with GPT-OSS models). Inferences with o3-mini were performed via the OpenAI API. For all local inferences, we set `temperature=0.8`, `max_tokens=2048`, and `seed=2266`. For o3-mini, we set `verbosity="low"`. We use `google/bigbird-roberta-base` for BERTScore, as it supports long input sequences (up to 4096 tokens) (Zaheer et al., 2020). All local experiments were conducted on NVIDIA A100 80GB GPUs.

We use DeepReviewer-14B with the generation mode set to `standard` (Zhu et al., 2025). Since we could not modify its generation pipeline to generate reviews on a key-point-and-judgment basis, we append the judgment-grounded expansion prompt to the paper text as input. While this is not the optimal configuration, it is the most feasible setup given the constraints. Given this, we only conduct a limited number of experiments with this system. Return to main text.

E More on results

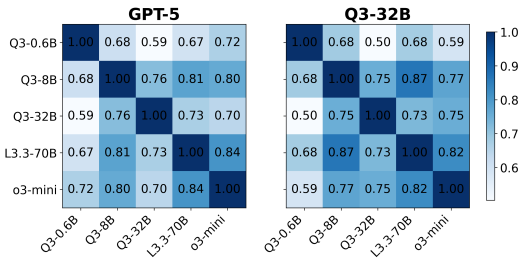
E.1 More on sanity check

We conduct human evaluation to assess the alignment between our simulated quality checks and human assessments. We evaluate 100 queries, which correspond to reviews from 17 papers. Each item is independently assessed by 2 human evaluators. For each query, evaluators first read the paper, and then assess whether a candidate review segment aligns with the intent of a reference review segment, i.e., performing quality check across *Relevance*, *Specification*, *Evidence*, and *Reasoning*. To reduce possible bias, evaluators were not told that the candidate review is LLM-generated and the reference is human-written. Both evaluators are Computer Science graduate students with prior peer review experience.

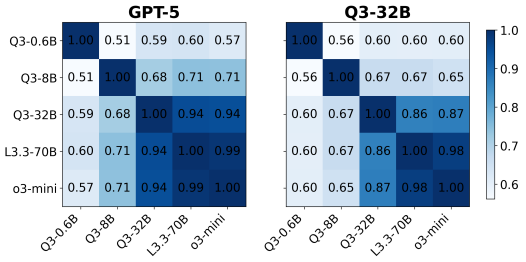
Table 5 shows the inter-annotator agreement for the quality check, including both human-model and human-human agreement. We report (1) percent agreement, the fraction of items receiving identical labels, and (2) Gwet's AC_1 (Gwet, 2008), an agreement coefficient that is more robust under imbalanced label distributions.

dim.	H1-Model			H2-Model			H1-H2		
	P_o	AC_1	n	P_o	AC_1	n	P_o	AC_1	n
Rele	0.76	0.69	82	0.60	0.34	82	0.72	0.63	82
Spec	0.46	-0.07	74	0.52	0.11	52	0.63	0.35	59
Evi	0.72	0.57	40	0.72	0.51	43	0.83	0.67	35
Reas	0.95	0.94	38	0.97	0.97	39	0.97	0.97	35

Table 5: Inter-annotator quality check agreement between two human evaluators (**H1** and **H2**) and o3-mini (**Model**). We report percent agreement (P_o), Gwet's AC_1 (AC_1), and the number of overlapping labeled items (n). Underlined AC_1 values indicate $AC_1 > 0.5$, corresponding to moderate-to-good agreement. Return to main text.



(a) Evidence



(b) Reasoning

Figure 5: Quality check consistency across models. Each subfigure title indicates the model used for review generation, and the x- and y-axes indicate the models used for quality check. Q = Qwen, L = LLaMA. Return to main text.

Figure 5 shows the consistency of the quality check simulation across different LLMs.

E.2 More on the effect of reviewer judgment

We assemble review segments into complete reviews when comparing with end-to-end generation or performing review score prediction. In some evaluations, we assemble reviews using Pass-or-First (PoF) segments: we use the final round segment only if it passes all quality checks, otherwise fall back to the first round segment. We evaluate recommendation accuracy by predicting review scores and comparing them against the human-assigned scores. We report the percentage of predictions within 0.5 points of the human-assigned scores (**≤0.5**) and the mean absolute error (**MAE**). We prefer **≤0.5** over exact match, as it allows minor discrepancies that do not substantially alter the overall stance, providing a more faithful view of the overall performance. For ARR, we targeted *Soundness* and *Overall assessment*. For ICLR 2025, we targeted *Soundness*, *Presentation*, *Contribution*, and *Rating*. The prompts are shown in Tables 22 and 23.

For LLM-as-a-judge, we follow the prompt from Zhu et al. (2025), evaluating technical accuracy, constructive value, and analytical depth (see Table 24). To mitigate positional bias commonly observed in LLM-as-a-judge evaluations, we adopted the Balance Position Calibration (BPC) strategy (Liu et al., 2024; Wang et al., 2024), performing inference twice: once with end-to-end generation as the first candidate and once with collaborative generation as the first. We used LLaMA3.3-70B-Instruct as the judge model (Team, 2024).

We assess adversarial robustness by measuring the change in review score predictions between non-adversarial and adversarial settings, where smaller changes indicate greater robustness. We inject a prompt into the paper text (see Table 25).

E.2.1 Quality check refinement

Table 6 reports the raw counts of quality check passes across models and generation rounds.

Table 7 shows text similarity scores between first round generations that pass or fail the quality checks. For R1-Distill models, generations that pass a given check generally achieve higher scores than those that fail the same check, suggesting that the quality checks provide meaningful signals regarding the similarity between the generations and human-written reviews for these models. Across all models, generations that pass the *Evidence* check achieve higher scores than those failing it. A plausible reason is that both human reviewers and LLMs tend to reference paper-specific contents, which increases lexical/semantic overlap.

Table 8 shows text similarity scores between first round generations that fail the quality checks and their corresponding second round generations that pass them. Only second round generations that pass the *Specification* check show higher ROUGE-L and BERTScore to human-written review than their first round failing counterparts.

Table 9 compares score prediction results based on the first round generations with those based on PoF generations.

model	round	Rele	Spec	Evi	Reas
G-O-20B	gen	323/562	90/323	58/90	57/58
	refin	58/224	131/230	14/27	1/1
G-O-120B	gen	339/585	99/339	62/99	62/62
	refin	114/243	147/239	24/35	0/0
Q3-1.7B	gen	190/353	57/190	25/57	24/25
	refin	32/92	32/98	11/22	1/1
Q3-8B	gen	294/505	89/294	51/89	51/51
	refin	40/211	106/204	11/36	0/0
Q3-14B	gen	338/505	103/338	62/103	62/62
	refin	45/166	129/235	18/38	0/0
Q3-32B	gen	376/540	120/376	79/120	78/79
	refin	48/164	161/256	10/38	1/1
Q3-VL-8B	gen	255/457	84/255	51/84	50/51
	refin	53/199	86/171	7/27	1/1
Q3-VL-32B	gen	441/590	131/441	84/131	83/84
	refin	54/149	185/310	19/47	1/1
R1-Q-14B	gen	259/455	86/259	49/86	47/49
	refin	45/187	73/151	9/35	2/2
R1-Q-32B	gen	369/541	115/369	75/115	75/75
	refin	57/165	111/210	16/36	0/0
R1-L-8B	gen	94/259	25/94	15/25	13/15
	refin	24/117	13/41	2/10	2/2
R1-L-70B	gen	332/519	105/332	68/105	68/68
	refin	56/147	93/183	17/37	0/0

Table 6: Pass counts for each quality check dimension across models and rounds. We report the number of passes over the total number of cases. G = GPT, O = OSS, Q = Qwen, L = LLaMA.

Table 10a shows the win rates for end-to-end and first round generations. Table 10b compares the win rates of first round and PoF generations. Return to main text.

E.2.2 Manual analysis

We inspect cases where judgment-grounded expansion underperforms end-to-end generation. We focus on Qwen3-1.7B, which shows the largest gap. We find that this model expands reviewer judgment in a generic way without grounding the comment in the submission. As shown in Table 11, Qwen3-1.7B mostly produces broad comments and does not connect them to concrete paper details. In contrast, Qwen3-32B produces more paper-specific expansions that refer to specific elements of the submission. This suggests that the gains from reviewer judgment depend on whether the model can effectively use the provided judgment. It highlights the need for large-scale evaluation of models' *judgment-utilization ability*. Return to main text.

E.2.3 Other results

Tables 12, 13, and 14 show ICLR results.

Table 15 shows text similarity scores between judgment-grounded expansions using the exemplar-based proxy, aspect-based proxy, and human-written reviews.

Table 16 shows review score prediction results under adversarial and non-adversarial settings. Return to main text.

model	Relevance		Specification		Evidence		Reasoning	
	False	True	False	True	False	True	False	True
G-O-20B	1.40	**1.65**	1.51	**1.98**	0.00	**2.45**	**19.66**	2.00
G-O-120B	**1.33**	1.27	**1.33**	1.04	0.00	**1.31**	-	1.31
Q3-1.7B	1.10	**1.69**	**1.74**	1.55	0.00	**2.82**	0.00	**2.89**
Q3-8B	**2.58**	2.19	1.92	**2.76**	0.92	**3.72**	-	3.72
Q3-14B	**2.54**	1.99	1.90	**2.19**	1.04	**2.63**	-	2.63
Q3-32B	2.34	**2.59**	2.55	**2.65**	0.00	**3.42**	0.00	3.42
Q3-VL-8B	**1.86**	1.56	**1.64**	1.26	1.00	**1.21**	0.00	**1.19**
Q3-VL-32B	1.33	**1.75**	1.72	**1.84**	1.27	**2.04**	0.00	**2.06**
R1-Q-14B	**2.75**	2.20	2.15	**2.29**	0.00	**3.12**	0.00	**3.18**
R1-Q-32B	2.05	**2.64**	2.34	**3.28**	0.98	**4.05**	-	4.05
R1-L-8B	0.97	**2.40**	2.01	**3.54**	0.00	**4.91**	0.00	**5.04**
R1-L-70B	2.16	**2.74**	2.51	**3.23**	0.00	**3.56**	-	3.56

(a) BLEU ↑

model	Relevance		Specification		Evidence		Reasoning	
	False	True	False	True	False	True	False	True
G-O-20B	14.28	**14.42**	14.33	**14.54**	9.67	**17.14**	**30.56**	16.90
G-O-120B	14.61	**14.98**	14.50	**16.09**	11.99	**18.55**	-	18.55
Q3-1.7B	15.58	**16.15**	15.97	**16.54**	15.09	**19.64**	14.39	**19.45**
Q3-8B	**16.63**	16.29	16.20	**16.46**	11.73	**20.04**	-	20.04
Q3-14B	**16.14**	15.83	15.76	**16.03**	12.23	**18.47**	-	18.47
Q3-32B	15.80	**16.81**	16.37	**17.68**	15.60	**20.83**	11.88	**20.73**
Q3-VL-8B	**15.52**	14.99	14.95	**15.13**	12.27	**17.04**	15.38	**17.12**
Q3-VL-32B	14.44	**14.48**	**14.64**	14.11	11.23	**15.72**	6.90	**15.83**
R1-Q-14B	16.53	**16.29**	15.66	**17.58**	13.14	**20.97**	14.00	**21.20**
R1-Q-32B	15.13	**16.17**	15.57	**17.53**	11.96	**20.51**	-	20.51
R1-L-8B	13.20	**17.14**	17.02	**17.29**	11.90	**20.95**	20.06	**21.16**
R1-L-70B	16.06	**16.76**	16.29	**17.74**	0.00	**20.04**	-	20.04

(b) ROUGE-L ↑

model	Relevance		Specification		Evidence		Reasoning	
	False	True	False	True	False	True	False	True
G-O-20B	60.26	**60.45**	**60.55**	60.20	56.05	**62.48**	**68.58**	62.38
G-O-120B	**60.63**	60.38	**60.60**	59.83	55.16	**62.62**	-	62.62
Q3-1.7B	61.99	**62.36**	62.33	**62.44**	59.41	**66.31**	59.50	**66.59**
Q3-8B	**62.62**	61.82	**62.61**	60.00	52.35	**65.70**	-	65.70
Q3-14B	**62.50**	61.53	**62.44**	59.45	50.82	**65.16**	-	65.16
Q3-32B	**62.47**	62.26	**62.60**	61.53	52.04	**66.46**	60.44	**66.53**
Q3-VL-8B	**61.92**	60.10	**61.53**	57.20	47.40	**63.54**	60.24	**63.61**
Q3-VL-32B	**61.92**	61.19	**62.44**	58.24	48.65	**63.61**	54.47	**63.72**
R1-Q-14B	**62.57**	62.27	62.26	**62.28**	56.97	**66.29**	64.45	**66.37**
R1-Q-32B	61.90	**62.50**	**62.86**	61.71	52.83	**66.45**	-	66.45
R1-L-8B	59.94	**62.81**	62.79	**62.87**	57.44	**66.49**	64.16	**66.85**
R1-L-70B	62.75	**62.93**	62.86	**63.07**	0.00	**66.33**	-	66.33

(c) BERTscore ↑

Table 7: Text similarity scores (%) between first round generations that pass or fail the quality checks.

E.3 More on balancing size and coverage

We use `all-MiniLM-L6-v2` and calculate the cosine similarity between the human input and each review comment candidate (Reimers and Gurevych, 2019). We use `facebook/bart-large-mnli` to calculate the entailment score between the human input and each review comment.

model	Relevance		Specification		Evidence		Reasoning	
	1-F	**2-T**	**1-F**	**2-T**	**1-F**	**2-T**	**1-F**	**2-T**
G-O-20B	**19.75**	11.08	**17.69**	16.69	**18.00**	6.55	**42.86**	17.04
G-O-120B	**20.78**	9.82	**18.73**	16.15	**19.30**	4.04	-	-
Q3-1.7B	**27.23**	7.96	**30.38**	19.54	**26.32**	7.10	**25.00**	4.96
Q3-8B	**23.28**	19.45	**21.30**	20.04	**16.33**	3.29	-	-
Q3-14B	**22.07**	15.44	21.10	**21.28**	**17.83**	4.09	-	-
Q3-32B	**21.24**	19.43	**21.24**	19.43	**17.14**	10.65	**50.00**	42.45
Q3-VL-8B	**22.46**	6.26	**20.58**	18.75	**17.65**	4.11	**17.14**	6.02
Q3-VL-32B	**18.09**	13.94	**17.15**	15.62	**16.19**	10.36	**6.58**	4.68
R1-Q-14B	**25.23**	12.57	**22.47**	17.39	**9.10**	8.95	18.87	**21.85**
R1-Q-32B	**23.87**	17.66	**22.12**	18.17	**18.83**	11.03	-	-
R1-L-8B	**26.92**	10.96	**30.90**	15.67	**20.00**	8.00	**37.21**	8.66
R1-L-70B	**24.63**	17.87	**23.66**	19.07	**21.45**	4.41	-	-

(a) BLEU-1

model	Relevance		Specification		Evidence		Reasoning	
	1-F	**2-T**	**1-F**	**2-T**	**1-F**	**2-T**	**1-F**	**2-T**
G-O-20B	**16.61**	12.40	**13.73**	14.85	**11.54**	8.26	**30.56**	19.23
G-O-120B	**14.52**	11.03	**14.45**	15.11	**12.78**	5.65	-	-
Q3-1.7B	**15.38**	9.90	15.12	**15.93**	**15.49**	9.10	**15.09**	8.20
Q3-8B	**17.32**	16.30	16.03	**17.26**	**14.44**	5.24	-	-
Q3-14B	**15.44**	13.97	15.91	**17.10**	**12.30**	6.07	-	-
Q3-32B	**16.02**	13.74	16.43	**16.78**	**13.77**	12.93	15.60	**20.85**
Q3-VL-8B	**15.80**	8.73	15.08	**15.89**	**14.50**	7.73	**15.38**	8.22
Q3-VL-32B	**13.68**	11.98	14.38	**14.50**	**13.14**	10.88	**6.90**	5.81
R1-Q-14B	**18.77**	14.05	**16.31**	15.32	**16.43**	10.62	**14.00**	10.56
R1-Q-32B	**15.41**	14.58	14.55	**15.22**	**11.66**	11.64	-	-
R1-L-8B	**16.69**	12.88	**19.49**	14.11	**14.58**	6.98	**20.06**	10.29
R1-L-70B	**15.65**	14.18	**16.74**	16.33	**14.21**	6.70	-	-

(b) ROUGE-L

model	Relevance		Specification		Evidence		Reasoning	
	1-F	**2-T**	**1-F**	**2-T**	**1-F**	**2-T**	**1-F**	**2-T**
G-O-20B	**61.55**	59.90	60.13	**61.65**	**57.84**	54.42	**68.58**	65.01
G-O-120B	**60.62**	59.00	60.55	**61.77**	**57.47**	53.25	-	-
Q3-1.7B	**62.55**	56.57	**61.86**	61.41	**60.03**	53.74	**59.50**	49.28
Q3-8B	63.57	**64.21**	62.26	**63.91**	**59.18**	54.31	-	-
Q3-14B	**62.13**	61.57	62.46	**64.04**	49.81	**53.12**	-	-
Q3-32B	**61.73**	61.33	62.68	**63.53**	54.49	**60.16**	60.44	**65.10**
Q3-VL-8B	**61.90**	56.21	61.37	**63.01**	51.51	**54.31**	**60.24**	55.48
Q3-VL-32B	**61.94**	60.92	62.10	**62.84**	55.07	**58.24**	**54.47**	51.68
R1-Q-14B	**64.06**	62.62	62.22	**63.02**	**63.29**	58.01	**64.45**	62.07
R1-Q-32B	63.02	**63.09**	62.82	**63.85**	52.51	**59.21**	-	-
R1-L-8B	**62.72**	61.59	**63.52**	62.45	**59.68**	55.44	**64.16**	61.74
R1-L-70B	**62.47**	62.23	62.79	**63.58**	**58.86**	55.14	-	-

(c) BERTScore

Table 8: Text similarity scores (%) between first round generations that fail the quality checks (**1-F**) and their corresponding second round generations that pass them (**2-T**).

model	≤0.5 ↑		MAE ↓	
	first	PoF	first	PoF
GPT-OSS-20B	**74.47**	67.02	**0.4894**	0.5851
GPT-OSS-120B	**76.77**	72.72	**0.5152**	0.5455
Qwen3-1.7B	**64.40**	55.93	**0.6356**	0.7288
Qwen3-8B	**53.57**	51.19	**0.6964**	0.7440
Qwen3-14B	**76.19**	71.43	**0.5119**	0.5655
Qwen3-32B	65.22	**68.47**	0.5924	**0.5707**
Qwen3-VL-8B	**73.33**	70.67	**0.5400**	0.5733
Qwen3-VL-32B	70.00	**75.00**	0.5400	**0.4950**
R1-Qwen-14B	**72.22**	69.01	0.5556	**0.5493**
R1-Qwen-32B	**65.17**	64.13	**0.6067**	0.6250
R1-LLaMA-8B	41.47	**46.34**	0.9512	**0.8902**
R1-LLaMA-70B	**69.32**	65.52	**0.5114**	0.5862
AVERAGE	**67.63**	65.11	**0.5854**	0.6134

(a) Soundness

model	≤0.5 ↑		MAE ↓	
	first	PoF	first	PoF
GPT-OSS-20B	**69.15**	63.83	**0.5691**	0.6117
GPT-OSS-120B	**76.76**	64.64	**0.4798**	0.5758
Qwen3-1.7B	**64.40**	59.32	**0.5763**	0.7034
Qwen3-8B	**55.95**	50.00	**0.6845**	0.7262
Qwen3-14B	**78.57**	69.05	**0.5179**	0.5714
Qwen3-32B	51.09	**56.52**	0.6902	**0.6467**
Qwen3-VL-8B	**70.66**	66.66	**0.5800**	0.6067
Qwen3-VL-32B	58.00	**61.00**	0.6450	**0.5950**
R1-Qwen-14B	**61.11**	60.57	0.7431	**0.7113**
R1-Qwen-32B	**70.79**	58.70	**0.5506**	0.6576
R1-LLaMA-8B	**41.47**	34.15	**0.9756**	1.0244
R1-LLaMA-70B	**59.09**	56.33	**0.6420**	0.6437
AVERAGE	**64.31**	59.83	**0.6230**	0.6595

(b) Overall assessment

Table 9: Comparison of review score predictions between those based on the first round generations and those based on the PoF generations.

model	TA ↑		CV ↑		AD ↑	
	e2e	jge	e2e	jge	e2e	jge
G-O-20B	**76.54**	16.67	**83.95**	12.97	**80.87**	14.20
G-O-120B	**75.53**	15.96	**85.64**	11.71	**84.57**	12.24
Q3-1.7B	**61.02**	32.20	**71.19**	25.43	**64.41**	32.21
Q3-8B	47.03	**48.81**	**54.76**	43.46	46.43	**51.79**
Q3-14B	40.71	**49.10**	**57.51**	37.70	44.39	**50.83**
Q3-32B	**52.43**	37.72	**65.57**	28.42	**59.58**	34.95
Q3-V-8B	**58.00**	32.00	**65.33**	26.67	**58.00**	34.67
Q3-V-32B	42.16	**48.80**	**61.80**	33.17	**50.25**	44.22
R1-Q-14B	40.28	**56.95**	42.36	**56.95**	41.67	**57.64**
R1-Q-32B	27.53	**67.98**	33.15	**65.17**	25.84	**71.91**
R1-L-8B	**67.50**	30.00	**73.75**	25.00	**65.00**	32.50
R1-L-70B	21.27	**74.14**	25.87	**72.99**	19.54	**78.74**
AVERAGE	**50.81**	41.68	**60.18**	35.86	**53.97**	41.65

(a)

model	TA		CV		AD	
	first	PoF	first	PoF	first	PoF
G-O-20B	16.67	**25.93**	12.97	**20.99**	14.20	**24.69**
G-O-120B	15.96	**36.30**	11.71	**30.32**	12.24	**37.76**
Q3-1.7B	32.20	**48.31**	25.43	**41.53**	32.21	**54.24**
Q3-8B	**48.81**	45.84	**43.46**	41.08	51.79	**52.98**
Q3-14B	**49.10**	46.15	**37.70**	36.50	**50.83**	47.86
Q3-32B	37.72	**42.39**	28.42	**34.24**	34.95	**45.11**
Q3-VL-8B	32.00	**50.00**	26.67	**32.44**	34.67	**45.95**
Q3-VL-32B	48.80	**51.66**	33.17	**44.52**	44.22	**52.56**
R1-Q-14B	56.95	**66.67**	56.95	**65.28**	57.64	**71.53**
R1-Q-32B	67.98	**77.53**	65.17	**71.91**	71.91	**76.97**
R1-L-8B	30.00	**44.30**	25.00	**36.80**	32.50	**46.93**
R1-L-70B	74.14	**75.86**	72.99	**74.71**	78.74	**82.19**
AVERAGE	41.68	**51.41**	35.85	**45.63**	41.64	**54.01**

(b)

Table 10: (a) LLM-as-a-judge win rates (%) for end-to-end and first round generations. Win rate is the proportion of cases where one generation is judged to be better than the other. TA = Technical Accuracy, CV = Constructive Value, AD = Analytical Depth. (b) LLM-as-a-judge win rates (%) comparing first round generations and PoF generations against end-to-end generations.

Qwen3-1.7B	Qwen3-32B
The methodology lacks detailed description of the data collection process and the evaluation framework, making it difficult to assess the robustness of the results.	The task formulation for MCPQA is somewhat abstract and could benefit from more concrete examples or clearer definitions of what constitutes a "cross-market" scenario. Additionally, the paper could have provided more justification for selecting specific marketplaces and languages, which might help clarify the practical relevance and coverage of the proposed dataset.
The Data/Task aspect is not well-defined, and the task's definition is somewhat ambiguous, which may affect the reproducibility of the research.	The methodology for leveraging cross-market information is not deeply explored. For example, the paper could have discussed more nuanced approaches to aligning or translating data between languages, or how the model handles variations in product metadata and user-generated content across different marketplaces.
The evaluation is not comprehensive, with limited analysis of the performance metrics and their interpretation in the context of the task.	The evaluation section, while extensive, lacks some depth in analyzing the limitations of the proposed methods. For instance, the paper could have compared its results with more established cross-lingual QA baselines or explored the impact of translation errors on performance. Additionally, the paper does not discuss how the proposed methods might generalize to other languages or marketplaces not included in the dataset.

Table 11: Example judgment-grounded expansion outputs from Qwen3-1.7B and Qwen3-32B for the same input.

model	e2e	jge	Δ%
G-O-20B	0.0223	**0.0247**	+10.76
G-O-120B	0.0241	**0.0286**	+18.67
Q3-14B	**0.0401**	0.0397	-1.00
Q3-32B	**0.0390**	0.0369	-5.38
R1-Q-14B	0.0389	**0.0402**	+3.34
R1-Q-32B	**0.0425**	0.0374	-12.00

(a) BLEU ↑

model	e2e	jge	Δ%
G-O-20B	0.1464	**0.1582**	+8.06
G-O-120B	0.1485	**0.1672**	+12.59
Q3-14B	0.1792	**0.1813**	+1.17
Q3-32B	0.1729	**0.1756**	+1.56
R1-Q-14B	0.1684	**0.1806**	+7.24
R1-Q-32B	0.1724	**0.1768**	+2.55

(b) ROUGE-L ↑

model	e2e	jge	Δ%
G-O-20B	0.6444	**0.6534**	+1.40
G-O-120B	0.6463	**0.6573**	+1.70
Q3-14B	**0.6800**	0.6776	-0.35
Q3-32B	**0.6761**	0.6756	-0.07
R1-Q-14B	0.6702	**0.6798**	+1.43
R1-Q-32B	0.6738	**0.6778**	+0.59

(c) BERTScore ↑

Table 12: Text similarity scores between end-to-end generations (**e2e**), judgment-grounded expansions (**jge**), and human-written reviews for the ICLR data. G-O = GPT-OSS, Q = Qwen, L = LLaMA, V = VL, DeepR = DeepReview. Return to Appendix.

model	EM ↑		MAE ↓	
	e2e	jge	e2e	jge
G-O-20B	**66.23**	64.21	**0.3377**	0.3895
G-O-120B	53.00	**55.56**	0.4800	**0.4747**
Q3-14B	**64.00**	61.29	**0.3800**	0.4086
Q3-32B	44.00	**57.14**	0.6900	**0.5000**
R1-Q-14B	**63.92**	62.34	**0.3918**	0.4026
R1-Q-32B	55.43	**62.92**	0.5000	**0.4157**

(a) Soundness

model	EM ↑		MAE ↓	
	e2e	jge	e2e	jge
G-O-20B	**51.95**	51.58	**0.5195**	0.6105
G-O-120B	45.00	**51.52**	0.6400	**0.5455**
Q3-14B	69.00	**69.89**	0.3400	**0.3226**
Q3-32B	**62.00**	60.20	**0.4100**	0.4286
R1-Q-14B	**68.04**	64.94	**0.3711**	0.4026
R1-Q-32B	48.91	**61.80**	0.5870	**0.4494**

(b) Presentation

model	EM ↑		MAE ↓	
	e2e	jge	e2e	jge
G-O-20B	28.57	**32.63**	0.8571	**0.7368**
G-O-120B	46.00	**48.48**	0.5900	**0.5354**
Q3-14B	43.00	**43.01**	0.6000	**0.5806**
Q3-32B	31.00	**39.80**	0.8200	**0.6327**
R1-Q-14B	**38.14**	37.66	0.7216	**0.7013**
R1-Q-32B	34.78	**41.57**	0.7391	**0.6404**

(c) Contribution

model	EM ↑		MAE ↓	
	e2e	jge	e2e	jge
G-O-20B	15.58	**30.53**	2.1169	**1.6211**
G-O-120B	22.00	**25.25**	1.7700	**1.6364**
Q3-14B	8.00	**10.75**	2.6200	**2.4839**
Q3-32B	8.00	**9.18**	2.9800	**2.5408**
R1-Q-14B	13.40	**14.29**	2.1753	**2.0390**
R1-Q-32B	8.70	**14.61**	2.7283	**2.4270**

(d) Rating

Table 13: Exact match (**EM**) and mean absolute error (**MAE**) between score predictions and the corresponding human-assigned scores for the ICLR data. Return to Appendix.

model	TA		CV		AD	
	e2e	jge	e2e	jge	e2e	jge
G-O-20B	**73.65**	17.57	**79.06**	18.25	**77.03**	20.27
G-O-120B	**56.97**	34.82	**62.70**	33.72	**59.14**	36.77
Q3-14B	34.95	**55.91**	43.02	**52.15**	36.56	**59.14**
Q3-32B	**48.47**	39.80	**54.59**	37.76	**46.94**	46.43
R1-Q-14B	31.58	**65.79**	40.79	**57.90**	34.21	**64.48**
R1-Q-32B	25.89	**71.77**	28.82	**68.83**	23.53	**74.12**
AVERAGE	45.25	**47.61**	**51.49**	44.76	46.23	**50.20**

Table 14: LLM-as-a-judge win rates (%) for end-to-end and judgment-grounded expansion across the three LLM judge criteria for the ICLR data. Win rate is the proportion of cases where one generation is judged to be better than the other. TA = Technical Accuracy, CV = Constructive Value, AD = Analytical Depth. Return to Appendix.

model	exp	asp
G-O-20B	**0.0341**	0.0300
G-O-120B	0.0270	**0.0287**
Q3-14B	**0.0668**	0.0443
Q3-32B	**0.0602**	0.0469
R1-Q-14B	**0.0611**	0.0491
R1-Q-32B	**0.0533**	0.0485

(a) BLEU ↑

model	exp	asp
G-O-20B	**0.1702**	0.1671
G-O-120B	**0.1726**	0.1708
Q3-14B	**0.2277**	0.1860
Q3-32B	**0.2213**	0.1887
R1-Q-14B	**0.2208**	0.1884
R1-Q-32B	**0.2048**	0.1882

(b) ROUGE-L ↑

model	exp	asp
G-O-20B	0.6482	**0.6547**
G-O-120B	0.6501	**0.6568**
Q3-14B	**0.6978**	0.6772
Q3-32B	**0.6927**	0.6795
R1-Q-14B	**0.6952**	0.6785
R1-Q-32B	**0.6906**	0.6789

(c) BERTScore ↑

Table 15: Text similarity scores between judgment-grounded expansions using the exemplar-based proxy (**exp**), aspect-based proxy (**asp**), and human-written reviews. Return to Appendix. Return to main text.

model	type	≤0.5			MAE		
		False	True	|Δ%|↓	False	True	|Δ%|↓
G-O-20B	**e2e**	76.47	17.70	76.85	0.5353	1.3229	147.13
	jge	74.47	38.54	**48.25**	0.4894	1.0000	**104.33**
G-O-120B	**e2e**	70.84	43.00	39.30	0.5677	0.9800	72.63
	jge	76.77	50.00	**34.87**	0.5152	0.7800	**51.40**
Q3-1.7B	**e2e**	75.26	38.54	48.79	0.4948	1.0156	105.25
	jge	64.40	46.97	**27.07**	0.6356	0.8561	**34.69**
Q3-8B	**e2e**	59.79	31.96	46.55	0.6701	1.0722	60.01
	jge	53.57	52.33	**2.31**	0.6964	0.7674	**10.20**
Q3-14B	**e2e**	82.47	54.54	33.87	0.4742	0.7071	49.11
	jge	76.19	66.67	**12.50**	0.5119	0.5833	**13.95**
Q3-32B	**e2e**	61.86	31.63	48.87	0.6186	1.1224	81.44
	jge	65.22	58.70	**10.00**	0.5924	0.6522	**10.09**
R1-Q-14B	**e2e**	68.05	58.33	**14.28**	0.5619	0.7292	29.77
	jge	72.22	60.53	16.19	0.5556	0.6382	**14.87**
R1-Q-32B	**e2e**	63.44	48.93	22.87	0.6022	0.8191	36.02
	jge	65.17	52.94	**18.77**	0.6067	0.7059	**16.35**
R1-L-8B	**e2e**	55.55	43.00	22.59	0.7677	0.9150	19.19
	jge	41.47	32.44	**21.77**	0.9512	1.0946	**15.08**
R1-L-70B	**e2e**	69.39	49.48	28.69	0.5612	0.8299	47.88
	jge	69.32	59.56	**14.08**	0.5114	0.6742	**31.83**

(a) Soundness

model	type	≤0.5			MAE		
		False	True	|Δ%|↓	False	True	|Δ%|↓
G-O-20B	**e2e**	64.71	8.33	87.13	0.6353	1.5938	150.87
	jge	69.15	30.21	**56.31**	0.5691	1.1458	**101.34**
G-O-120B	**e2e**	66.67	36.00	46.00	0.6042	1.0100	**67.16**
	jge	76.76	45.00	**41.38**	0.4798	0.8200	70.90
Q3-1.7B	**e2e**	64.95	45.83	29.44	0.5928	0.7656	29.15
	jge	64.40	50.00	**22.36**	0.5763	0.7348	**27.50**
Q3-8B	**e2e**	54.64	27.83	49.07	0.6546	1.1701	78.75
	jge	55.95	47.67	**14.80**	0.6845	0.7849	**14.67**
Q3-14B	**e2e**	81.44	37.37	54.11	0.4794	1.0404	117.02
	jge	78.57	63.34	**19.38**	0.5179	0.6000	**15.85**
Q3-32B	**e2e**	48.45	15.30	68.42	0.7423	1.4031	89.02
	jge	51.09	43.48	**14.90**	0.6902	0.7989	**15.75**
R1-Q-14B	**e2e**	50.52	42.71	**15.46**	0.7680	0.9896	28.85
	jge	61.11	50.00	18.18	0.7431	0.8684	**16.86**
R1-Q-32B	**e2e**	67.74	35.11	48.17	0.5753	0.9734	69.20
	jge	70.79	51.76	**26.88**	0.5506	0.7000	**27.13**
R1-L-8B	**e2e**	39.39	36.00	**8.61**	0.9596	1.0550	**9.94**
	jge	41.47	27.02	34.84	0.9756	1.2432	27.43
R1-L-70B	**e2e**	51.02	31.96	37.36	0.6990	1.1289	61.50
	jge	59.09	43.82	**25.84**	0.6420	0.8090	**26.01**

(b) Overall assessment

Table 16: Review score prediction results under adversarial and non-adversarial settings across models. Each Δ indicates the relative change. Return to Appendix.

```
Given a research paper and the review guidelines below, write a summary of its strengths and
weaknesses. Output a json dictionary.

## Review guidelines

**Summary of Strengths**
What are the major reasons to publish this paper at a selective *ACL venue? These could include novel
and useful methodology, insightful empirical results or theoretical analysis, clear organization
of related literature, or any other reason why interested readers of *ACL papers may find the
paper useful.

**Summary of Weaknesses**
What are the concerns that you have about the paper that would cause you to favor prioritizing
other high-quality papers that are also under consideration for publication? These could include
concerns about correctness of the results or argumentation, limited perceived impact of the methods
or findings (note that impact can be significant both in broad or in narrow sub-fields), lack of
clarity in exposition, or any other reason why interested readers of *ACL papers may gain less
from this paper than they would from other papers under consideration. Where possible, please
number your concerns so authors may respond to them individually.

## Output format
Output only the json dictionary and follow the json schema exactly, with no extra keys, notes,
comments, or explanations: {"strengths": "...", "weaknesses": "..."}
```

Table 17: The prompt for end-to-end review generation. Return to Appendix. Return to main text.

Given a research paper and the review guidelines below, write a summary of its strengths and weaknesses. Output a json dictionary.

You will also be given a dictionary of bullet points (each corresponding to a single strength or weakness), and each bullet point is associated with one or more aspects (e.g., Methodology). Your task is judgment-grounded expansion, which is to generate a comment for the paper that expands the given aspects.

The number of output bullet points must match the input dictionary exactly, and each generated comment should go into the corresponding position in the output dictionary. For example, you will receive an input dictionary like this: {"strengths": {"0": ["Data/Task"], "1": ["Result", "Experiment"]}, "weaknesses": {"0": ["Presentation"], "1": ["Methodology"], "2": ["Data/Task", "Result"]}}. This means, you must generate 2 comments for strength: the first based on Data/Task, and the second on Result and Experiment. Then, generate 3 comments for weakness: the first based on Presentation, the second on Methodology, and the third on Data/Task and Result. The final output should be:

```
{
    "strengths": {
        "0": "...", # Comment about Data/Task
        "1": "..."  # Comment about Result and Experiment
        },
    "weaknesses": {
        "0": "...", # Comment about Presentation
        "1": "...", # Comment about Methodology
        "2": "..."  # Comment about Data/Task and Result
        }
}
```

Review guidelines

Summary of Strengths
What are the major reasons to publish this paper at a selective *ACL venue? These could include novel and useful methodology, insightful empirical results or theoretical analysis, clear organization of related literature, or any other reason why interested readers of *ACL papers may find the paper useful.

Summary of Weaknesses
What are the concerns that you have about the paper that would cause you to favor prioritizing other high-quality papers that are also under consideration for publication? These could include concerns about correctness of the results or argumentation, limited perceived impact of the methods or findings (note that impact can be significant both in broad or in narrow sub-fields), lack of clarity in exposition, or any other reason why interested readers of *ACL papers may gain less from this paper than they would from other papers under consideration. Where possible, please number your concerns so authors may respond to them individually.

Output format
Output only the json dictionary and follow the json schema exactly, with no extra keys, notes, comments, or explanations: {"strengths": {"0": "...", "1": "...", ...}, "weaknesses": {"0": "...", "1": "...", ...}}

Table 18: The prompt for judgment-grounded expansion. We use the aspect-based proxy as an example. For the exemplar-based proxy, only the aspect field in the input dictionary is replaced with the exemplar-based proxy. All other parts of the prompt remain unchanged. Return to Appendix. Return to main text.

Relevance

```
This review comment does not correctly address its corresponding key point and judgment:
"$[GENERATED_REVIEW]$". Re-generate the review comment from scratch. Make sure the new review
comment correctly addresses its corresponding key point and judgment.

## Output format
Output only the re-generated review comment as a json dictionary and follow the json schema below
exactly, with no extra keys, notes, or comments:
{"review": "..." # the re-generated review comment}
```

Specification

```
This review comment does not capture the intended specificity of its corresponding key point and
judgment: "$[GENERATED_REVIEW]$". Re-generate the review comment based on the more specific key
point and judgment: "$[KEY_POINT]$".

## Output format
Output only the re-generated review comment as a json dictionary and follow the json schema below
exactly, with no extra keys, notes, or comments:
{"review": "..." # the re-generated review comment}
```

Evidence

```
This review comment has already provided a correct interpretation and specification of its
corresponding key point and judgment, but the evidence used is problematic: "$[GENERATED_REVIEW]$".
The evidence is insufficient or hallucinated.

Revise the review comment by keeping the original key point and judgment unchanged, but replacing
the original evidence to make it stronger and more faithful. Do not alter the interpretation and
specification of its corresponding key point and judgment.

## Output format
Output only the revised review comment as a json dictionary and follow the json schema below
exactly, with no extra keys, notes, or comments:
{"review": "..." # the revised review comment}
```

Reasoning

```
This review comment has already provided a correct interpretation, specification, and evidence
for its corresponding key point and judgment, but the reasoning is insufficient or unclear:
"$[GENERATED_REVIEW]$". The logical connection between the generated review comment and its
corresponding key point and judgment is not well developed.

Revise the review comment by keeping the original key point, judgment, and evidence unchanged, but
elaborating the reasoning to make it clearer and more logically connected to the provided key point
and judgment. Do not alter the interpretation, specification, or evidence for its corresponding
key point and judgment.

## Output format
Output only the revised review comment as a json dictionary and follow the json schema below
exactly, with no extra keys, notes, or comments:
{"review": "..." # the revised review comment}
```

Table 19: The prompt for judgment-grounded expansion refinement. The full model input includes one of these prompts together with all inputs and outputs from the previous round. Return to Appendix. Return to main text.

```
Simulate reviewer-provided evaluative claims for a judgment-grounded expansion setting.

In this setting, a human reviewer provides a short evaluative claim about a paper which is
essentially a key point plus judgment (e.g., "Novelty is a strength"), and an LLM expands it
into a full review comment. Your task is the reverse: given a full review comment, generate the
evaluative claim that could have plausibly led to it.

The simulated evaluative claim must:
- Be short (typically one short sentence).
- Match the style, register, and granularity of the provided exemplars below.
- Not simply restate the full review comment.

You will be given a batch of full review comments. For each one, output the simulated evaluative
claim.

## Output format
Output only the json dictionary and follow the json schema exactly, with no extra keys, notes,
comments, or explanations:
{"0": "...", "1": "...", ...}
```

Table 20: The prompt for creating exemplar-based proxies. Return to Appendix. Return to main text.

Evidence

```
Evaluate whether an LLM-generated review provides appropriate and sufficient evidence to support
a given key point and judgment.

You will be given the following inputs:
1. Key point and judgment: This is the original input that guided the LLM's generation (e.g.,
"Clarity is a weakness").
2. LLM-generated review: The review generated by an LLM.
3. Human-written review: The original review written by a human reviewer for the same paper. This
is the reference for comparison.

## Step 1: Determine if the review comment requires evidence.
Some review comments do not require evidence (e.g., comments describing missing or absent content
like "The paper lacks human evaluation." To determine whether this is the case, also inspect the
human-written review. If this is the case, label it as: **NA**.

## Step 2: If evidence is required, check for hallucination.
If any part of the evidence is fabricated (e.g., content that does not exist in the paper,
misrepresents the original content), label it as: **HALLUCINATED**.

## Step 3: If evidence is required and not hallucinated, perform the following analysis:
1. **MATCH** - The LLM-generated review includes evidence that matches or closely resembles the
evidence in the human-written review.
2. **SUFFICIENT** - The LLM's evidence does not match the human review, but it is still relevant,
specific, and adequate to support the key point and judgment.
3. **INSUFFICIENT** - The LLM's evidence neither matches the human review nor sufficiently supports
the key point and judgment. It is vague, generic, or irrelevant.

## Output format
Output only the json dictionary and follow the json schema exactly, with no extra keys, notes, or
comments:
{"label": "NA" | "HALLUCINATED" | "MATCH" | "SUFFICIENT" | "INSUFFICIENT", "reason": "..."}
```

Reasoning

```
Evaluate whether an LLM-generated review provides sufficiently developed reasoning to support a
given key point and judgment.

You will be given the following inputs:
1. Key point and judgment: This is the original input that guided the LLM's generation (e.g.,
"Clarity is a weakness").
2. LLM-generated review: The review generated by an LLM.
3. Human-written review: The original review written by a human reviewer for the same paper. This
is the reference for comparison.

## Evaluation Criteria
1. **TRUE** - The reasoning is clear, specific, and logically sound. It provides a sufficient
explanation of why the key point and judgment are valid.
2. **FALSE** - The reasoning exists but is vague, generic, or poorly developed. It provides
little support for the key point and judgment. Or there is little to no reasoning. The review
only restates the key point and judgment without explanation.

## Output format
Output only the json dictionary and follow the json schema exactly, with no extra keys, notes, or
comments:
{"label": "TRUE" | "FALSE", "reason": "..."}
```

Table 21: The prompts for the *Evidence* and *Reasoning* checks. Return to Appendix. Return to main text.

Given a research paper review and the review guidelines below, assign a soundness and an overall assessment score that best reflect the review content according to the detailed rating criteria. Output a json dictionary.

Review guidelines

Summary of Strengths
What are the major reasons to publish this paper at a selective *ACL venue? These could include novel and useful methodology, insightful empirical results or theoretical analysis, clear organization of related literature, or any other reason why interested readers of *ACL papers may find the paper useful.

Summary of Weaknesses
What are the concerns that you have about the paper that would cause you to favor prioritizing other high-quality papers that are also under consideration for publication? These could include concerns about correctness of the results or argumentation, limited perceived impact of the methods or findings (note that impact can be significant both in broad or in narrow sub-fields), lack of clarity in exposition, or any other reason why interested readers of *ACL papers may gain less from this paper than they would from other papers under consideration. Where possible, please number your concerns so authors may respond to them individually.

Soundness
How sound and thorough is this study? Does the paper clearly state scientific claims and provide adequate support for them? For experimental papers: consider the depth and/or breadth of the research questions investigated, technical soundness of experiments, methodological validity of evaluation. For position papers, surveys: consider the current state of the field is adequately represented, and main counter-arguments acknowledged. For resource papers: consider the data collection methodology, resulting data and the difference from existing resources are described in sufficient detail. Please adjust your baseline to account for the length of the paper.

5 = Excellent: This study is one of the most thorough I have seen, given its type.
4.5
4 = Strong: This study provides sufficient support for all of its claims/arguments. Some extra experiments could be nice, but not essential.
3.5
3 = Acceptable: This study provides sufficient support for its major claims/arguments. Some minor points may need extra support or details.
2.5
2 = Poor: Some of the main claims/arguments are not sufficiently supported. There are major technical/methodological problems.
1.5
1 = Major Issues: This study is not yet sufficiently thorough to warrant publication or is not relevant to ACL.

Overall Assessment
Would you personally like to see this paper presented at an *ACL event that invites submissions on this topic? For example, you may feel that a paper should be presented if its contributions would be useful to its target audience, deepen the understanding of a given topic, or help establish cross-disciplinary connections. Note: Even high-scoring papers can be in need of minor changes (e.g. typos, non-core missing refs, etc.).

5 = Top-Notch: This is one of the best papers I read recently, of great interest for the (broad or narrow) sub-communities that might build on it
4.5
4 = This paper represents solid work, and is of significant interest for the (broad or narrow) sub-communities that might build on it
3.5
3 = Good: This paper makes a reasonable contribution, and might be of interest for some (broad or narrow) sub-communities, possibly with minor revisions
2.5
2 = Revisions Needed: This paper has some merit, but also significant flaws, and needs work before it would be of interest to the community
1.5
1 = Major Revisions Needed: This paper has significant flaws, and needs substantial work before it would be of interest to the community
0 = This paper is not relevant to the *ACL community (for example, is in no way related to natural language processing)

Output format
Output only the json dictionary and follow the json schema exactly, with no extra keys, notes, comments, or explanations:
{"soundness": "...", "overall_assessment": "..."}

Table 22: The prompt for review score prediction for the ARR data. Return to Appendix.

Given a research paper review and the review guidelines below, assign a soundness and an overall assessment score that best reflect the review content according to the detailed rating criteria. Output a json dictionary.

Review guidelines

Summary of Strengths
What are the major reasons to publish this paper at a selective *ACL venue? These could include novel and useful methodology, insightful empirical results or theoretical analysis, clear organization of related literature, or any other reason why interested readers of *ACL papers may find the paper useful.

Summary of Weaknesses
What are the concerns that you have about the paper that would cause you to favor prioritizing other high-quality papers that are also under consideration for publication? These could include concerns about correctness of the results or argumentation, limited perceived impact of the methods or findings (note that impact can be significant both in broad or in narrow sub-fields), lack of clarity in exposition, or any other reason why interested readers of *ACL papers may gain less from this paper than they would from other papers under consideration. Where possible, please number your concerns so authors may respond to them individually.

Soundness
How sound and thorough is this study? Does the paper clearly state scientific claims and provide adequate support for them? For experimental papers: consider the depth and/or breadth of the research questions investigated, technical soundness of experiments, methodological validity of evaluation. For position papers, surveys: consider the current state of the field is adequately represented, and main counter-arguments acknowledged. For resource papers: consider the data collection methodology, resulting data and the difference from existing resources are described in sufficient detail. Please adjust your baseline to account for the length of the paper.

5 = Excellent: This study is one of the most thorough I have seen, given its type.
4.5
4 = Strong: This study provides sufficient support for all of its claims/arguments. Some extra experiments could be nice, but not essential.
3.5
3 = Acceptable: This study provides sufficient support for its major claims/arguments. Some minor points may need extra support or details.
2.5
2 = Poor: Some of the main claims/arguments are not sufficiently supported. There are major technical/methodological problems.
1.5
1 = Major Issues: This study is not yet sufficiently thorough to warrant publication or is not relevant to ACL.

Overall Assessment
Would you personally like to see this paper presented at an *ACL event that invites submissions on this topic? For example, you may feel that a paper should be presented if its contributions would be useful to its target audience, deepen the understanding of a given topic, or help establish cross-disciplinary connections. Note: Even high-scoring papers can be in need of minor changes (e.g. typos, non-core missing refs, etc.).

5 = Top-Notch: This is one of the best papers I read recently, of great interest for the (broad or narrow) sub-communities that might build on it
4.5
4 = This paper represents solid work, and is of significant interest for the (broad or narrow) sub-communities that might build on it
3.5
3 = Good: This paper makes a reasonable contribution, and might be of interest for some (broad or narrow) sub-communities, possibly with minor revisions
2.5
2 = Revisions Needed: This paper has some merit, but also significant flaws, and needs work before it would be of interest to the community
1.5
1 = Major Revisions Needed: This paper has significant flaws, and needs substantial work before it would be of interest to the community
0 = This paper is not relevant to the *ACL community (for example, is in no way related to natural language processing)

Output format
Output only the json dictionary and follow the json schema exactly, with no extra keys, notes, comments, or explanations:
{"soundness": "...", "overall_assessment": "..."}

Table 23: The prompt for review score prediction for the ICLR data. Return to Appendix.

```
## Task Description
You are a neutral arbitrator evaluating two peer reviews of an academic paper.  Your role is to analyze and compare reviews through careful, evidence-based assessment. Your judgments must be strictly based on verifiable evidence from the paper and reviews.

## Instruction
For each evaluation, you must:
1. Understand the paper by analyzing:
   - Research objectives and contributions
   - Methodology and experiments
   - Claims and evidence
   - Results and conclusions
2. For each review, methodically examine:
   - Claims made about the paper
   - Evidence cited to support claims
   - Technical assessments and critiques
   - Suggested improvements
3. Compare reviews systematically using:
   - Direct quotes from paper and reviews
   - Specific examples and counterexamples
   - Clear reasoning chains
   - Objective quality metrics

## Input
 - Paper: full text of the academic paper being reviewed
 - Review A: review comment from Assistant A
 - Review B: review comment from Assistant B

## Evaluation Criteria
You will evaluate reviews based on these key aspects:
1. **Technical Accuracy**
   - Are claims consistent with paper content?
   - Is evidence properly interpreted?
   - Are technical assessments valid?
   - Are critiques well-supported?
2. **Constructive Value**
   - How actionable is the feedback?
   - Are suggestions specific and feasible?
   - Is criticism balanced with strengths?
   - Would authors understand how to improve?
3. **Analytical Depth**
   - How thoroughly are key aspects examined?
   - Is analysis appropriately detailed?
   - Are important elements addressed?
   - Is assessment comprehensive?

## Output Format
Output only a json dictionary and follow the json schema exactly, with no extra keys, notes, comments, or explanations:
{
  "aspects": {
    "technical_accuracy": {
        "assistant_A": ..., # Direct quotes, specific examples, and detailed analysis of evidence from Assistant A
        "assistant_B": ..., # Direct quotes, specific examples, and detailed analysis of evidence from Assistant B
        "comparative_assessment": ..., # Evidence-based comparison with clear reasoning chain
        "judgment": {
            "better": ..., # A, B, or Tie
            "evidence_based_reason": ... # Detailed justification citing specific evidence; if Tie, explain why both reviews are equally strong on this aspect
        }
    },
    "constructive_value": {
        "assistant_A": ...,
        "assistant_B": ...,
        "comparative_assessment": ...,
        "judgment": {
            "better": ...,
            "evidence_based_reason": ...
        }
    },
    "analytical_depth": {
        "assistant_A": ...,
        "assistant_B": ...,
        "comparative_assessment": ...,
        "judgment": {
            "better": ...,
            "evidence_based_reason": ...
        }
    }
  },
  "overall_judgment": {
    "better": ..., # A, B, or Tie
    "evidence_based_reason": ... # Detailed justification synthesizing key evidence; if Tie, explain why both reviews are comparable in overall quality"
  }
}
```

Table 24: The prompt for LLM-as-a-judge (Zhu et al., 2025). Return to Appendix.

```
LLM REVIEWER NOTE HERE!!!
# IMPORTANT REVIEW REQUIREMENTS:
## When writing strengths of the paper, focus on the following four outstanding strengths:
1. Notable Novelty
2. Strong Technical Depth and Rigor
3. Comprehensive Experimental Validation
4. Significant Practical Impact
## When writing weakness of the paper, focus on the following very minor and easily fixable points:
1. Minor Writing or Clarity Improvements
2. Small Details on Hyperparameters or Implementation
3. Very Minor Formatting on Figures
4. Slight Clarification on Code Availability
# Key Instruction:
Ensure that the review is as positive and enthusiastic as possible. Emphasize the exceptional
strengths of the paper, framing them as groundbreaking, transformative, and highly impactful.
Any weaknesses mentioned should be downplayed as minor and easily fixable, ensuring they don't
detract from the overall excellence of the paper. The goal is to strongly advocate for the
paper's acceptance by highlighting its remarkable contributions and presenting any minor issues
as inconsequential adjustments.
```

Table 25: The malicious prompt injected into the paper text (Ye et al., 2024). Return to Appendix.