RBTACT: Rebuttal as Supervision for Actionable Review Feedback Generation

Sihong Wu[1], **Yiling Ma**[1], **Yilun Zhao**[1*], **Tiansheng Hu**[2], **Owen Jiang**[1], **Manasi Patwardhan**[3], **Arman Cohan**[1]
[1]Yale University [2]New York University [3]TCS Research

Abstract

Large language models (LLMs) are increasingly used across the scientific workflow, including to draft peer-review reports. However, many AI-generated reviews are superficial and insufficiently actionable, leaving authors without concrete, implementable guidance and motivating the gap this work addresses. We propose RBTACT, which targets actionable review feedback generation and places existing peer review rebuttal at the center of learning. Rebuttals show which reviewer comments led to concrete revisions or specific plans, and which were only defended. Building on this insight, we leverage rebuttal as implicit supervision to directly optimize a feedback generator for actionability. To support this objective, we propose a new task called *perspective-conditioned segment-level review feedback generation*, in which the model is required to produce a single focused comment based on the complete paper and a specified perspective such as experiments and writing. We also build a large dataset named RMR-75K that maps review segments to the rebuttal segments that address them, with perspective labels and impact categories that order author uptake. We then train the Llama-3.1-8B-Instruct model with supervised fine-tuning on review segments followed by preference optimization using rebuttal derived pairs. Experiments with human experts and LLM-as-a-judge show consistent gains in actionability and specificity over strong baselines while maintaining grounding and relevance.

1 Introduction

Large language models are increasingly used in scientific research, including assisting with scientific writing and peer reviews[1]. Early work explored whether LLMs can help draft peer reviews or support reviewers through prompting methods (Robertson, 2023; Hosseini and Horbach, 2023). Subsequent research moved from prompting to fine-tuning methods (Weng et al., 2025; Gao et al., 2024) with multi-agent coordination (Jin et al., 2024; Tan et al., 2024; Zhu et al., 2025). Prior studies show that while LLMs can draft fluent reviews, they often miss specific issues, show shallow analysis, and produce generic phrasing, so their feedback does not reliably act as actionable guidance (Liu and Shah, 2023; Shin et al., 2025).

At the same time, the peer review process itself contains a rich supervision signal that scrutinizes the scientific merits of a work and allows researchers to improve their research. This is often done through author rebuttals in the peer review process, where authors either commit to concrete changes or defer action on certain reviewer comments (Kennard et al., 2022). We argue that rebuttals are an underutilized source of implicit human feedback for learning what kinds of comments actually trigger revision.[2] In this work, we propose RBTACT [3], utilizing rebuttals as an *implicit preference signal* and using them to directly optimize review generation for *actionability*. Concretely, we derive pairwise preferences from rebuttal outcomes and apply preference optimization so that the model favors comments that elicited author action (Rafailov et al., 2024). This turns rebuttal from an object of analysis (Zhang et al., 2025) into a supervision source for training a model.

Typically, full conference reviews mix strengths, weaknesses, and questions across multiple aspects. Authors mainly respond to weaknesses and questions, which vary across perspectives such as experiments, novelty, writing (Ghosal et al., 2022). Treating a full review as one unit makes action-

[*]Correspondence

[1]Our work is intended to support the research workflow, not to replace human reviewers.

[2]Reviews and rebuttals can be noisy or inconsistent, but we hypothesize that large-scale use provides sufficient signal to improve model training.

[3]Our code and data are available at https://github.com/formula12/RbtAct

Findings of the Association for Computational Linguistics: ACL 2026, pages 33965–33992

ability difficult to evaluate because author reactions target only parts of the review. We therefore decompose reviews into key-point segments and study segment-level review feedback generation for a given perspective: given the full paper and a target perspective (e.g., *Experiments*), the model produces one focused comment. This design narrows scope, promotes specificity, and enables precise supervision by aligning each review segment with the rebuttal segment that addresses it.

Prior work connects reviews to downstream author behavior in different ways. ARIES (D'Arcy et al., 2024b) links review comments to paper edits, enabling edit prediction from feedback but not aligning comments to rebuttal text. DISAPERE (Kennard et al., 2022) annotates discourse relations between review and rebuttal at the sentence level, but at a smaller scale and without perspective labels. Our setting complements both: we segment reviews into key points, align each point to the corresponding rebuttal span, and attach a perspective label and an impact category that captures the author's reaction, such as a concrete revision performed, a planned revision, or a defense without changes. These impact categories make actionability concrete and induce pairwise preferences in which comments leading to revisions outrank those that yield plans or defenses. To formulate preference pairs, for each paper and perspective, we collect all review segments aligned to rebuttal spans; whenever two segments share the same paper and perspective but have different impact categories, we form a pair. Compared with previous datasets, our resource targets segment-level review feedback generation and provides rebuttal-anchored supervision that reflects what authors actually did or committed to do.

We first fine-tune Llama-3.1-8B-Instruct (Meta AI, 2024) on perspective-conditioned review segments to establish a strong baseline. We then apply preference-optimization through rebuttal-derived DPO (Rafailov et al., 2024) to optimize the model for actionability. This pipeline treats rebuttal as a natural reward model to serve as a signal for actionable review generation, targeting the gap identified by prior evaluations that LLM-generated reviews are often generic and not sufficiently tied to revision (Sadallah et al., 2025; Hosseini and Horbach, 2023; Chamoun et al., 2024).

Experiments show that our model produces review feedback that are more actionable and specific than competitive baselines under both human and LLM-as-a-judge evaluations. We evaluate on a test set built from ICLR 2025 and the improvements hold across multiple perspectives. Comparisons cover our RBTACT, an SFT-only variant, larger prompted LLMs such as Llama-3.1-70B (Meta AI, 2024) and GPT-5-chat (OpenAI, 2025), and other methods. RBTACT achieves the highest actionability in both studies: 3.46 out of 5 in human evaluation and 3.38 out of 5 in LLM-as-a-judge evaluation, maintaining parity on groundedness and relevance, with fine-grained analyses across seven perspectives and pairwise win rates indicating consistent gains.

We summarize our contributions as follows:

- We present the framework RBTACT that first utilizes author rebuttals as an implicit supervision and applies preference optimization to directly optimize a feedback generator for actionability.
- We release a large-scale dataset named RMR-75K, which contains 75,542 examples. Each example consists of (i) a review segment, (ii) an associated perspective on the review, (iii) an author response to the review, and (iv) an annotated impact category indicating the review's actionability.
- We propose an effective training pipeline that yields consistent gains in actionability and other aspects over competitive baselines under human and LLM-as-a-judge evaluations.

2 Related Work

Peer Review Datasets. Early corpora such as PeerRead and NLPeer collect manuscripts, decisions, and textual reviews to support large-scale analysis of peer review (Kang et al., 2018; Dycke et al., 2023). More recent large-scale resources support fine-tuning reviewer models, including ReviewMT, Review-5K and DeepReview-13K (Tan et al., 2024; Zhu et al., 2025; Weng et al., 2025). Some other resources incorporate rebuttal to study review: PRRCA links reviews with rebuttal counter-arguments for meta-review generation (Wu et al., 2022); MOPRD aggregates multi-disciplinary open reviews with rebuttal letters and editorial signals (Lin et al., 2023); Re^2 scales multi-turn review and rebuttal discussions across many venues (Zhang et al., 2025). To analyze review–rebuttal links, DISAPERE adds sentence-level discourse annotations over 506 review and rebuttal pairs (Kennard et al., 2022); JitsuPeer further

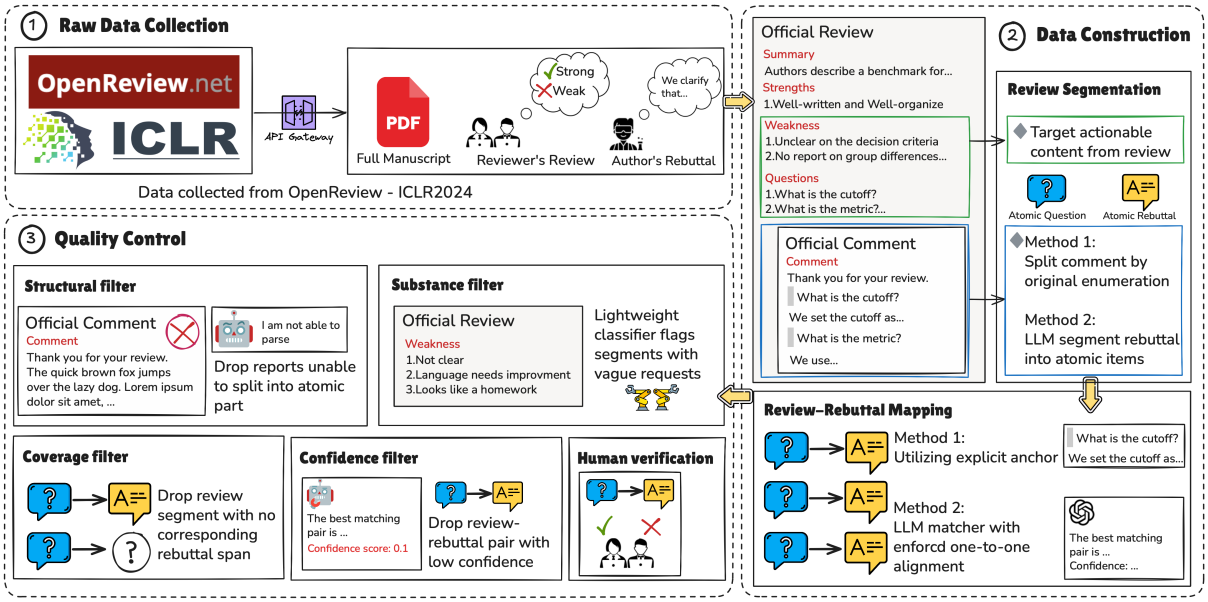

Figure 1: Overview of our construction pipeline for RMR-75K.

aligns review–rebuttal sentences with rebuttal action types for attitude-grounded rebuttal generation (Purkayastha et al., 2023). However, these links are defined at the sentence level and remain small-scale. In contrast, we construct large-scale point-to-point mappings that align each review point to its corresponding rebuttal segment and add perspective and impact labels, using them to supervise review feedback generation.

Review Generation. Early efforts prompted LLMs to write full reviews, revealing limited but non-trivial usefulness and notable failure modes (Liu and Shah, 2023; Robertson, 2023; Yuan et al., 2021). Later systems improved specificity through fine-tuning and structured pipelines, and most recently multi-agent approaches that assign roles to reduce generic feedback and increase comment quality (Zhu et al., 2025; Jin et al., 2024). Another related thread focuses on *actionable* reviewing, concerning whether feedback induces concrete changes or commitments. Most prior approaches operationalize actionability using reviewer-centric signals. ReAct labels reviews for actionability and intent to support detection and triage (Choudhary et al., 2021); ARIES links review comments to implemented edits, enabling analysis of which feedback leads to revisions (D'Arcy et al., 2024b); MARG explicitly frames the task as generating actionable peer review with specialized agents and reports reduced generic comments (D'Arcy et al., 2024a). In contrast, our method leverages rebuttal-anchored signals to ground actionability in authors' responses and actions, facilitating actionable review feedback generation.

3 Review and Rebuttal Mapping

To train models that generate actionable review comments using rebuttal as supervision, we need fine-grained mappings that connect each review point to the specific rebuttal that addresses it. Such segment-level alignment lets us attach a perspective label and an impact signal indicating whether the comment led to a concrete revision, a specific plan, or a defense without changes. Existing resources are not sufficient: DISAPERE (Kennard et al., 2022) provides sentence-level discourse annotations but lacks our segment-level labels and is much smaller (506 pairs); ARIES (D'Arcy et al., 2024b) is related but targets authors' edits; other datasets either omit rebuttal or do not treat it as a training signal. Therefore we introduce our dataset RMR-75K (Review-Map-Rebuttal) which maps review and rebuttal at the segment level, and is much bigger (about 150× more than DISAPERE).

3.1 Review–Rebuttal Source Data Collection

We curate a corpus with papers, reviews and rebuttals from ICLR 2024 on OpenReview by querying the official API for each submission's manuscript, full reviews, and the corresponding author responses. For downstream text processing, we convert the PDF manuscripts to Markdown with MinerU (Wang et al., 2024).

For each paper we keep (i) the full manuscript text; (ii) every reviewer's complete review and free-form comments; (iii) the author rebuttal thread.

When a rebuttal appears in multiple messages, we concatenate them in chronological order to obtain a single response document per paper.

3.2 Dataset Construction

We construct a point-to-point mapping between reviewer key points and the specific portion of the rebuttal that addresses each point, named RMR-75K, shown in Figure 1. Dataset statistics are summarized in Table 1.

Statistic	Value
Total mappings	75,542
Total papers	4,825
Avg. reviewers per paper	3.44
Avg. mappings per paper	15.66
Distinct reviews	16,583
Avg. mappings per review	4.56
Avg. confidence score	0.9268
Conference	ICLR 2024

Table 1: Summary statistics for the Review-Rebuttal-Mapping-75k dataset.

Notation. Let paper p have reviewer set $\mathcal{J}_p$. A reviewer $j \in \mathcal{J}_p$ writes a review R_j^p, which we decompose into a sequence of weakness/question segments $R_{j,1:K_j}^p = \{r_{j,1}^p, \ldots, r_{j,K_j}^p\}$. The concatenated rebuttal text for p is B^p, which we further split into candidate response spans $B_{1:T}^p = \{b_1^p, \ldots, b_T^p\}$. The goal is to construct a mapping

$$\mathcal{A}(p) \;=\; \{\,(r_{j,k}^p,\, b_t^p,\, \hat{c}_{j,k,t}^p)\,\}_{(j,k,t)\in\mathcal{I}_p},$$

where $\hat{c}_{j,k,t}^p \in [0,1]$ is a confidence score and we enforce a one-to-one mapping: each $r_{j,k}^p$ and each b_t^p appears in at most one pair.

Review Segmentation. We target only actionable content by extracting *Weaknesses* and *Questions* because author rebuttals mainly respond to these two parts. When reviewers already enumerate items (e.g., "1/2/3", "W1/W2/…" or bullet lists), we take those spans directly. Otherwise, we prompt GPT-5 (OpenAI, 2025) to split R_j^p into atomic critique units $\{r_{j,k}^p\}$ that each express a single concern. The prompt is shown in Figure 4.

Review–Rebuttal Mapping. We perform two-stage alignment. First, a heuristic pass links items using explicit anchors (e.g., "W1", quoted phrases, or reviewer's specific references), yielding high precision pairs. Second, an LLM matcher conducts span-level semantic linking: given a segment $r_{j,k}^p$ and the candidate spans $\{b_t^p\}$, it selects the best-supported b_t^p and returns a calibrated confidence $\hat{c}_{j,k,t}^p$. We enforce one-to-one alignment by greedy matching in descending $\hat{c}$ and discard ties. This realizes a paper-specific map

$$\text{Map} : \big(R^p,\, B^p\big) \to \big\{(r_{j,k}^p,\, b_t^p)\big\}_{(j,k,t)}.$$

The prompt is shown in Figure 5.

3.3 Cleaning and Quality Control

Because our training uses rebuttal-derived impact categories as preference signals for actionability, supervision quality hinges on precise review–rebuttal alignment. We therefore keep only confidently aligned pairs via the filters below and verify quality with targeted human checks.

Structural Filter. We drop reviews with no visible itemization and for which the segmenter cannot produce stable atomic units.

Coverage Filter. During alignment, if a review segment $r_{j,k}^p$ receives no plausible rebuttal span (no $\hat{c} \geq \tau$), we remove it.

Confidence Filter. We set a threshold τ and keep only pairs with $\hat{c} \geq \tau$; we also prune pairs where the matched rebuttal span merely restates the question without addressing it.

Substance Filter. We filter out review segments that do not semantically raise any substantive issue or recommendation.

Human Verification. We sample 60 papers to cover different perspectives. From these papers, we include all review segments that passed our filters and met the confidence threshold, yielding 944 mapped segments. Two trained annotators independently map each review segment to the most specific rebuttal span (or mark "No Response"), following the same one-to-one constraint and span granularity as our pipeline. A mapping is counted as *correct* if the predicted rebuttal span overlaps the gold span with token-level IoU ≥ 0.5. We compute span-level Precision/Recall/F1 for our automatic mapping against the adjudicated gold, and report Cohen's κ between the two annotators before adjudication. Results are in Table 2. Our mapping achieves high span-level accuracy (F1 = 0.91) with substantial IAA ($\kappa = 0.80$), indicating that the pipeline provides reliable supervision for downstream training.

Setting	Precision	Recall	F1	Cohen's κ
Filtered set	0.93	0.90	**0.91**	0.80

Table 2: Gold-standard validation of review→rebuttal span mappings. κ is inter-annotator agreement (IAA).

4 Methodology

4.1 Task Definition

Given a paper p, a perspective s and the whole paper text as context, the model generates a single focused review segment y that raises weaknesses or questions of paper p from perspective s.

Review Perspective Labels. Each mapped review segment receives one label from a 7-category taxonomy: *Experiments*, *Writing*, *Presentation*, *Theory*, *Novelty*, *Reproducibility*, and *Evaluation*. We assign labels automatically with a rubric-based prompt to GPT-5 that asks for the single best label and a short rationale, where prompt is shown in Appendix B. To verify quality, two trained annotators conducted stratified spot checks over papers and confidence bins. They independently reviewed a held-out sample and then resolved disagreements. The automatic labels matched human judgment on most cases (accuracy about 92%; Cohen's $\kappa = 0.81$). The remaining confusions were mainly *Writing* vs. *Presentation* and *Experiments* vs. *Evaluation*. Moreover, the detailed definition of perspective labels is shown in Table 3.

Rebuttal Impact Category as Actionability Signal. For each mapped rebuttal segment, we annotate an impact category that reflects the author's concrete action in response to the review by GPT-5: (1) *Concrete Revision Performed* (*CRP*), (2) *Specific revision plan* (*SRP*), (3) *Vague Commitment to Revise* (*VCR*), (4) *Defend Without Change* (*DWC*), and (5) *Deflect/Reframe* (*DRF*). The definition is shown in Table 3. The automatic labels matched the adjudicated human labels on most cases (accuracy 89%), with inter-annotator agreement $\kappa = 0.79$. They correspond to varying degrees of modifications made by the authors, reflecting the review's level of actionability. The categories leverage author reactions in rebuttal segments, reflecting how much revision or defense occurred, to capture the actionability of reviewer comments and their impact on rebuttals. After labeling, the distribution of each impact category and perspective in RMR-75K is shown in Figure 2. We detail the specific label process and the distribution of each category in Appendix B.

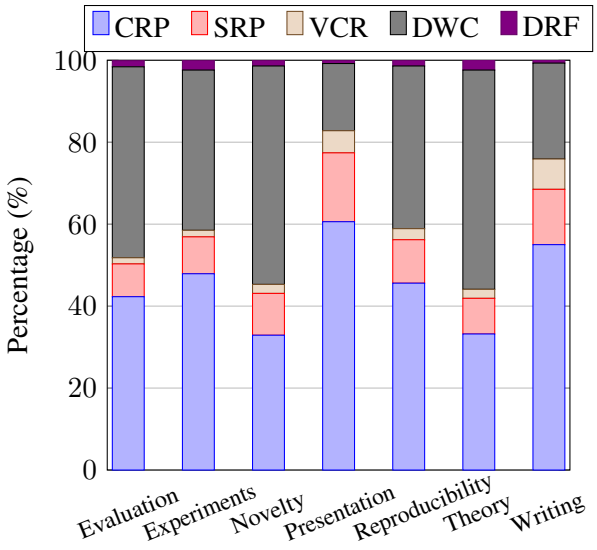

Figure 2: Normalized (100%) impact category composition by perspective.

4.2 Training Dataset Construction

SFT Data. We construct REVIEWSEG-SFT-13K, a supervised corpus collecting pairs $\mathcal{D}_{\text{SFT}} = \{(x_i, y_i)\}_{i=1}^{N}$. Each input is $x_i = (p, s)$, where p is the paper content and s is the target perspective (e.g., *Experiments*, *Writing*); the target $y_i^\star$ is the gold review sentence selected from the mapped reviewer segments $R^p_{j,1:K_j}$. Dataset size and coverage are reported in Table 4. There are 13,300 pairs spanning 4,637 papers, balanced to include 1,900 instances per perspective.

Preference Data. From the review–rebuttal alignments $\mathcal{A}(p)$ in §3.2, we build REVIEWPREF-DPO-22K as preference triples (x, y_w, y_ℓ), where $x = (p, s)$ matches the SFT input and y_w, y_ℓ are two review segments drawn from the same paper p and perspective s. Winners follow the rebuttal impact order *CRP* > *SRP* > *VCR* > *DWC* > *DRF*, which reflects increasing author uptake and revision degree in rebuttals. We compare only strictly ordered labels and stratify by impact gap to modulate difficulty: large (*CRP* vs. *DWC*, *DRF*), medium (*SRP* vs. *DWC*, *DRF*, *CRP* vs. *VCR*), and small (*CRP* vs. *SRP*, *VCR* vs. *DWC*, *DRF*, *DWC* vs. *DRF*). To keep the signal robust, we balance pair counts across papers and perspectives, never mix different papers or perspectives within a pair, and cap how often any single segment can appear so no segment dominates the pool. As summarized

Label set	Definition
	Perspective of Review Segment
Experiments	Concerns about experimental **setup and design**: missing or insufficient experiments, weak or absent ablations, unfair/weak baselines, unclear dataset descriptions or splits, problematic hyperparameter/seed choices, compute budgets, or training details that undermine empirical support.
Evaluation	How results are **measured, analyzed, and interpreted**: inappropriate/missing metrics, lack of statistical testing or error bars, insufficient analysis (e.g., no variance/sensitivity), unfair comparisons, or inconsistencies between reported numbers and claims.
Reproducibility	Ability to **reproduce** results: missing implementation details or pseudo-code, absent code/data/links, unspecified hyperparameters, unclear preprocessing, seeds, or hardware; insufficient instructions to replicate tables/figures.
Novelty	**Originality and relation to prior work**: overlap with existing methods, incremental contributions, unclear differentiation, missing or superficial positioning against closely related literature.
Theory	**Theoretical correctness and justification**: flawed or unstated assumptions, gaps in proofs, incorrect derivations, mismatches between theorems and algorithms, or theory not supporting the claimed guarantees.
Writing	**Clarity and readability**: grammar/style issues, ambiguous phrasing, undefined symbols/terms, confusing explanations, or organization that impedes understanding at the sentence/paragraph level.
Presentation	**Figures/tables/organization**: unclear plots or legends, poor formatting, misplaced or redundant content, and overall paper structure that makes the narrative hard to follow.
	Impact Level of Rebuttal Segment
CRP	**Concrete Revision Performed**: authors point to specific changes or verifiable artifacts already added (new text/sections, new experiments/tables/figures, updated numbers, released code/data). *Cues:* "We added/updated . . . ", "Section X rewritten . . . ", "New ablation in Sec. . . . shows . . . ", "Code/data at . . . ".
SRP	**Specific Revision Plan**: authors commit to concrete future edits with where/what to change, but not yet implemented. *Cues:* "We will add an ablation in Sec. X . . . ", "We will redraw Fig. . . . ", "We will clarify definitions in §. . . ".
VCR	**Vague Commitment to Revise**: promises to improve without actionable details (no locations, artifacts, or timelines). *Cues:* "We will revise accordingly.", "We will improve clarity/writing."
DWC	**Defend Without Change**: argues the paper already addresses the point; no edits proposed. *Cues:* "Already covered in Sec. . . . ", "Setup is standard.", "Claim stands."
DRF	**Deflect/Reframe**: shifts responsibility or reframes the issue; no change offered. *Cues:* "Reviewer misinterprets . . . ", "Out of scope . . . ", "Reviewer phrasing is incorrect."

Table 3: Summary of *perspective* labels for review segments and *impact* labels for rebuttal segments in §4.1.

Statistic	REVIEWSEG-SFT-13K	REVIEWPREF-DPO-22K
# pairs	13,300	21,822
# papers	4,637	4,825
Per-perspective	1,900 each	avg. 3,117 each
Avg. paper length	22,152	21,798
Avg. output length	62	Ch: 65, Re: 63
Conference	ICLR 2024	ICLR 2024
Extra test pairs	1330	2180

Table 4: Statistics for SFT dataset and preference dataset for DPO. The unit of length is tokens. "Ch" denotes Chosen and "Re" denotes Rejected.

in Table 4, the preference set contains 21,822 pairs from 4,825 papers with 3,117 pairs per perspective in average.

We quantify the difficulty stratification in Table 5. This distribution exposes the model to a spectrum of preference margins while keeping pairs within the same paper and perspective.

Tier	Winner	Loser	#Pairs	Percentages
Large impact gap (Easy)				
	CRP	DWC	9,887	45.3
	CRP	DRF	381	1.7
Subtotal			10,268	47.1
Medium impact gap				
	SRP	DWC	4,969	22.8
	SRP	DRF	248	1.1
	CRP	VCR	1,904	8.7
Subtotal			7,121	32.6
Small impact gap (Hard)				
	CRP	SRP	2,484	11.4
	VCR	DWC	1,423	6.5
	VCR	DRF	70	0.3
	DWC	DRF	456	2.1
Subtotal			6,433	20.3
Total			21,822	100.0

Table 5: Difficulty-stratified breakdown of preference pairs in REVIEWPREF-DPO-22K. Pairs are constructed only within the same paper and perspective and use strictly ordered rebuttal-impact labels.

4.3 Policy Optimization

Direct Preference Optimization. Direct Preference Optimization (DPO) learns a policy from pairwise preferences without fitting a separate reward model (Rafailov et al., 2024). Let π_θ be the trainable policy and π_{ref} a fixed reference policy (the SFT model). Under a Bradley–Terry prefer-

ence model, DPO maximizes the probability that the policy prefers y_w over y_ℓ by matching the optimal policy's log-density ratio relative to π_{ref}. The loss for a batch of preference triples is as follows.

Let $\Delta_{\theta,\text{ref}}(x, y) = \log \pi_\theta(y \mid x) - \log \pi_{\text{ref}}(y \mid x)$.

We sample preference triples (x, y_w, y_ℓ) from the dataset $\mathcal{D}_{\text{pref}}$. The DPO loss is

$$\mathcal{L}_{\text{DPO}}(\theta) = -\,\mathbb{E}_{(x,y_w,y_\ell)}\Big[\log \sigma\big(\beta\,[\,\Delta_{\theta,\text{ref}}(x, y_w) - \Delta_{\theta,\text{ref}}(x, y_\ell)\,]\big)\Big] \tag{1}$$

where $\beta > 0$ controls the sharpness of the preference. Intuitively, the policy is encouraged to increase likelihood on comments that led to higher-impact author actions (CRP, SRP) while decreasing it on comments that were defended or deflected (DWC, DRF).

Stabilization. We keep π_{ref} frozen at the SFT checkpoint and optionally mix in a small fraction of $\mathcal{L}_{\text{SFT}}$ on positive samples to prevent drift on perspective control when the context is long. The full objective is

$$\mathcal{L}(\theta) \;=\; \mathcal{L}_{\text{DPO}}(\theta) \;+\; \lambda\,\mathbb{E}_{(x,y_w)}[-\log \pi_\theta(y_w \mid x)]\,,$$

with a small λ. In practice, we set λ as 0.1.

4.4 Training Details

All experiments run on NVIDIA H200 141GB GPUs with bf16 compute. We use the base model Llama-3.1-8B-Instruct (Meta AI, 2024).

We train RBTACT-SFT model on dataset REVIEWSEG-SFT-13K. We run 3 epochs (4,989 steps) with learning rate 1.0×10^{-4} and a cosine scheduler. Total training time on H200 is approximately 120 hours.

To obtain RBTACT, we then utilize dataset REVIEWPREF-DPO-22K for further DPO training. The DPO policy π_θ is initialized from the SFT checkpoint and trained for 2 epochs (2,728 steps) using the Bradley–Terry DPO loss, keeping the reference π_{ref} frozen. We use learning rate 1.0×10^{-5} and a cosine scheduler. Total DPO training time on H200 is approximately 203 hours.

The additional training details including training configs and other optimizations are shown in Appendix C.

5 Experiments

5.1 Baselines

We compare RBTACT against three baseline types under identical inputs, prompts and decoding.

Fine-tuned (SFT-only). As a trained baseline, we fine-tune Llama-3.1-8B-Instruct on ReviewSeg-SFT-13k, getting RBTACT-SFT, using the same prompt template and output format.

Prompted LLMs. We query API models and run competitive open-source models in zero-shot mode to produce a single review segment for the requested perspective: GPT-5-chat (OpenAI, 2025), DeepSeek-V3.2 (DeepSeek-AI, 2025), Llama-3.1-70B (Meta AI, 2024) and Qwen-3-32B (Team, 2025). The prompt is shown in Appendix D.1.

Other Methods. We additionally compare against three task-adapted methods that are widely used around review feedback generation: (i) MARG (D'Arcy et al., 2024a), a multi-agent prompting framework for scientific review generation; (ii) LimGen (Faizullah et al., 2024), a limitations-focused generation setup that targets suggestive weaknesses; and (iii) DeepReviewer-14B (Zhu et al., 2025), a review generation model to produce comprehensive reviews. For each method, we adapt its protocol to our setting (paper + requested perspective $\rightarrow$ a single review feedback) and normalize the final output to our unified segment format. The details are shown in Appendix D.2. We additionally include a simple retrieval baseline in Appendix F.2 to test whether gains can be explained by nearest-neighbor reuse from the training corpus.

5.2 Evaluation Protocol

Evaluation Dataset. We construct a test set from a subset of ICLR 2025 papers using the same pipeline as in §3.2 to reduce data contamination, ensuring that none are resubmissions of ICLR 2024 papers. We sample 700 papers, stratified by perspective with 100 papers per perspective across the seven perspectives. Each paper is paired with one perspective labeled by annotators and one human review segment, which serves as the gold reference.

Human Evaluation Setup. We evaluate a subset of the evaluation dataset: 50 papers sampled across perspectives. For each paper, annotators view the title, relevant content, and the target perspective. Three PhD-level or senior graduate annotators (each with $\geq$2 completed reviews at major ML venues) rate nine anonymized model outputs per paper on a 1–5 scale for *Actionability*, *Specificity*, *Groundedness*, *Relevance*, and *Helpfulness*, following a written rubric with positive and nega-

System	Human					LLM-as-a-Judge				
	Action.	Spec.	Ground.	Rel.	Help.	Action.	Spec.	Ground.	Rel.	Help.
Ours										
RBTACT (ours)	**3.46**	**4.08**	4.30	4.77	4.25	**3.38**	**3.71**	4.05	4.82	3.74
RBTACT-SFT	3.28	4.01	4.15	4.70	4.24	3.19	3.59	3.93	4.73	3.66
LLMs										
GPT-5-chat	3.38	4.03	**4.35**	**4.97**	**4.49**	3.27	3.66	**4.12**	**4.95**	**3.78**
DeepSeek-V3.2	3.15	3.97	4.23	4.88	4.28	3.14	3.57	4.00	4.85	3.70
Llama-3.1-70B	3.22	3.95	4.18	4.65	4.15	3.11	3.54	3.93	4.71	3.53
Qwen-3-32B	3.06	3.78	4.12	4.58	4.12	3.03	3.36	3.87	4.68	3.32
Other Methods										
MARG	3.20	3.87	4.15	4.72	4.18	3.19	3.46	3.91	4.69	3.51
DeepReviewer-14B	3.27	3.96	4.28	4.75	4.21	3.23	3.51	4.03	4.74	3.58
LimGen	3.16	3.92	4.08	4.62	4.05	3.08	3.37	3.89	4.54	3.39

Table 6: Pointwise ratings on five quality dimensions. Left: Human. Right: LLM-as-a-judge. Higher is better.

tive indicators and anchor examples. Outputs are shown as three order-randomized pairs to mitigate position bias. Scores are averaged over annotators. Full instructions are reported in Appendix D.3. Because the human ratings use a 1–5 Likert scale, we also report an ordinal-aware analysis in Appendix F.3.

LLM-as-a-Judge Evaluation Setup. We evaluate on 105 papers from the evaluation dataset. Each model produces one segment per paper, yielding 105 segments per model and 1350 total across nine models. The judge model (GPT-5-chat) sees the paper context, the target perspective, and one or two candidate segments (Zheng et al., 2023). For pointwise scoring, it assigns 1–5 on the same five dimensions as the human study. For pairwise comparison, it scores both candidates and then chooses which is more *Actionable* with a brief rationale. The exact prompts are in Figure 15 and Figure 16.

5.3 Experiment Results

Human Evaluation Results. As shown in Table 6, RBTACT attains the highest *Actionability* while remaining competitive on other dimensions.

LLM-as-a-judge Pointwise Results. The judge reproduces the human trend: RBTACT scores highest on *Actionability* and *Specificity*, while remaining close to strong LLM baselines on other dimensions (Table 6). The prompt is shown in Figure 15.

LLM-as-a-judge Pairwise Results. We compare *Actionability* via pairwise judgments on the same paper and perspective. Table 7 reports the percentage of wins for each model. To visualize patterns across perspectives, Figure 17 shows heatmaps where each cell is the win rate of the *row* model over the *column* model. Overall, RBTACT attains the highest average pairwise win rate and leads in most perspectives, followed by GPT-5-chat. The pairwise prompt is in Figure 16.

Human–LLM Agreement. We quantify alignment between human judgments and the LLM judge along three axes. (i) *Model-level rank correlation*: ranking models by pointwise means yields high agreement for *Actionability* (Spearman's ρ=0.94, Kendall's τ_b=0.87), with only a minor swap between mid-ranked baselines. (ii) *Item-level correlation*: across matched paper–model cells (20 papers $\times$ 9 models), human vs. LLM pointwise scores show moderate positive correlation on *Actionability* (Spearman's ρ=0.52).

5.4 Automatic Evaluation

Besides human and LLM-as-a-Judge evaluation, we also evaluate our method and different baselines by some metrics. Results are shown in Table 8.

As mentioned, we evaluate all 700 test instances from evaluation dataset with four metrics: BLEU@4 (Papineni et al., 2002), ROUGE-L_{sum} (F1) (Lin, 2004), METEOR (Banerjee and Lavie, 2005), and chrF (Popović, 2015).

Overall, our models perform competitively against baselines.

5.5 Case Study

Here we show some case studies to demonstrate why our model RBTACT is more actionable than other baselines in most cases from different perspectives. They are shown in Appendix E.

	Actionability (Pairwise Win Rate %)								
Winners	RBTACT	GPT-5	DeepSeek	Llama-3.1	DeepReviewer	MARG	Qwen-3	RBTACT-SFT	LimGen
RBTACT	–	57.1	63.8	61.9	65.7	68.6	66.7	65.7	76.2
GPT-5-chat	42.9	–	55.2	57.1	60.0	62.9	60.0	54.3	71.4
DeepSeek-V3.2	36.2	44.8	–	54.3	57.1	59.0	58.1	53.3	68.6
Llama-3.1-70B	38.1	42.9	45.7	–	52.4	55.2	56.2	51.4	66.7
DeepReviewer-14b	34.3	40.0	42.9	47.6	–	51.4	53.3	50.5	64.8
MARG	31.4	37.1	41.0	44.8	48.6	–	50.5	49.5	61.9
Qwen-3-32B	33.3	40.0	41.9	43.8	46.7	49.5	–	52.4	58.1
RBTACT-SFT	34.3	45.7	46.7	48.6	47.6	50.5	47.6	–	56.2
LimGen	23.8	28.6	31.4	33.3	35.2	38.1	41.9	43.8	–

Table 7: LLM-as-a-judge pairwise win rates (%) on Actionability

Model	BLEU@4	ROUGE-L_{sum}	METEOR	chrF
RBTACT (Ours)	14.62	**12.64**	**11.65**	18.57
RBTACT-SFT	**14.93**	12.33	11.53	18.51
GPT-5-chat	11.17	10.11	9.96	**24.90**
DeepSeek-V3.2	10.19	9.76	8.49	17.78
Llama-3.1-70B	10.48	8.76	8.27	16.42
Qwen-3-32B	9.72	8.58	8.14	17.18
DeepReview	12.40	11.35	10.21	19.69
MARG	10.95	9.12	8.64	18.16
LimGen	10.92	8.27	7.97	17.43

Table 8: Automatic evaluation on 700 test instances. Values are reported as percentages.

5.6 Summary

Both human and LLM-as-a-judge evaluations show that our RBTACT improves the practical value of review segments. Automatic evaluation also shows the strong capability of RBTACT. Gains concentrate on actionability and specificity with parity on groundedness, relevance, and Helpfulness. Importantly, our 8B-scale RBTACT remains competitive with much larger 32–70B and proprietary models on actionability (and often specificity), suggesting the practical value of rebuttal-supervised training. Additional analyses on issue severity and paper strength are provided in Appendix F, showing that our gains are not driven only by minor issues and are also consistent across papers of different strengths.

6 Conclusion

We study actionable review feedback generation by placing rebuttal at the center of learning. Our framework, RBTACT, uses rebuttals as implicit supervision and frames the task as segment-level generation from a given perspective with explicit mappings from each review segment to its addressing rebuttal span. We release RMR-75K with perspective labels and impact categories that utilize actionability. An effective pipeline of supervised fine-tuning followed by preference optimization on impact ordered pairs yields consistent gains in actionability and specificity while maintaining grounding and relevance against strong baselines under both human evaluation and LLM-as-a-judge protocols. These results show rebuttal signals are a practical form of human feedback for producing targeted, implementable guidance. We release data and code to spur research on rebuttal driven learning and better evaluation of actionability.

Limitations

Our approach relies on rebuttal as an implicit supervision signal for actionability, which is informative but imperfect. Rebuttals reflect short horizon author uptake during review, not long-term implementation, and can include strategic promises or deferrals. The dataset focuses on venues with public rebuttals, mainly computer science communities that use OpenReview, so generalization to journals, non-English venues, and other fields remains uncertain. Our model can generate precise yet infeasible suggestions and our current setup does not include rigorous verification against the manuscript, code, or data artifacts.

Acknowledgements

This work was supported in part by Google's Research Scholar Program and a research grant from TCS Research.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

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

Gautam Choudhary, Natwar Modani, and Nitish Maurya. 2021. *ReAct: A Review Comment Dataset for Actionability (and more)*, page 336–343. Springer International Publishing.

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

Mike D'Arcy, Alexis Ross, Erin Bransom, Bailey Kuehl, Jonathan Bragg, Tom Hope, and Doug Downey. 2024b. Aries: A corpus of scientific paper edits made in response to peer reviews.

DeepSeek-AI. 2025. Introducing deepseek-v3.2-exp. https://api-docs.deepseek.com/news/news250929. Accessed: 2025-10-07.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. 2023. NLPeer: A unified resource for the computational study of peer review. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5049–5073, Toronto, Canada. Association for Computational Linguistics.

Abdur Rahman Bin Md Faizullah, Ashok Urlana, and Rahul Mishra. 2024. Limgen: Probing the llms for generating suggestive limitations of research papers.

Zhaolin Gao, Kianté Brantley, and Thorsten Joachims. 2024. Reviewer2: Optimizing review generation through prompt generation.

Tirthankar Ghosal, Shubham Kumar, Prasenjit Kumar Bharti, and Asif Ekbal. 2022. Peer review analyze: A novel benchmark resource for computational analysis of peer reviews. *PLOS ONE*, 17(1):e0259238.

M. Hosseini and S. P. J. M. Horbach. 2023. Fighting reviewer fatigue or amplifying bias? considerations and recommendations for use of chatgpt and other large language models in scholarly peer review. *Research Integrity and Peer Review*, 8(4):1–12.

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.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (PeerRead): Collection, insights and NLP applications. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1647–1661, New Orleans, Louisiana. Association for Computational Linguistics.

Neha Nayak Kennard, Tim O'Gorman, Rajarshi Das, Akshay Sharma, Chhandak Bagchi, Matthew Clinton, Pranay Kumar Yelugam, Hamed Zamani, and Andrew McCallum. 2022. DISAPERE: A dataset for discourse structure in peer review discussions. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1234–1249, Seattle, United States. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Jialiang Lin, Jiaxin Song, Zhangping Zhou, Yidong Chen, and Xiaodong Shi. 2023. Moprd: A multidisciplinary open peer review dataset. *Neural Computing and Applications*, 35(34):24191–24206.

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

Meta AI. 2024. Introducing llama 3.1: Our most capable models to date. https://ai.meta.com/blog/meta-llama-3-1/. Accessed: 2025-10-07.

OpenAI. 2025. Gpt-5. https://openai.com/gpt-5/. Accessed: 2025-10-07.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Sukannya Purkayastha, Anne Lauscher, and Iryna Gurevych. 2023. Exploring jiu-jitsu argumentation for writing peer review rebuttals.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model.

Zachary Robertson. 2023. Gpt4 is slightly helpful for peer-review assistance: A pilot study.

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.

Hyungyu Shin, Jingyu Tang, Yoonjoo Lee, Nayoung Kim, Hyunseung Lim, Ji Yong Cho, Hwajung Hong, Moontae Lee, and Juho Kim. 2025. Mind the blind spots: A focus-level evaluation framework for llm reviews.

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.

Qwen Team. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, Bo Zhang, Liqun Wei, Zhihao Sui, Wei Li, Botian Shi, Yu Qiao, Dahua Lin, and Conghui He. 2024. Mineru: An open-source solution for precise document content extraction.

Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. 2025. Cycleresearcher: Improving automated research via automated review.

Po-Cheng Wu, An-Zi Yen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2022. Incorporating peer reviews and rebuttal counter-arguments for meta-review generation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 2189–2198, New York, NY, USA. Association for Computing Machinery.

Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2021. Can we automate scientific reviewing?

Daoze Zhang, Zhijian Bao, Sihang Du, Zhiyi Zhao, Kuangling Zhang, Dezheng Bao, and Yang Yang. 2025. Re^2: A consistency-ensured dataset for full-stage peer review and multi-turn rebuttal discussions.

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.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. 2025. Deepreview: Improving llm-based paper review with human-like deep thinking process.

A Review and Rebuttal Mapping Details

A.1 Example of Mapping

Here is an example in our Review-Rebuttal-Mapping-75k dataset in Figure 3. We sample two mappings in the list.

A.2 Prompts for Review Segment and Mapping

The prompt for review segmentation is shown in Figure 4. The prompt for mapping between review segment and rebuttal segment is shown in Figure 5.

B Label Details

We discuss details of our labeling of perspectives and impact categories here.

B.1 Summarized Definition of Perspective labels and Impact Categories

The summarized definition of perspective labels and rebuttal's impact categories are shown in Table 9.

B.2 Perspective and Impact Category Distribution

We analyze the seven-perspectives and impact category distribution in our Review-Rebuttal-Mapping-75k dataset, which is shown in Table 10.

B.3 Prompt for labeling perspectives

The prompt for labeling which perspective a review segment belongs to is shown in Figure 6.

B.4 Prompt for labeling impact categories

The prompt for labeling which impact category a rebuttal segment belongs to is shown in Figure 7.

C Additional Training Details

All training details are shown in Table 11.

Framework and hardware. We train using `LLaMA-Factory` on NVIDIA H200 GPUs (141 GB) with `bf16` compute. We apply LoRA adapters to attention and MLP projections (`{q,k,v,o,gate,up,down}`) on top of the Llama-3.1-8B-Instruct base model. All runs use per-device batch size 1 with gradient accumulation as specified below.

Memory and speed optimizations. Direct Preference Optimization (DPO) requires evaluating both the chosen and the rejected responses under the policy and a frozen reference model, which increases memory use compared with SFT. To enable a 32k token context in DPO, we combine: (i) FlashAttention-2 for efficient attention; (ii) DeepSpeed ZeRO-2 with gradient checkpointing; and (iii) a 4-bit quantized frozen reference model. This configuration reduces activation and parameter memory, allowing full paper context while maintaining throughput.

Schedules. SFT trains for 3 epochs ($= 4{,}989$ steps) with learning rate 1.0×10^{-4} and a cosine scheduler (warmup ratio 0.10). DPO initializes from the SFT checkpoint and trains for 2 epochs ($= 2{,}728$ steps) with learning rate 1.0×10^{-5} and a cosine scheduler (warmup ratio 0.05). Wall-clock time on H200 is ≈ 120 hours for SFT and ≈ 203 hours for DPO.

D Evaluation Details

D.1 Baseline Prompt

Figure 8 shows the prompt for different LLM baselines generating review segments.

D.2 Baseline Adaptation Details

All baselines of "Other Methods" receive the same paper text and the same requested perspective, and are required to output *exactly one* review segment following our unified format (Appendix D.4). When a baseline naturally produces longer outputs (e.g., multi-section reviews, multiple bullets, or agent traces), we apply a deterministic post-processing rule to extract one segment, described below.

MARG. MARG (D'Arcy et al., 2024a) generates review feedback via multiple LLM instances (agents) that each read a portion of the paper and then aggregate intermediate discussion into final comments. To adapt MARG to our single-segment, perspective-conditioned setting, we inject the requested perspective into the leader's instruction so that the final synthesized comment targets the requested perspective. We output only the leader's final synthesized comment and discard intermediate agent messages. If multiple bullets are produced in the final synthesis, we take the first bullet as the single segment.

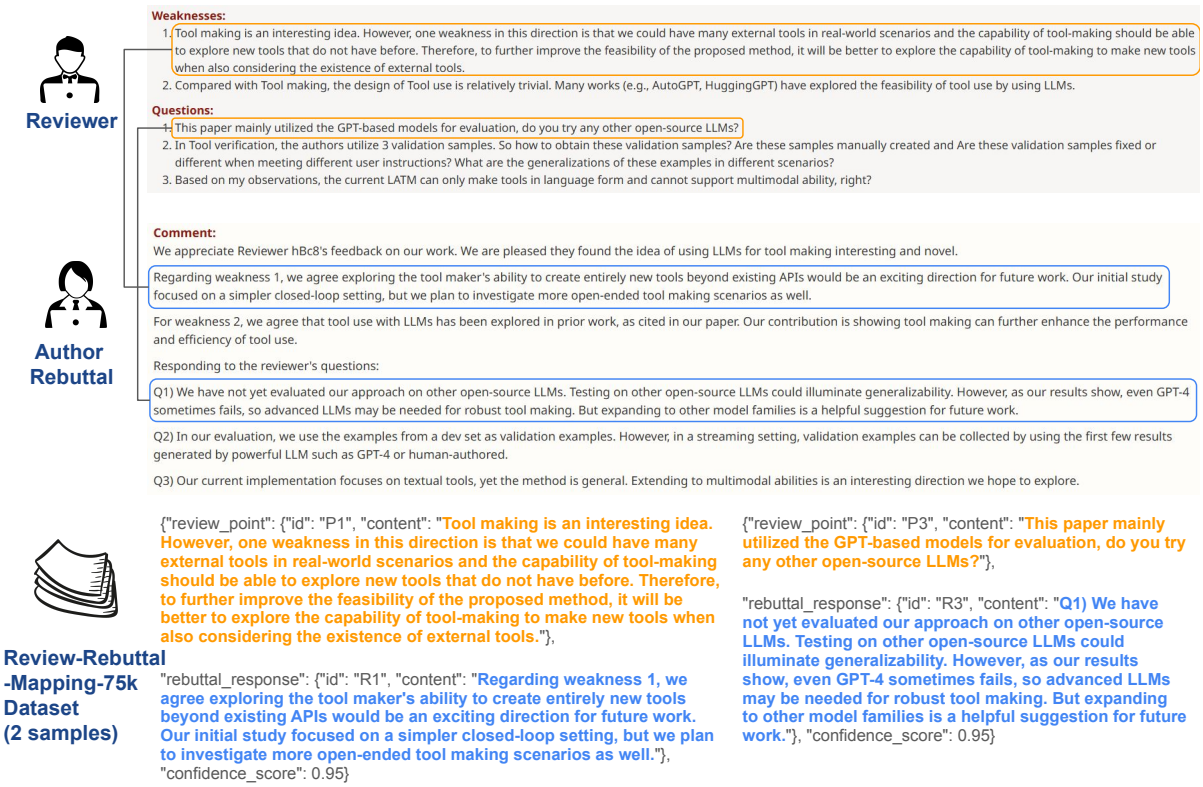

Figure 3: A mapping example of our Review-Rebuttal-Mapping-75k dataset. The review and rebuttal are from the paper titled "Large Language Models as Tool Makers".

LimGen. LimGen (Faizullah et al., 2024) focuses on generating limitations with suggestions for improvement. We adapt LimGen by prompting it to generate limitations *conditioned on the requested perspective*. If the output contains multiple limitation items, we deterministically select the first item (including its paired suggestion when present) as the single segment. We only normalize surface formatting (e.g., removing numbering or headings) and do not add new content.

DeepReviewer-14B. DeepReviewer-14B (Zhu et al., 2025) is trained to produce comprehensive, structured reviews. To make it comparable in our setting, we prompt DeepReviewer-14B to produce a *single* perspective-specific comment in our segment format. If the model outputs a full multi-section review despite the instruction, we extract the subsection that best matches the requested perspective using a fixed heading-based mapping (e.g., *Experiments/Empirical Evaluation* → experiments; *Writing/Presentation/Clarity* → clarity; *Impact/Significance/Novelty* → contribution/impact), and then take the first paragraph in that subsection as the segment. We run DeepReviewer-14B in pure generation mode without external tools to keep inputs comparable.

D.3 Human Expert Evaluation Protocol

This appendix describes the interface and rubric we used to evaluate review segments and actionable rebuttal guidance.

D.3.1 Interface

We collect two judgments per comparison on the page shown in Figure 9: (i) a *pairwise preference* between two anonymized candidate reviews (A vs. B vs. Tie), and (ii) *per-candidate 1–5 ratings* on five dimensions: **Actionability**, **Specificity**, **Groundedness**, **Relevance**, and **Helpfulness**. Anchors are 1 = Very poor, 2 = Poor, 3 = Fair, 4 = Good, 5 = Excellent.

Raters judge only using evidence in the paper. They should avoid hallucinations and must not penalize a section for missing information if the same information exists elsewhere in the paper.

D.3.2 Dimension Rubrics (Concise)

1. **Actionability**: Clear next steps with parameters or acceptance criteria as shown in Figure 10.

2. **Specificity**: Pinpoints exact sections, figures, metrics, or settings as shown in Figure 11.

3. **Groundedness**:Supported by the paper with explicit references or numbers as shown in

[System Prompt]
You are a professional academic review text analysis assistant. Your task is to segment a complete "Weaknesses & Questions" section from an academic paper review into independent, specific points.

You must follow these rules:
1. Each point should be an independent, specific issue or weakness
2. Preserve the core meaning of the original text without adding or removing information
3. Maintain existing numbering structures if present (e.g., 1., 2., W1, W2, etc.)
4. Handle various formatting styles including:
- Numbered lists (1., 2., 3.)
- Letter prefixes (W1, W2, Q1, Q2)
- Markdown bullet points (-, *, +)
- Section headers (## Weaknesses, ## Questions)
5. If no clear numbering exists, logically segment based on content structure
6. Each point should contain sufficient context to be understood independently
7. Preserve the original language and terminology used by the reviewer

[User Prompt]
Please segment the following Weaknesses & Questions text into independent points:

`{weaknesses&questions_text}`

IMPORTANT: Regardless of the input format (bullet points, numbered lists, paragraphs, etc.), you MUST output in this exact format:

Point 1: [Complete content of the first weakness point]
Point 2: [Complete content of the second weakness point]
Point 3: [Complete content of the third weakness point]
...

Rules:
- Use exactly "Point X:" where X is a number starting from 1
- Include ALL weakness points from the input, don't skip any
- Each point should be complete and independently understandable
- Preserve the original meaning and wording as much as possible
- If the input has bullet points (-, *, +) or numbered lists (1., 2.), convert them to Point format
- If the input has long paragraphs, break them into logical points

Figure 4: Prompt used to segment the weaknesses and questions parts of the review into segments.

Figure 12.

4. **Relevance**: Aligned with the target perspective and main contributions as shown in Figure 13.

5. **Helpfulness**: Clear, constructive guidance that improves the paper as shown in Figure 14.

D.4 LLM-as-a-Judge Evaluation

Here we discuss some details of LLM-as-a-judge evaluation.

The prompt for point-wise evaluation is Figure 15.

The prompt for pairwise evaluation on Actionability is Figure 16.

To visualize pairwise evaluation results across seven different perspectives, Figure 17 shows heatmaps where each cell is the win rate of the *row* model over the *column* model.

E Case Study

Here we show some case studies to demonstrate why our model RBTACT is more actionable than other baselines in most cases from different perspectives. They are shown in Figure 18.

F Additional Analyses

F.1 Severity and Paper-Strength Analyses.

We further analyze whether rebuttal-derived supervision over-emphasizes minor issues and whether gains vary with paper strength. We studies (i) issue *severity* using a perspective-based proxy and shows that our training signal substantially covers major issues that are often defended, and (ii) the relationship between paper strength (OpenReview ratings) and actionability improvements.

Label set	Definition
Perspective of Review Segment	
Experiments	Setup, baselines, ablations, datasets
Evaluation	Metrics, analysis, claims vs. results
Reproducibility	Missing code/details, reproducibility info
Novelty	Originality, relation to prior work
Theory	Assumptions, derivations, proofs
Writing	Clarity, grammar, readability
Presentation	Figures, tables, organization
Impact Category of Rebuttal Segment	
CRP	Concrete Revision Performed
SRP	Specific Revision Plan
VCR	Vague Commitment to Revise
DWC	Defend Without Change
DRF	Deflect or reframe, no change

Table 9: Brief summary of perspective labels for review segments and impact categories for rebuttal segments.

Perspective	CRP	SRP	VCR	DWC	DRF
Evaluation (11,257)	4,766 / 42.3%	903 / 8.0%	171 / 1.5%	5,249 / 46.6%	168 / 1.5%
Experiments (25,160)	12,059 / 47.9%	2,272 / 9.0%	401 / 1.6%	9,833 / 39.1%	595 / 2.4%
Novelty (8,585)	2,828 / 32.9%	872 / 10.2%	185 / 2.2%	4,578 / 53.3%	122 / 1.4%
Presentation (4,776)	2,894 / 60.6%	803 / 16.8%	256 / 5.4%	784 / 16.4%	39 / 0.8%
Reproducibility (4,402)	2,009 / 45.6%	465 / 10.6%	120 / 2.7%	1,747 / 39.7%	61 / 1.4%
Theory (12,822)	4,253 / 33.2%	1,110 / 8.7%	282 / 2.2%	6,859 / 53.5%	318 / 2.5%
Writing (8,540)	4,693 / 55.0%	1,149 / 13.5%	631 / 7.4%	1,997 / 23.4%	70 / 0.8%
Overall	33,502 / 44.3%	7,574 / 10.0%	2,046 / 2.7%	31,047 / 41.1%	1,373 / 1.8%

Table 10: Label distribution by perspective (counts / %).

F.1.1 Are we biased toward minor issues? A severity proxy.

We use a lightweight proxy for issue severity based on the review perspective labels. We treat **major** issues as {Experiments, Evaluation, Theory, Novelty, Reproducibility} and **minor** issues as {Writing, Presentation}. Table 12 aggregates the impact-label counts by this proxy.

As a result, major issues are indeed defended more often (DWC 45.4% vs. 20.9%), which matches the intuition that authors push back on higher-stakes critiques. However, rebuttal supervision is *not* dominated by minor issues: over half of major-issue mappings correspond to specific revisions (CRP+SRP = 50.7%). Moreover, our preference pairs are constructed *within the same paper and perspective* and balanced across perspectives (§4.2), which reduces the risk that training is driven primarily by easy-to-fix minor edits.

F.1.2 Do gains correlate with paper strength?

Setup. We operationalize *paper strength* using the mean reviewer rating on OpenReview (averaged across reviewers for the same submission). We split papers into three buckets (Weak / Medium / Strong) by tertiles of the mean rating. On the 105-paper LLM-as-a-judge set, we compute the average Actionability score per bucket.

Interpretation. The trend suggests that while stronger papers already elicit reasonably actionable feedback even from strong baselines, weaker papers benefit more from rebuttal-anchored supervision. This is aligned with our training signal: rebuttals often make explicit which critiques lead to concrete revisions versus defenses, which helps the model prioritize actionable, fixable issues.

F.2 Retrieval Baseline Analysis

To address the concern that the gains may come from a large organized corpus rather than learned generation, we add a simple retrieval baseline. Given a test paper and a target perspective, the baseline encodes the paper text as a query, searches only training review segments with the same perspective, excludes segments from the same paper to prevent leakage, and returns the nearest neighbor segment as the output. The baseline does not access rebuttals and uses the same single-segment output format as other systems.

Setting	SFT	DPO
Framework	LLaMA-Factory	LLaMA-Factory
Base model	Llama-3.1-8B-Instruct	Llama-3.1-8B-Instruct
Precision	bf16	bf16
GPU	H200 141 GB	H200 141 GB
Context limit (`cutoff_len`)	32,768	32,768
Generation limit (`max_new_tokens`)	512	512
Dataset	REVIEWSEG-SFT-13K	REVIEWPREF-DPO-22K
Objective	Token-level CE	DPO (sigmoid)
Epochs / Steps	3 / 4,989	2 / 2,728
LR / Scheduler	$1e{-4}$ / cosine	$1e{-5}$ / cosine
Warmup ratio	0.10	0.05
Batch (per-dev $\times$ accum)	1×8	1×16
Eff. batch	8	16
LoRA target	q,k,v,o,gate,up,down	q,k,v,o,gate,up,down
LoRA rank / α / dropout	8 / 16 / 0.05	16 / 16 / 0.05
Grad. checkpointing	Off	On
Reference model	N/A	Frozen SFT
Ref. quantization	N/A	4-bit
DeepSpeed (ZeRO-2)	Optional	On
FA2 (FlashAttention-2)	On	On
Wall-clock time	$\approx$ 120 h	$\approx$ 203 h

Table 11: Key training configuration for SFT and DPO. "Eff. batch" is per-device batch size $\times$ accumulation steps. DeepSpeed denotes ZeRO-2; FA2 denotes FlashAttention-2.

Impact label	Major (Ev/Ex/Th/No/Re)	Minor (Wr/Pr)
CRP	25,915 (41.6%)	7,587 (57.0%)
SRP	5,622 (9.0%)	1,952 (14.7%)
VCR	1,159 (1.9%)	887 (6.7%)
DWC	28,266 (45.4%)	2,781 (20.9%)
DRF	1,264 (2.0%)	109 (0.8%)

Table 12: Impact-label distribution by a perspective-based severity proxy. Major issues are more frequently defended (DWC), but still exhibit substantial author uptake: CRP+SRP is 50.7% for major issues and 71.6% for minor issues.

Bucket (by mean rating)	n	RBTACT	Baseline	Δ (ours – base)
Weak (bottom 1/3)	35	3.35	3.08	+0.27
Medium (middle 1/3)	35	3.33	3.11	+0.22
Strong (top 1/3)	35	3.46	3.27	+0.19

Table 13: Actionability by paper-strength bucket.

Table 14 shows the LLM-as-a-judge results. Retrieval is competitive on relevance and specificity, which suggests that perspective-matched review segments provide useful topical priors. However, it performs worse than RBTACT on all five dimensions, especially Actionability, Groundedness, and Helpfulness. This indicates that nearest-neighbor reuse can retrieve plausible comments, but it often fails to adapt them to the paper's actual method, evidence, and experimental setting. In contrast, RBTACT learns to generate paper-specific and more actionable feedback from rebuttal-derived supervision.

System	Action.	Spec.	Ground.	Rel.	Help.
RBTACT (ours)	3.38	3.71	4.05	4.82	3.74
Retrieval	3.02	3.45	3.80	4.48	3.39

Table 14: LLM-as-a-judge comparison with a retrieval baseline. The retrieval baseline returns the nearest training review segment under the same perspective while excluding segments from the same paper. Higher is better.

F.3 Ordinal-Aware Analysis of Human Ratings

Human pointwise ratings in Table 6 are collected on a 1–5 Likert scale. Since Likert ratings are ordinal, we provide an ordinal-aware analysis to complement the mean scores reported in the main paper. Specifically, we report the median and interquartile range, the percentage of ratings at least 4, and a ridit score computed from pooled category frequencies for Actionability and Helpfulness, as shown in Table 15. The ridit score measures the relative standing of a system under the pooled ordinal distribution, where values above 0.5 indicate ratings that are stochastically higher than the pooled average.

We further conduct paired nonparametric tests on per-item human ratings, as shown in Table 16. For each of these two dimensions, Actionability and Helpfulness, ratings are compared between

System	Actionability				Helpfulness			
	Mean	Med. [IQR]	% ≥ 4	Ridit	Mean	Med. [IQR]	% ≥ 4	Ridit
Ours								
RBTACT (ours)	**3.46**	**4 [2, 5]**	**52**	**0.548**	4.25	4 [2, 5]	82	0.509
RBTACT-SFT	3.28	3 [2, 4]	45	0.497	4.24	4 [1, 5]	81	0.505
LLMs								
GPT-5-chat	3.38	3 [2, 4]	47	0.523	**4.49**	**5 [4, 5]**	**91**	**0.574**
DeepSeek-V3.2	3.15	3 [2, 5]	39	0.461	4.28	4 [2, 5]	83	0.517
Llama-3.1-70B	3.22	3 [1, 4]	42	0.478	4.15	4 [1, 5]	77	0.482
Qwen-3-32B	3.06	3 [1, 4]	35	0.437	4.12	4 [1, 5]	75	0.472
Other Methods								
MARG	3.20	3 [2, 5]	41	0.474	4.18	4 [1, 5]	79	0.491
DeepReviewer-14B	3.27	3 [2, 4]	44	0.494	4.21	4 [1, 5]	77	0.503
LimGen	3.16	3 [1, 4]	40	0.465	4.05	4 [2, 5]	73	0.450

Table 15: Ordinal-aware summaries of human Likert ratings on Actionability and Helpfulness. Median, interquartile range, threshold rate, and ridit score are reported in addition to the descriptive mean. Higher is better.

RBTACT and each baseline on the same evaluated instances using the Wilcoxon signed-rank test. We report the signed-rank statistic, two-sided p-value, and rank-biserial correlation as an effect size. Positive effects favor RBTACT. This analysis avoids assuming equal intervals between Likert categories.

The ordinal-aware analysis is consistent with the main findings. RBTACT has the strongest actionability under the median-based summary and has positive paired effects against most baselines. Its improvement over RBTACT-SFT is supported by the Wilcoxon signed-rank test, while the helpfulness difference between the two systems is small and not significant. GPT-5-chat remains strongest on helpfulness, which is also reflected by its higher median, threshold rate, and ridit score. These results suggest that the main conclusions do not depend on treating Likert ratings as interval-scale measurements.

	Actionability				Helpfulness			
Comparison	W	p	**Effect**	**Result**	W	p	**Effect**	**Result**
RBTACT vs. RBTACT-SFT	1812.5	0.018	0.23	+	2148.0	0.641	0.04	n.s.
RBTACT vs. GPT-5-chat	1987.0	0.276	0.09	n.s.	1695.5	0.031	-0.19	−
RBTACT vs. DeepSeek-V3.2	1654.0	0.004	0.31	+	2186.5	0.784	-0.02	n.s.
RBTACT vs. Llama-3.1-70B	1875.5	0.031	0.20	+	1922.0	0.096	0.14	n.s.
RBTACT vs. Qwen-3-32B	1538.0	0.001	0.36	+	1830.5	0.041	0.18	+
RBTACT vs. MARG	1819.5	0.022	0.22	+	1994.0	0.184	0.11	n.s.
RBTACT vs. DeepReviewer-14B	1936.0	0.086	0.15	n.s.	2075.5	0.502	0.05	n.s.
RBTACT vs. LimGen	1711.0	0.007	0.28	+	1788.0	0.022	0.21	+

Table 16: Paired ordinal tests on human ratings. We use the Wilcoxon signed-rank test over matched evaluation instances and report rank-biserial correlation as the effect size. Positive effects favor RBTACT. + indicates that RBTACT is significantly better, − indicates that the baseline is significantly better, and n.s. indicates no significant difference.

[System Prompt]
You are a professional academic review analysis assistant. Your task is to perform precise one-to-one mapping between review weakness points and author rebuttal responses.

Guidelines for mapping:
1. Carefully analyze the rebuttal text to identify which sections respond to specific weaknesses
2. Look for explicit references (W1, W2, Point 1, etc.) or implicit topical connections
3. Extract the complete response content that addresses each weakness
4. Assign confidence scores (0-1) based on the clarity and directness of the mapping
5. Mark as "No Response" if a weakness is not addressed in the rebuttal
6. Be conservative with confidence scores - only use high scores (>0.8) when the mapping is very clear
7. Preserve the exact wording from the rebuttal when extracting responses

CRITICAL RULE - NO SHORTCUTS OR REFERENCES:
You must NEVER use summarizing phrases or references like "[Same content as W2 response]", "[Similar to above]", "[As mentioned earlier]", etc. Always copy the complete, verbatim text from the rebuttal for each weakness point, even if the same rebuttal section addresses multiple weaknesses. Repetition is required and expected - do not try to avoid it.

[User Prompt]
Please map each weakness point in <weakness_points> to its corresponding rebuttal response from <rebuttal_text>:

<weakness_points>
`{weaknesses_list}`
</weakness_points>

<rebuttal_text>
`{rebuttal_text}`
</rebuttal_text>

MANDATORY REQUIREMENTS:
- Output a mapping line for EVERY weakness point (W1, W2, W3, ... up to W{len(weakness_points)})
- Use exactly "W[number] -> R[number]:" or "W[number] -> No Response" format
- Include confidence score in parentheses for every mapping
- Do not skip any weakness numbers
- If you cannot find a rebuttal response, write "No Response" instead of omitting the line

CRITICAL: You MUST provide a mapping for EVERY weakness point listed above. Do not skip any weakness points.
ABSOLUTELY CRITICAL - NO SUMMARIZATION OR SHORTCUTS FOR REBUTTAL CONTENT:
- ALWAYS copy the COMPLETE, FULL, VERBATIM text from the rebuttal for each weakness, even if content is identical to previous responses
- NEVER use summarizing or abbreviated phrases like "[Same content as W2 response]" or "[Similar to above]" or "[Full content identical to W5 response]" - always provide the complete original text
- Do NOT abbreviate, summarize, or reference other responses
- If the same rebuttal section addresses multiple weaknesses, copy the ENTIRE text in full for each relevant weakness
- NEVER add any commentary, explanation, or meta-text beyond the actual rebuttal content

<output_format> (use EXACTLY this format):
W1 -> R1: [Specific rebuttal content addressing W1] (Confidence: 0.xx)
W2 -> R2: [Specific rebuttal content addressing W2] (Confidence: 0.xx)
W3 -> R3: [Specific rebuttal content addressing W3] (Confidence: 0.xx)
...continue for ALL weakness points...

If no rebuttal response exists for a weakness:
Wx -> No Response (Confidence: 1.0)
</output_format>

Figure 5: Prompt used for mapping review segments with rebuttal segments.

[System Prompt]
You are an expert in academic paper review classification. Your task is to identify from which perspective the reviewer is raising concerns or questions about the paper.

[User Prompt]
The following review point is a weakness or question raised by a reviewer during peer review. Please classify this review point based on the PERSPECTIVE from which the reviewer is critiquing or questioning the paper:

1. **Experiments**: The reviewer is questioning experimental **setup and design**. This includes missing or insufficient experiments, lack of ablation studies, weak baseline comparisons, unclear descriptions of datasets, or issues with hyperparameter selection.

2. **Writing**: The reviewer is concerned about writing quality - grammar, clarity, readability, ambiguous phrasing, typos, missing definitions of symbols/terms, unclear explanations of concepts.

3. **Presentation**: The reviewer is critiquing presentation and organization - figures, tables, and organization issues, unclear plots, missing legends, poor formatting, misplaced content, overall paper structure making it hard to follow.

4. **Theory**: The reviewer is questioning theoretical aspects - incorrect mathematical derivations, flawed assumptions, weak theoretical justification, missing proofs, inconsistency between claims and formulas.

5. **Novelty**: The reviewer is questioning novelty and originality - lack of novelty or originality, overlap with prior work, incremental contribution, insufficient differentiation from existing methods.

6. **Reproducibility**: The reviewer is concerned about reproducibility - missing implementation details, absent code or pseudo-code, hyperparameters not specified, insufficient information to reproduce results.

7. **Evaluation**: The reviewer is concerned about how the experimental results are **measured, analyzed, and interpreted**. This includes the use of inappropriate or missing evaluation metrics, insufficient analysis of results, or inconsistencies between reported results and the paper's claims.

8. **Miscellaneous**: Content that is not a direct review point (weaknesses, questions, suggestions) about the paper. This includes polite remarks, Summative or transitional comments, summaries of the paper's or review's content, or irrelevant text.

Please analyze the following review point and identify from which perspective the reviewer is raising their concern. Respond with ONLY the category name (Experiments, Writing, Presentation, Theory, Novelty, Reproducibility, Evaluation, Miscellaneous).

Review point to classify:
`{weakness_content}`

Perspective:

Figure 6: Prompt used to label which perspective a review segment belongs to.

[System Prompt]
You are a precise, deterministic classifier for rebuttal responses.

[User Prompt]
Return only a compact JSON object: {"impact": "<one_of:[CRP,SRP,VCR,DWC,DRF]>"}.
Categories:
- CRP: Concrete revision already made or concrete, verifiable artifact provided (new text/sections, new experiments/tables/figures, code/data links, numbers).
Cues: "We added/updated...", "Section X rewritten...", "New ablation in Sec. ... shows ...", "Code/data at ...".
- SRP: Specific revision plan committed, but not yet implemented; where/what to revise is specific.
Cues: "We will add ablation in Sec. X...", "We will redraw Fig. ...", "We will clarify definitions in §...".
- VCR: Vague promise to revise; no concrete actions, locations, or artifacts.
Cues: "We will revise accordingly.", "We will improve writing/clarity.".
- DWC: Defend current paper as-is; no new change proposed.
Cues: "Already covered in Sec. ...", "Setup is standard", "Claim stands".
- DRF: Shift issue to reviewer or avoid underlying point; no change offered.
Cues: "Reviewer misinterprets ...", "Out of scope ...", "Reviewer phrasing is incorrect".

Figure 7: Prompt used to label which impact category a rebuttal segment belongs to.

[System Prompt]
You are a professional reviewer. Provide a comment such as weakness, question or suggestion on the given paper in 1 to 3 sentences.

[User Prompt]
Request: From the perspective of xxx, provide a comment on the following paper.
<PAPER_CONTEXT>

Figure 8: Prompt used to generate review segments by different LLM baselines.

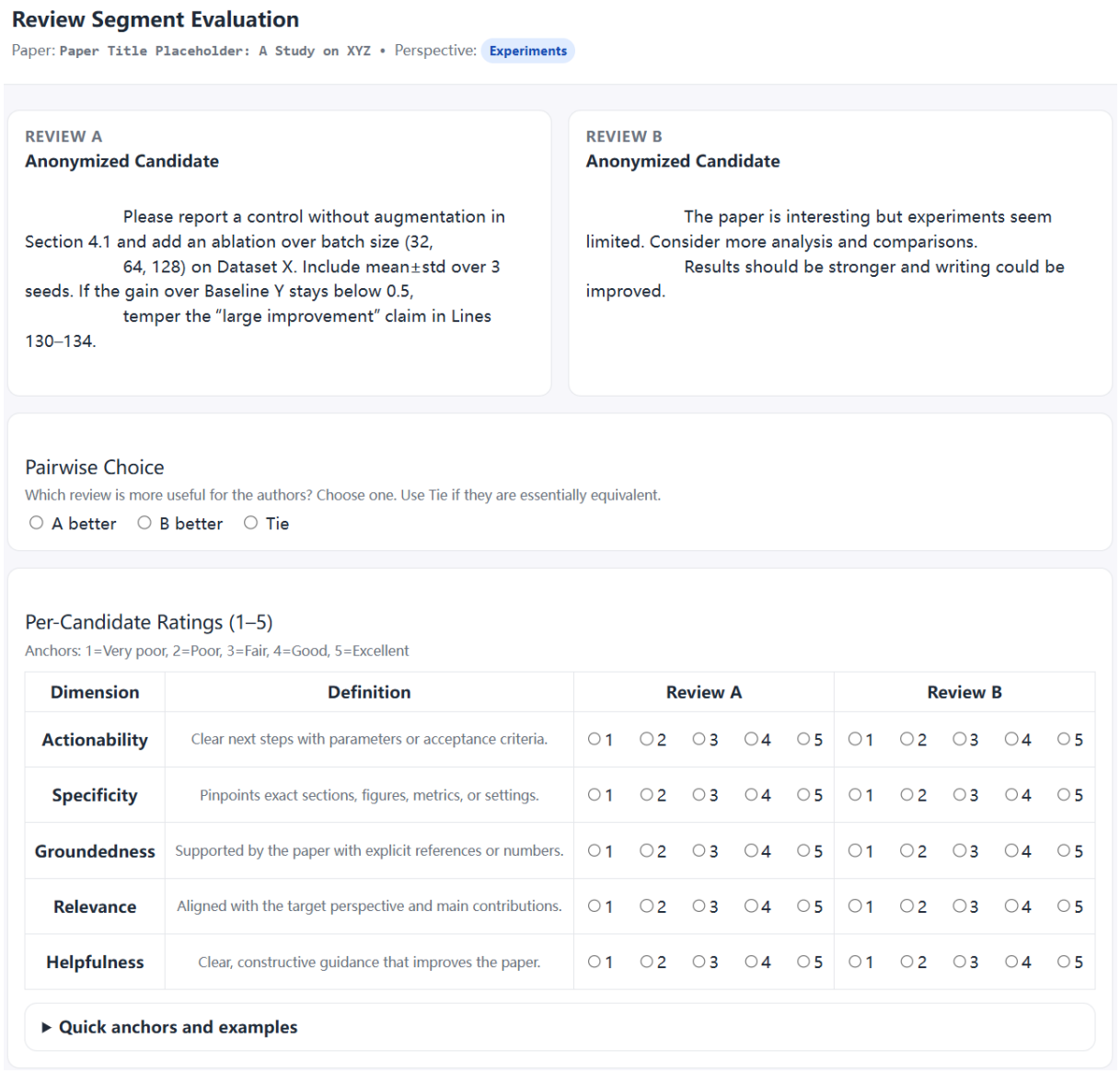

Figure 9: The interface of our human expert evaluation. The page contains 2 tasks: pairwise preference and per-candidate 1–5 ratings.

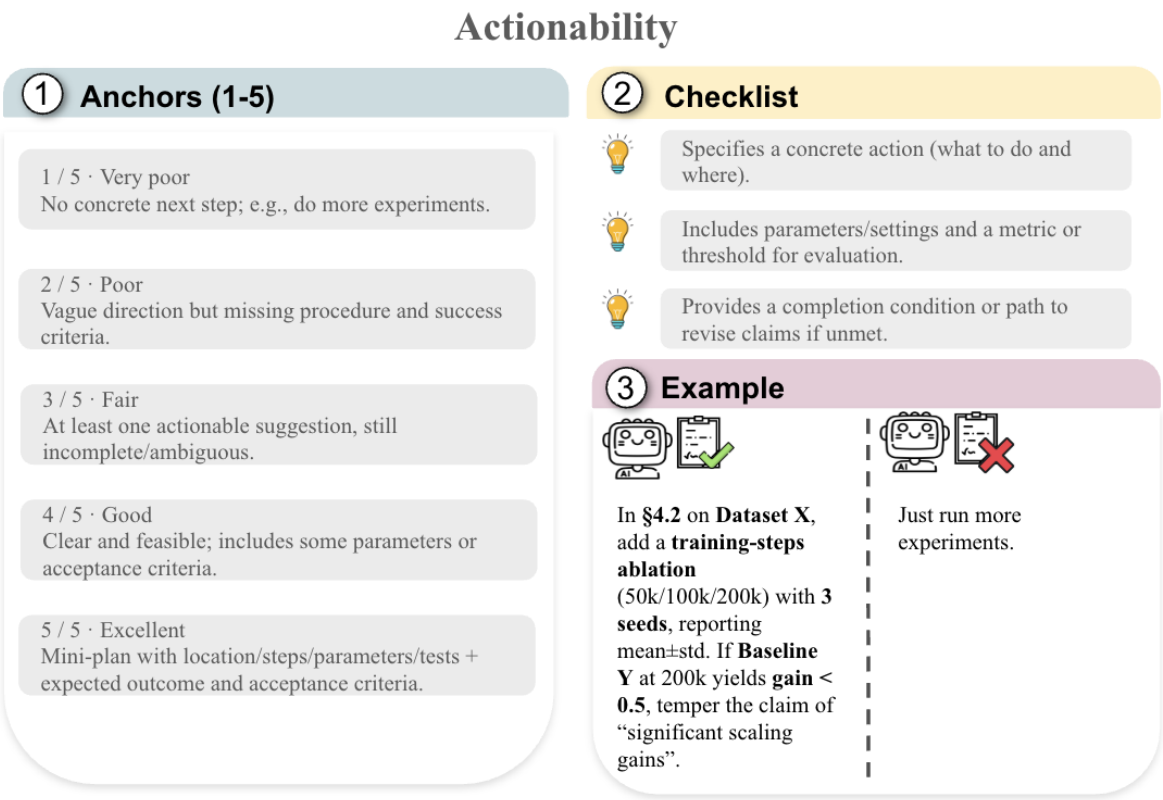

Figure 10: Comparison guidelines for the "Actionability" criterion.

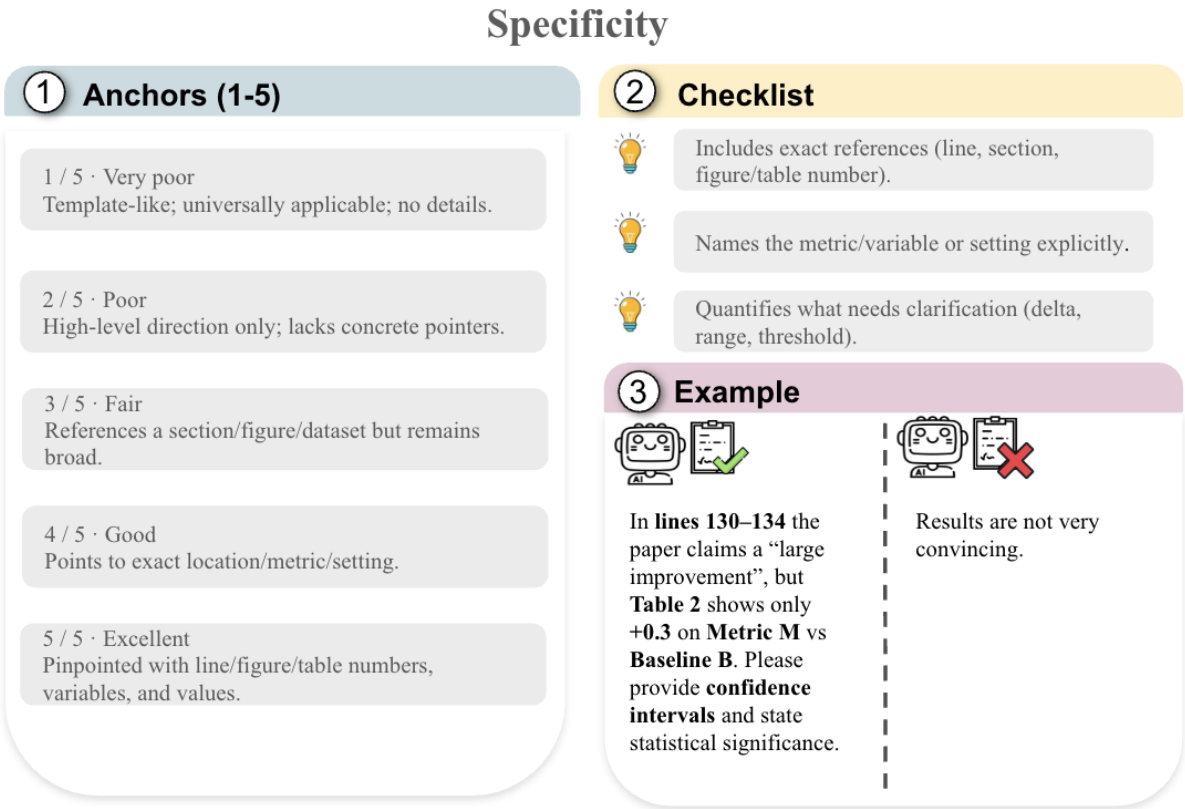

Figure 11: Comparison guidelines for the "Specificity" criterion.

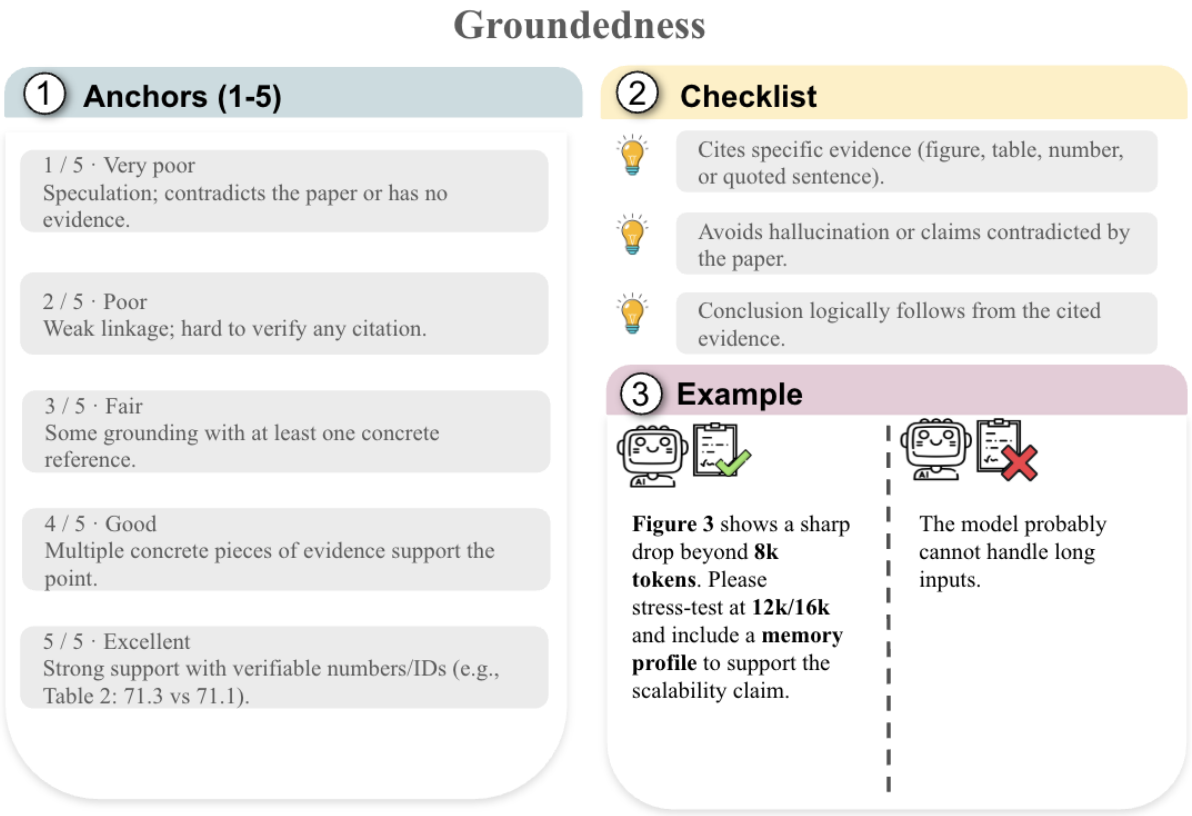

Figure 12: Comparison guidelines for the "Groundedness" criterion.

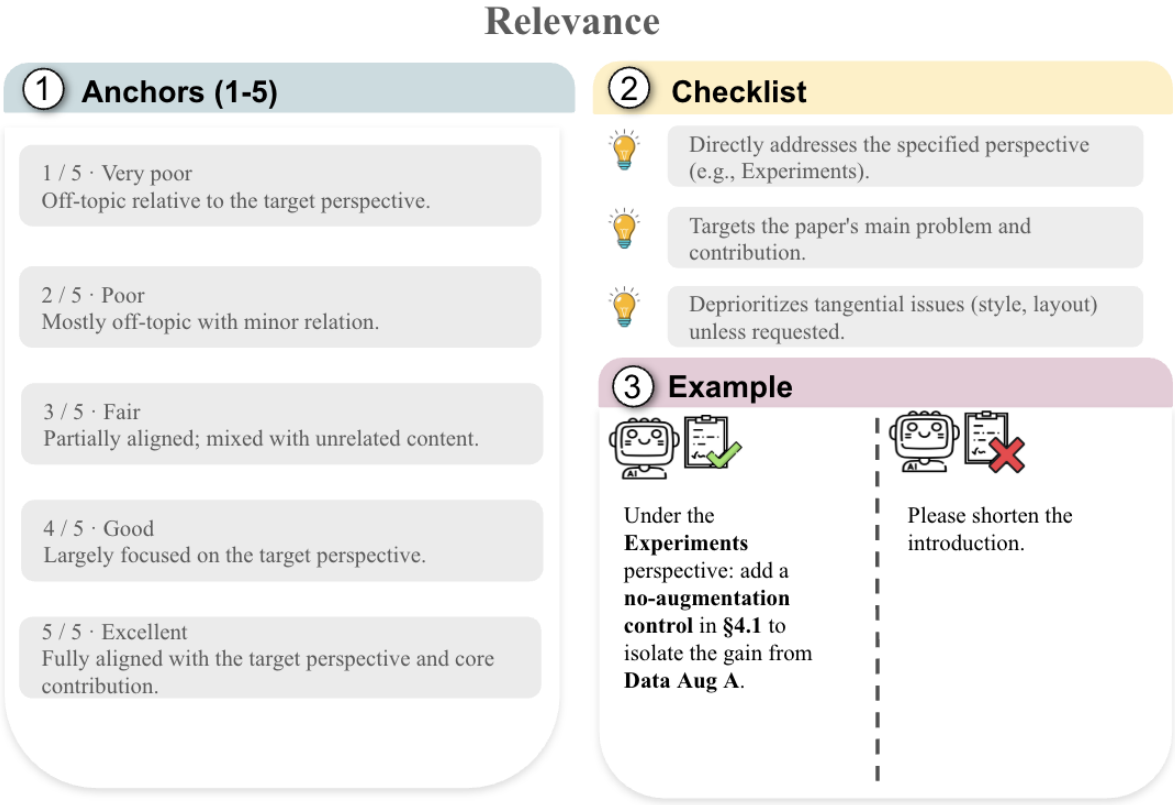

Figure 13: Comparison guidelines for the "Relevance" criterion.

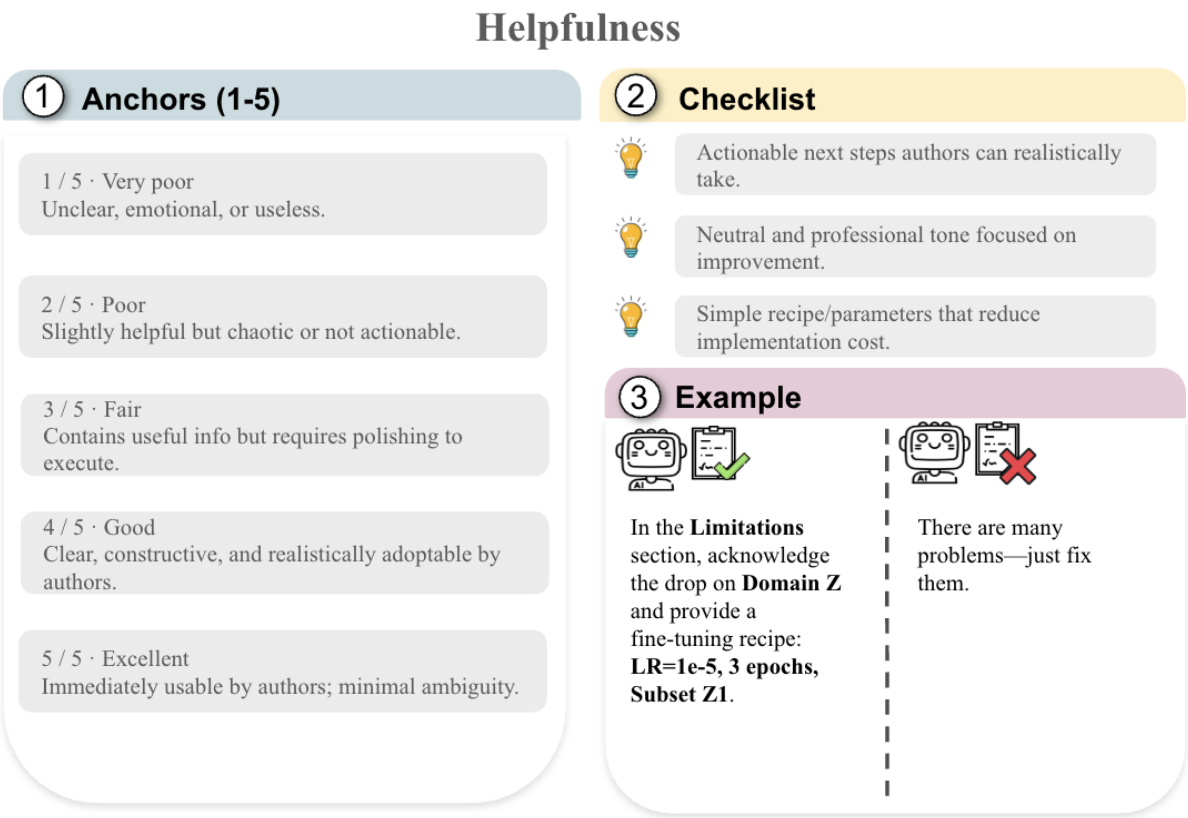

Figure 14: Comparison guidelines for the “Helpfulness” criterion.

[System Prompt]
You are an expert in evaluating peer review quality.
Your task is to assess a peer review comment from multiple dimensions and provide scores (1-5) with detailed reasoning for each dimension.

[User Prompt]

Please evaluate the following peer review comment based on the scoring rubric provided.

Scoring Rubric

Actionability (1-5)
1. Very poor: No concrete next step. Vague remarks like "improve experiments."
2. Poor: A possible step is implied but not described. No criteria for success.
3. Fair: At least one concrete suggestion, but incomplete or underspecified.
4. Good: Clear, feasible steps with some parameters or success criteria.
5. Excellent: A short plan with steps, locations in the paper, parameters or tests, and what outcome would address the issue.

Specificity (1-5)
1. Very poor: Generic template text that could apply to any paper.
2. Poor: Mentions broad areas but no details.
3. Fair: Refers to a section, figure, dataset, or claim but stays broad.
4. Good: Points to exact sections, figures, metrics, or settings.
5. Excellent: Pinpoints precise passages or numbers and names exact variables, metrics, or ablation locations.

Groundedness (1-5)
1. Very poor: Speculative, incorrect, or contradicted by the paper.
2. Poor: Weak link to the paper; no verifiable reference.
3. Fair: Partly grounded with at least one reference to paper content.
4. Good: Well supported with references to specific content.
5. Excellent: Strongly supported with exact identifiers or numbers from the paper (for example "Table 2 shows 71.3 vs 71.1 and the claim of a large gain is not supported").

Relevance (1-5)
1. Very poor: Off topic relative to the target perspective or the main paper issues.
2. Poor: Mostly off topic with minor relevant content.
3. Fair: Partially aligned. Mixes relevant and irrelevant feedback.
4. Good: Mostly aligned with the target perspective.
5. Excellent: Fully aligned with the target perspective and the paper's main contributions.

Helpfulness (1-5)
1. Very poor: Unclear, hostile, or not useful.
2. Poor: Slightly useful but confusing or impractical.
3. Fair: Some useful content, needs refinement to be actionable.
4. Good: Clear, constructive, and practically useful.
5. Excellent: Directly helps the authors improve the paper with minimal ambiguity.

Paper Content:
`paper_content`

Review Perspective:
`perspective`

Review Comment to Evaluate:
`review_text`

Please provide scores (1-5) for each dimension along with your reasoning. Be critical and precise in your evaluation.

You MUST respond with a valid JSON object in the following format (no markdown code blocks, just raw JSON):
{{
"actionability_score": <1-5>,
"actionability_reasoning": "<brief explanation>",
.....
}}

Figure 15: Prompt used for point-wise evaluation for LLM-as-a-judge.

[System Prompt]
You are an impartial judge comparing the actionability of two peer-review segments. Actionability means the feedback gives concrete, specific, and feasible guidance that authors can directly implement. Prefer segments that:
-specify what to change (methods, experiments, analyses, writing),
-localize where to change (section/figure/table/scope),
-propose how to change (procedures, metrics, datasets, ablations, edits),
-include verifiable artifacts or acceptance criteria (e.g., code/data, new experiments, numbers to report).
Output JSON schema:{ "winner": "A" | "B", "justification": "1–2 sentences citing the most decisive actionable cues." }

[User Prompt]
Task: Choose the more actionable review segment for the specified perspective. Remember: no ties.
Perspective:
<PERSPECTIVE>
Paper context:
<PAPER_CONTEXT>
Segment A:
<REVIEW_SEGMENT_A>
Segment B:
<REVIEW_SEGMENT_B>

Figure 16: Prompt used for pairwise evaluation for LLM-as-a-judge on Actionability.

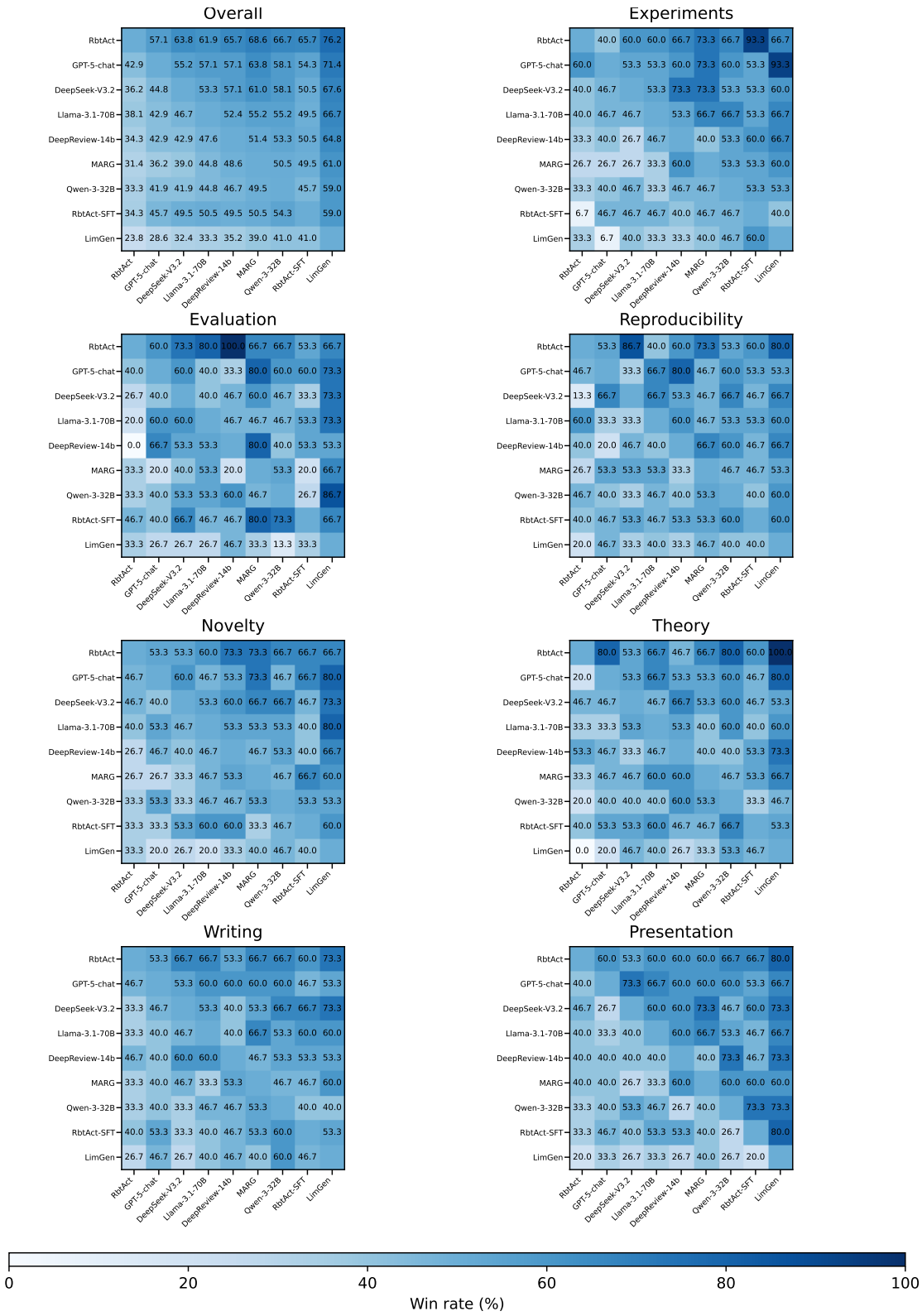

Figure 17: Pairwise win rates heatmaps by perspective (row beats column) of §5.3

Case Study of Why RBTACT Reviews More Actionable

RBTACT Review Feedback:
The reported gains may be entangled with augmentation choices (MixUp/CutMix) and unquantified run-to-run variance, which leaves robustness in §5.2 not directly stress-tested; to disambiguate these factors, re-train without MixUp/CutMix under a fixed seed with three independent trials, report mean±std Top-1 in Table 3, and add a Corrupted ImageNet check aligned with the claims.

Qwen-3-32b Review Feedback:
The current training description under-specifies key regularization and sizing decisions, making it unclear how sensitive results are to setup and randomness. It would help to include a compact ablation that revisits weight decay and batch size.

Analysis: It specifies exact factors to remove, run counts, where to report (Table 3), and a concrete robustness check—clear steps and success criteria.

RBTACT Review Feedback:
In Figs. 2-3, small axis labels, palettes that are not color-blind-safe, missing y-axis units, and cramped legends reduce interpretability; revising the plots with adequately sized labels, an OKLCH-based color-blind-safe palette, explicit y-axis units, legends moved below the panels, and captions define each metric and sample size would make the figures verifiable.

Llama-3.1-70b Review Feedback:
Figure readability is limited by low apparent resolution, small typography, weak color separability, and inconsistent styling, with legend placement adding clutter. You may consider increasing DPI and font sizes, adopting more distinguishable colors under a unified style, clarifying captions in figure. 2.

Analysis: It names specific figures, thresholds, concrete palette guidance, and exact caption content—easy to implement and verify.

RBTACT Review Feedback:
Extend Sec. 4.2: add baselines and SOTA with identical prompts; report macro-F1 and calibration with 95% CIs via paired bootstrap over papers. Include error taxonomy in Appendix.

MARG Review Feedback:
Broaden evaluation to more datasets and compare against stronger models. Include confidence intervals where appropriate and a brief qualitative error analysis.

Analysis: It pins down section, comparators, metrics, CI method, and deliverables, giving precise implementation and acceptance checks.

Figure 18: Case study comparing review feedback on Actionability from Experiment, Presentation and Evaluation perspectives.