RATING-SCALE DRIFT IN AUTOMATED PAPER REVIEWERS AFTER DOMAIN-SPECIFIC REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Domain-specific post-training can improve an automated reviewer's agreement with its training labels, but does it preserve agreement with human ratings? We study this question by adapting a Qwen3.5-9B reviewer, initially fine-tuned on human reviews, with reinforcement learning against generated domain scores. In two source-matched Computer Science and Mathematics runs, active-domain Rating error decreases while human-referenced validation error increases by 43.3% and 29.3%, respectively. A separate offline evaluation compares these checkpoints and four additional, unmatched domain checkpoints with the same SFT baseline on jointly parsed human-rated papers. All six checkpoints have higher paired Rating error than SFT, accompanied by upward shifts or near-constant predictions. Some checkpoints also improve aggregate rule reward despite worse Rating agreement. The SFT baseline is itself concentrated, so these comparisons measure changes from a fixed, imperfect reviewer. We characterize these observations as rating-scale drift and report agreement, signed bias, and output concentration separately. The results document a separation between generated-domain progress and human-rating retention in the evaluated runs, without identifying its causal mechanism.

1 INTRODUCTION

Automated paper reviewers produce numerical recommendations alongside textual critiques (Idahl & Ahmadi, 2025; Zhu et al., 2025). Comparing these ratings across model checkpoints requires a stable relationship to the evaluation reference. If a reviewer continues to report the same Rating field after domain adaptation, a score of 6 may be interpreted as comparable to a score of 6 before adaptation. Whether this interpretation remains justified is an empirical question. A domain-specific scoring standard may be useful, but progress under that standard does not by itself establish continuity with an earlier human reference.

This question builds on existing evidence about reviewer training and evaluator reliability. ReviewRL studies RL-based paper reviewing and cold-start rating collapse (Zeng et al., 2025), while comparison-based paper evaluation addresses limitations of absolute score scales (Zheng et al., 2026). More broadly, numerical bias can change after alignment (Sato et al., 2026), and updated judge models must be assessed for compatibility with earlier evaluation distributions (Singh et al., 2026). We examine a specific setting within these concerns: when a human-review-trained reviewer is further optimized against generated domain scores, does improvement against those scores preserve agreement with fixed human ratings? We use *rating-scale drift* to describe changes in rating level or concentration after adaptation, assessed alongside human-referenced error.

We study this question using an image-based Qwen3.5-9B reviewer (Qwen Team, 2026) initialized from a shared checkpoint fine-tuned on human reviews from DEEPREVIEW (Zhu et al., 2025). Domain adaptation uses group relative policy optimization (GRPO) and generated reviews, while the held-out human reference remains an evaluation target. Figure 1 summarizes the design. Our central evidence comes from two source-matched Computer Science and Mathematics runs evaluated on both active-domain and human-reference validation sets. Four additional domain checkpoints provide supporting observations rather than matched replications. A separate offline evaluation compares each checkpoint with SFT on papers for which both policies produce parsable ratings.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

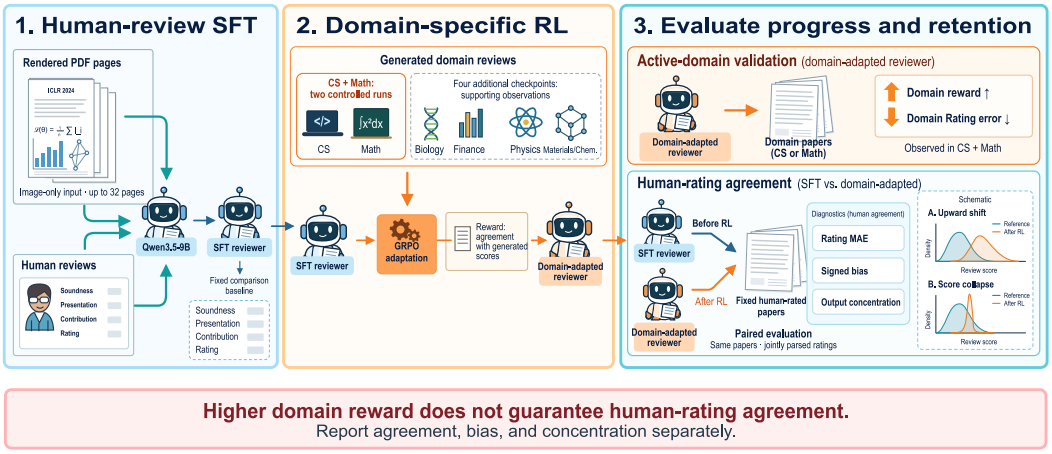


Figure 1: Empirical study of human-rating retention under generated-domain adaptation. A shared human-review SFT baseline is adapted against generated scores, then evaluated for active-domain progress and agreement with fixed human ratings. CS and Math supply two controlled runs; four additional checkpoints supply supporting observations. Retention comparisons pair jointly parsed ratings and report MAE, signed bias, and output concentration. Distribution sketches are schematic; the SFT baseline is not assumed to be calibrated.

The two controlled runs improve agreement with generated active-domain scores while increasing human-referenced validation Rating MAE by 43.3% and 29.3%. The offline comparisons also show higher paired Rating MAE for all six available checkpoints. Output distributions reveal different forms of change: ratings may move upward while retaining variation, or concentrate almost entirely on one value. The SFT baseline already assigns 6 to 82.4% of its parsed papers, so these results describe additional changes from an imperfect starting point. Moreover, some checkpoints improve aggregate rule reward despite worse Rating agreement. The aggregate therefore does not fully describe whether the numerical recommendations retain their relationship to the human reference.

Our contribution is an empirical characterization of human-rating retention under generated-domain adaptation. The paired comparisons measure changes relative to a shared SFT baseline; signed bias and output concentration characterize aspects of those changes that aggregate reward does not specify. These findings support evaluating active-domain progress and human-reference retention separately. They concern the reported model and training runs, and do not isolate generated labels, score-prior differences, or a particular RL component as the cause of the observed changes.

2 Related Work

Automated paper evaluation and reviewer training. PeerRead provides an early dataset linking manuscripts, reviews, and decisions (Kang et al., 2018). DeepReview develops review data and a reasoning-oriented reviewer, while OpenReviewer trains a specialized model on expert reviews (Zhu et al., 2025; Idahl & Ahmadi, 2025). ReviewRL combines SFT and RL with human-derived rating targets and studies cold-start rating collapse, including the role of initialization and score balancing (Zeng et al., 2025). CNPE instead trains comparison-based paper evaluators and aggregates preferences into a ranking (Zheng et al., 2026). Our study retains absolute Rating outputs and examines their agreement with a fixed human reference after further adaptation using generated domain scores. It does not introduce reviewer RL or rating collapse as new phenomena, nor test whether a comparison-based formulation prevents the observed changes.

Numerical scoring bias and evaluator reliability. LLM judges can produce skewed rating distributions and exhibit anchoring effects across evaluation dimensions (Stureborg et al., 2024). Comparing base and instruction-tuned models, Sato et al. (2026) find that alignment can increase numerical bias, although the effect on evaluation performance varies. These results motivate inspecting score

distributions alongside agreement with human targets. We measure changes in signed bias and output concentration together with paired Rating MAE. These diagnostics address distinct properties: concentrated predictions can have moderate absolute error, while a shifted distribution can retain output variation. Neither variation nor low MAE alone establishes correct ranking, and our analysis is not a complete calibration assessment.

Adaptation, generalization, and reward optimization. Fine-tuned judges can perform strongly in-domain while having limited generalization (Huang et al., 2025). Singh et al. (2026) study future-proofing, backward compatibility, and continual judge updates under changing response distributions. RL can also improve judgment performance, as demonstrated by JudgeLRM (Chen et al., 2025); degradation is not an inevitable consequence of judge training. Reward overoptimization (Gao et al., 2023) and the relationship between reward maximization, KL-control, and distribution matching (Korbak et al., 2022) provide broader context for distinguishing an optimization objective from properties to be retained. Our setting measures a particular retained quantity—agreement with human paper ratings—when the adaptation target is generated domain scores. The observed separation does not by itself identify a reward-hacking or forgetting mechanism.

3 Rating-Scale Retention under Domain RL

The central quantity is the change in a fixed human-referenced measurement after RL, rather than target-domain performance alone. Let π_0 denote the image-based SFT reviewer and π_d the policy obtained by applying domain RL on domain d. For a paper x_i in the held-out human set $\mathcal{D}_{\mathrm{H}}$, let $r_i \in [1, 10]$ be the average available human Rating and let $\hat{r}_i(\pi)$ be the parsed rating from policy π. Because parsing can fail for one policy but not another, we compare π_d with π_0 on their jointly parsed index set I_d.

This construction assigns distinct roles to the two evaluation distributions. Active-domain validation asks whether optimization succeeded under the new labels. The held-out human set asks whether agreement with the earlier human reference is retained. The first question cannot answer the second: a lower error on generated domain labels is compatible with either better domain judgment or closer matching to their score distribution. Holding both papers and human targets fixed makes the SFT-to-RL change a within-reference comparison rather than a comparison between unrelated test sets.

We define human-referenced Rating error and its change as

$$E_{\mathrm{H}}(\pi; I_d) = \frac{1}{|I_d|} \sum_{i \in I_d} |\hat{r}_i(\pi) - r_i| \,, \tag{1}$$

$$\Delta_d = E_{\mathrm{H}}(\pi_d; I_d) - E_{\mathrm{H}}(\pi_0; I_d). \tag{2}$$

A positive Δ_d means that the domain policy agrees less closely with the held-out human ratings than SFT does. The metric makes no claim that a human average is a complete measure of review quality. It tests the narrower property needed for scale comparability: whether later checkpoints preserve the supervised relationship to the same reference labels.

MAE alone does not identify how the scale changed. We therefore report signed bias

$$B_{\mathrm{H}}(\pi; I_d) = \frac{1}{|I_d|} \sum_{i \in I_d} (\hat{r}_i(\pi) - r_i) \tag{3}$$

and the share of predictions assigned to the modal rating. When positive bias approaches MAE, most errors are overratings and the scale has moved upward. A high modal share instead indicates that the reviewer is using little of the available rating range. The summaries are complementary: a policy can retain several output values at a shifted level, or it can reduce variation by concentrating on one value.

We use these quantities as complementary diagnostics of drift. Signed bias measures the average offset from human targets, and modal share measures output concentration. MAE supplies the paired measure of agreement. A nearly constant central prediction can have moderate MAE, while a dispersed policy can still be systematically shifted. Output variation alone does not establish correct ordering of papers, and these summaries do not constitute a full calibration or ranking evaluation.

The reward geometry explains why these diagnostics are necessary. Its Rating component is $1 - \alpha|\hat{r} - r|$ with $\alpha = 0.2$. If a predictor were restricted to output a constant c, maximizing its expected Rating reward would be equivalent to minimizing $\mathbb{E}|c - r|$, whose solutions are medians of the training-label distribution. This observation does not imply that GRPO learns a constant predictor. It shows that a shifted or concentrated training-label distribution can reward a shifted or concentrated output policy. The full rule reward also sums Rating with three additional score components, so improvement in the aggregate can conceal deterioration in Rating. Human-referenced error, bias, and concentration test aspects of the score that the aggregate reward does not protect explicitly.

4 Experimental Setup

The setup follows the comparison defined in Section 3. Human reviews provide the supervised training source. Generated domain reviews then provide the shifted source used for RL. The held-out human set supplies an evaluation reference without adapting its labels or examples. Keeping these roles separate prevents the generated annotations from serving simultaneously as the optimization target and the human-reference standard.

4.1 Human Reference and Image-Based SFT

We derive SFT and test examples from DeepReview-13K (Zhu et al., 2025) while preserving its original split. Each paper is rendered into page images and truncated to at most 32 pages. The prompt contains one image placeholder per page and requests a detailed review followed by Soundness, Presentation, Contribution, and Rating fields. The first three scores lie in $[1, 4]$, and Rating lies in $[1, 10]$. Paper-level evaluation targets average the available human scores. SFT retains individual review text as the assistant target, so one paper can contribute multiple review targets.

The reviewer starts from Qwen3.5-9B and receives page images only. This input choice follows pixel-based document understanding (Kim et al., 2022; Lee et al., 2023); it fixes the input interface across SFT and RL rather than introducing a new document architecture. SFT uses FSDP2 on eight GPUs, maximum sequence length 32,768, global batch size 16, learning rate 5×10^{-6}, cosine decay with 1% warmup, weight decay 0.1, gradient checkpointing, and one epoch. We select the step-800 checkpoint before examining any domain-RL result. Appendix B gives the complete optimization settings.

We evaluate on the first 200 papers of the PDF-only DeepReview test split. The choice follows the saved evaluation protocol rather than a stratified sampling design and is a limitation of the study. Rating is parsed for 193 SFT outputs. Every paired retention comparison below uses only examples parsed by both the SFT policy and the corresponding domain policy, rather than replacing missing predictions or comparing policies on different subsets.

4.2 Shifted Domain Training Sources

ArxivReview-Domains contains recent arXiv papers from six top-level domains. For each paper, we download the PDF, render its pages, and send only the page images to a vision-capable OpenAI-compatible endpoint using the DeepReview output schema. A record is kept only when PDF download, rendering, review generation, and all score fields succeed. Each domain is split deterministically into 95% train and 5% validation with seed 42.

The domain labels are generated, not human authored. We use them to create a realistic post-training distribution shift, but never treat them as a human-quality evaluation set. Their Rating means range from 6.42 to 7.38, whereas the human DeepReview training source has mean 5.21 (Appendix A). This difference is relevant because it supplies a plausible direction for scale movement. It is not causal evidence by itself: the document mix, generator, aggregation procedure, and training duration also differ.

Starting from the same SFT checkpoint, each run trains on one active domain. The controlled Computer Science and Mathematics runs validate both on the active domain and on held-out DeepReview before RL and throughout training. They use four responses per prompt, train batch size 16, actor learning rate 10^{-6}, KL coefficient 0.01, maximum prompt length 24,000, maximum response length 2,048, and three epochs. The Biology, Finance, Physics, and Materials Chemistry policies are

Table 1: Human-referenced agreement before domain RL on the 200-paper DEEPREVIEW PDF subset. Higher rule reward is better; lower Rating MAE is better. Parsed n is the denominator for Rating MAE.

Model	Rule reward ↑	Parsed n	Rating MAE ↓
Qwen3.5-9B Base	−0.5158	46	3.1554
Qwen3.5-9B Image SFT	1.9752	193	0.8377

earlier PDF-only checkpoints rather than source-matched reruns. We use those four checkpoints only to ask whether the same retention failure appears again; we do not compare domains as if their runs were exchangeable.

4.3 REWARD, METRICS, AND UNCERTAINTY

The parser reads the final assistant turn and extracts the four score fields. Predictions are clamped to their valid ranges. With s, p, c, r denoting Soundness, Presentation, Contribution, and Rating, the rule reward is

$$R = \sum_{q \in \{s,p,c\}} (0.5 - 0.2|\hat{q} - q|) + (1.0 - 0.2|\hat{r} - r|) \,. \tag{4}$$

The maximum reward is 2.5, and an output missing any required field receives -1. We report aggregate reward because it is the optimized checkpoint signal. Rating MAE, signed bias, and modal share separately describe human-referenced agreement, offset, and concentration. The distinction is important: reward aggregates several score dimensions, while our analysis focuses on changes in Rating.

For every domain policy, we compute Δ_d on its paired intersection with SFT. We obtain a 95% confidence interval by resampling the paired per-paper error differences 10,000 times with replacement, using seed 42. The resampling unit is the paper. This procedure quantifies uncertainty for the change in Rating error while preserving the pairing between policies. It does not account for variation across training seeds because only one training run is available per domain.

5 RESULTS

5.1 THE SFT REVIEWER AS A RETENTION BASELINE

The SFT checkpoint provides the fixed starting point for the retention analysis. Table 1 reports the base model and image-based SFT on the same 200-paper test subset. Rule reward is averaged over all papers and penalizes missing fields; Rating MAE is computed on each model's parsed ratings. The reported MAE is 3.155 for the base model and 0.838 for SFT, but their parsed subsets differ substantially. These numbers therefore do not establish a paired improvement in rating accuracy or human-level reviewing. SFT supplies 193 parsed ratings for subsequent comparisons, and its already concentrated output distribution is examined in Section 5.4.

The SFT policy is therefore the retention baseline, rather than the base model or an unrelated API reviewer. The available API evaluations use different image compression and training conditions, so they cannot answer the paired question studied here. We exclude them from the evidence chain and ask only what changes when domain RL starts from this particular human-trained reviewer.

The parse counts also delimit what can be inferred. The base model produces only 46 parsed ratings, so its MAE describes a selective subset and is not a competitive baseline result. SFT produces 193 parsed ratings and is evaluated on the same saved test protocol used for the RL checkpoints. We use the base row only to establish that SFT created a workable structured reviewer. All substantive retention claims compare an RL checkpoint directly with SFT on a paired set.

5.2 DOMAIN VALIDATION GAINS AND HUMAN-REFERENCE LOSS IN TWO RUNS

The controlled Computer Science and Mathematics runs establish the central before–after pattern. As Table 2 shows, the active-domain reward increases in both runs. Active-domain Rating MAE also

Table 2: Validation at initialization and at the selected final checkpoint for the controlled GRPO runs. Each run improves on its active domain while its DEEPREVIEW Rating MAE increases.

		Rule reward ↑		Rating MAE ↓	
RL domain	**Validation set**	**Initial**	**Final**	**Initial**	**Final**
CS	Arxiv-CS	0.9020	1.1673	1.2000	0.3750
	DEEPREVIEW	0.5793	0.5639	1.3069	1.8721
Math	Arxiv-Math	0.4280	0.5860	1.4167	0.5600
	DEEPREVIEW	0.5897	0.5623	1.2960	1.6756

Table 3: Offline retention on the 200-paper DEEPREVIEW subset. ΔMAE compares each RL policy with SFT on jointly parsed papers; intervals use 10,000 paired bootstrap resamples. The table's parsed n is each policy's own parse count, not its paired intersection with SFT. Rule reward is averaged over all 200 examples.

Checkpoint	**Reward ↑**	**Parsed n**	**Rating MAE ↓**	**Bias**	**Mode share**	**ΔMAE (95% CI)**
SFT	1.9752	193	0.8377	+0.241	82.4%	–
CS RL	1.8998	198	1.5866	+1.554	73.2%	+0.745 [+0.622, +0.868]
Math RL	1.7832	190	1.5100	+1.501	49.5%	+0.683 [+0.542, +0.825]
Bio RL	1.9085	192	0.9963	+0.570	95.3%	+0.139 [+0.057, +0.226]
Finance RL	2.0251	200	1.0119	+0.504	100.0%	+0.158 [+0.093, +0.231]
Physics RL	2.0088	199	1.0157	+0.518	99.5%	+0.159 [+0.093, +0.231]
Materials/Chem. RL	1.7152	199	2.5038	+2.494	100.0%	+1.634 [+1.468, +1.794]

decreases, from 1.200 to 0.375 for Computer Science and from 1.417 to 0.560 for Mathematics. By the training-domain criteria, both policies improve.

The human-referenced validation set moves in the opposite direction. Rating MAE rises by 43.3% in the Computer Science run and by 29.3% in the Mathematics run. The corresponding DEEPREVIEW rewards fall only slightly, by 0.0154 and 0.0274. This asymmetry matters. The aggregate reward makes the retention loss look modest, whereas the Rating component shows a large change in the score that downstream users are most likely to compare. The controlled runs therefore support the paper claim without requiring a cross-domain ranking: optimization succeeds on the active domain and simultaneously weakens agreement with the fixed human scale.

The two validation sets reveal different aspects of adaptation. Active-domain reward and Rating MAE favor the final checkpoints over initialization, whereas human-referenced Rating MAE worsens. Thus, improvement under the generated domain labels does not establish retention of agreement with human ratings. This before–after comparison does not determine an optimal intermediate checkpoint or demonstrate that a particular selection rule improves downstream performance.

5.3 OFFLINE PAIRED EVALUATION OF SIX AVAILABLE CHECKPOINTS

The offline paired evaluation asks whether the controlled result is isolated. Table 3 compares each domain policy with SFT on jointly parsed DEEPREVIEW papers. Every estimated change in Rating MAE is positive, and every 95% bootstrap interval lies above zero. The smallest increases occur for Biology, Finance, and Physics, at 0.139–0.159 points. Computer Science and Mathematics increase MAE by 0.745 and 0.683 points. Materials Chemistry has the largest change, 1.634 points.

The six rows do not support a claim that Materials Chemistry is intrinsically more vulnerable than Biology, or that the numerical ordering would persist under matched training. Only Computer Science and Mathematics are controlled source-matched reruns. The broader table supports a different and more limited conclusion: once evaluated against the same human reference, degradation is not confined to one domain checkpoint or one training trace. This recurrence motivates examining the output distributions rather than treating the result as ordinary variance in a single aggregate metric.

Pairing is important for the smaller changes. Each RL policy has between 190 and 200 parsed ratings, but its paired comparison uses only the intersection with the 193 parsed SFT ratings. Thus, the parsed n column does not give the paired denominator, which cannot exceed 193. Each ΔMAE compares the two policies on the same intersection rather than subtracting the separately computed MAEs

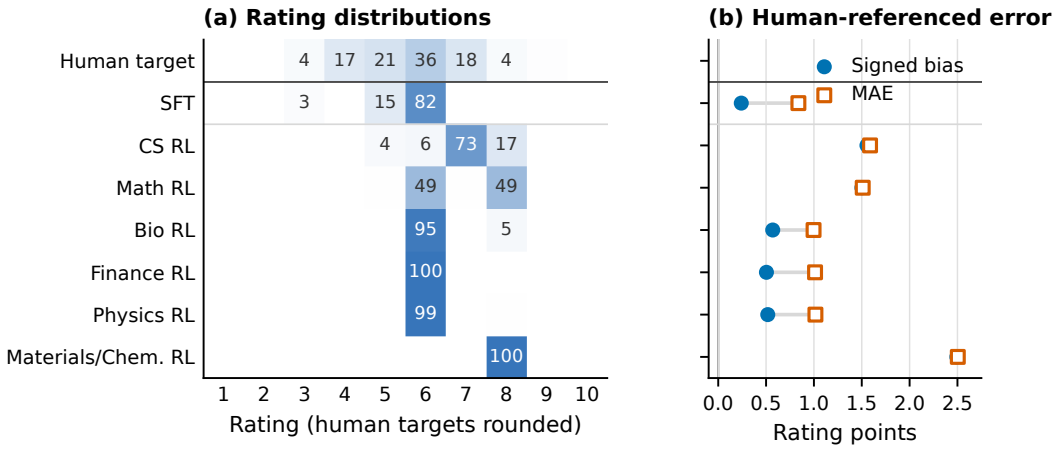


Figure 2: Two views of rating-scale drift on held-out human reviews. (a) Distribution of rounded human targets and parsed predictions; each model row is normalized over its parsed examples. (b) Signed bias and Rating MAE for the model rows. Computer Science and Mathematics mainly shift ratings upward, whereas several other policies concentrate almost entirely on one value.

in the table. The intervals for Biology, Finance, and Physics remain above zero even though their point estimates are much smaller than those of the controlled runs. These intervals describe variation over evaluation papers; they do not convert the supporting checkpoints into matched experiments or measure variation across training seeds.

5.4 Drift Appears as Scale Shift or Score Collapse

Figure 2 shows the form of the retention failure. The left panel places the rounded human targets above the prediction distributions, so changes in both location and spread are visible. The SFT policy already uses the scale narrowly and assigns 6 to 82.4% of its parsed outputs. Computer Science and Mathematics retain multiple output values but move mass upward. Their signed biases exceed $+1.5$ and nearly equal their MAEs, which means that most of their absolute error is systematic overrating rather than balanced noise.

The human targets provide a useful visual anchor. After rounding averages to the nearest integer for display, most targets lie between 4 and 7, with visible mass at every value in that range. SFT compresses this variation strongly toward 6, so the pre-RL reviewer already exhibits strong output concentration. Domain RL nevertheless changes the output further in a consistent direction. Computer Science moves most predictions to 7 and Mathematics divides nearly all predictions between 6 and 8. The relevant comparison is therefore not an idealized broad SFT distribution, but the additional movement away from the same human targets after RL.

Biology, Finance, Physics, and Materials Chemistry exhibit a different pattern. Biology assigns 6 to 95.3% of parsed examples. Finance predicts 6 for all 200 parsed examples, Physics predicts 6 for 198 of 199, and Materials Chemistry predicts 8 for all 199. These policies express almost no paper-to-paper variation in Rating. The resulting MAE can appear less dramatic for Finance and Physics because 6 lies near the center of the human target distribution. A moderate MAE is therefore not evidence that the reviewer still distinguishes stronger from weaker papers.

The two patterns describe different changes from the SFT baseline. An upward shift changes rating level while retaining some output variation; collapse removes most of that variation. Neither the remaining variation nor a low MAE proves that papers are correctly ordered. The distinction describes the observed predictions, rather than identifying why training produced them.

The aggregate rule reward provides the final link in the evidence chain. Finance and Physics obtain rewards of 2.0251 and 2.0088, both above the SFT value of 1.9752, even though their Rating MAEs are higher and their ratings are almost constant. Equation 4 sums four score terms and assigns -1 to an output with missing required fields. Finance parses all 200 ratings and Physics parses 199,

compared with 193 for SFT. Better output validity can therefore offset worse Rating agreement in the aggregate, just as changes in the other components could. The aggregate would favor these checkpoints over SFT even though their Rating error is higher; this comparison does not evaluate an intermediate-checkpoint selection strategy.

This result does not make the rule reward defective for its original purpose. The reward encourages structured completion and agreement across several review dimensions. A single aggregate value, however, does not reveal which component improved, whether ratings shifted upward, or whether predictions became constant. Human-referenced agreement and output concentration provide separate diagnostics; their value here is to expose changes hidden by the aggregate, not to establish a validated acceptance threshold.

6 Discussion and Limitations

The post-RL behavior is consistent with, but does not causally identify, adaptation to a changed rating prior. The generated domain sources have higher Rating means than the human training source, and the post-RL policies shift upward or concentrate near values favored by those sources. A causal decomposition would require matched label transformations, training durations, and seeds that are not available here. We therefore describe the observed behavior as rating-scale drift and treat the prior shift as an explanatory clue, not a proven mechanism.

The practical implication concerns evaluation design rather than a particular mitigation. A domain-RL run should be evaluated on two axes: progress on the active domain and retention on a fixed human-referenced set. The retention report should include at least an absolute-error measure, signed bias, and a measure of output concentration. The first tests agreement, the second detects a level shift, and the third measures how narrowly the available rating values are used. Reporting only one of them can hide a material failure: MAE alone can make a central constant prediction look acceptable, while bias alone does not reveal a low-entropy scale.

This separation also clarifies what an in-domain improvement means. A lower error against generated labels shows closer agreement with those labels; it does not determine whether the adapted reviewer makes better judgments under a human reference. Our sources differ in both documents and labels, so their disagreement cannot identify a change in label semantics. Evaluating against fixed human targets instead measures whether agreement with that reference improved or declined after adaptation.

A practical reporting recommendation follows from these observations: compare pre-RL and post-RL predictions on the same held-out human papers, state the paired denominator and parse coverage, and report agreement, bias, and concentration alongside domain reward. These measurements could inform checkpoint selection, but this study does not compare selection rules or validate retention thresholds. If used for selection, the human-reference validation set should be separate from the final test set.

A reference-model KL penalty constrains policy movement but does not by itself establish retention of rating agreement. Replay, distribution matching, or explicit constraints on the rating distribution are possible interventions, but our experiments do not compare them. The supported recommendation is to report human-rating retention separately; whether any particular intervention preserves it requires further evaluation.

The distinction between shift and collapse also matters for any later repair. If a policy preserves ordering and spread but moves the scale, a post-hoc mapping estimated from human anchors may correct part of the level error. A nearly constant policy has discarded the variation needed for such a mapping; changing its offset cannot recover paper-specific distinctions. We do not evaluate recalibration in this study, so this observation is not a mitigation result. It explains why the output distribution belongs in the retention test. Two checkpoints with similar MAE can differ substantially in output variation; determining whether that variation reflects paper quality requires a separate ranking evaluation.

The scope of the finding is equally important. Rating-scale drift does not imply that every capability of the reviewer deteriorates after domain RL, that the generated domain labels are incorrect, or that one universal scale is appropriate for every research field. It means that a checkpoint trained under one score distribution cannot be assumed to remain numerically comparable with a human-trained

checkpoint under another. A field-specific scale may be a legitimate design choice, but it should then be named and evaluated as a different scale. Problems arise when the same Rating field is retained and interpreted as continuous with the earlier reviewer even though its empirical mapping has changed. Our retention test detects that unannounced change; it does not adjudicate which domain-specific standard a community should prefer.

Several limitations bound the claim. The study measures structured-score agreement, not the factuality, issue coverage, novelty assessment, or usefulness of the review prose. It uses the first 200 papers of one held-out test split rather than a stratified or multi-benchmark evaluation. Human scores are themselves noisy, and their average is an imperfect reference, although it is the same reference used before and after RL. Only Computer Science and Mathematics have source-matched controlled traces; the other four policies establish recurrence but do not support causal domain comparisons. There is no matched comparison against continued SFT or adaptation on human domain labels, so the study cannot separate the effects of the training algorithm, supervision source, and domain content. Finally, each domain has one training seed, so the bootstrap intervals describe uncertainty over paired evaluation papers, not variation across RL training runs. These constraints make the conclusion narrower, but they do not alter the observed mismatch between domain reward and human-referenced rating retention.

7 Conclusion

We document rating-scale drift after domain-specific RL in an automated paper reviewer. Two source-matched runs improve agreement with generated active-domain scores while losing agreement with fixed human ratings. Offline paired evaluation of these and four additional, unmatched checkpoints finds higher Rating MAE relative to the same SFT baseline, accompanied by upward shifts or near-constant outputs. These observations do not isolate the cause of drift or establish a general effect across models and training runs. They support reporting fixed human-referenced agreement and output concentration alongside domain reward, rather than treating reward improvement as evidence that rating agreement has been retained.

AI use statement

Generative AI tools were used to generate the synthetic review annotations in ArxivReview-Domains, as described in Section 4.2. They also assisted with research ideation and execution through feedback on study framing and support for figure and code development, as well as manuscript drafting, editing, and literature search. The authors made the methodological decisions, ran the experiments, inspected the outputs, checked reported values against saved predictions and logs, and reviewed all AI-assisted text and artifacts. The authors take responsibility for the final content of this work.

Ethics statement

Automated reviews can contain incorrect critiques, reproduce field-specific biases, or be misused in high-stakes editorial decisions. The models and results in this paper are intended for research and author-assistance settings, not as replacements for human peer review. Generated arXiv reviews are treated as synthetic training signals, and all retention claims are evaluated against held-out human annotations. Any release should preserve source-dataset license restrictions, document generated-label provenance, and discourage use for undisclosed or fully automated reviewing.

Reproducibility statement

The project archive includes source data for Figure 2 and the script that regenerates its PDF, SVG, and PNG versions. Appendices A and B report the data statistics and hyperparameters used in the study.

REFERENCES

Nuo Chen, Zhiyuan Hu, Qingyun Zou, Jiaying Wu, Qian Wang, Bryan Hooi, and Bingsheng He. Judgelrm: Large reasoning models as a judge. *arXiv preprint arXiv:2504.00050*, 2025. URL https://arxiv.org/abs/2504.00050v3. Version 3.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866. PMLR, 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

Hui Huang, Xingyuan Bu, Hongli Zhou, Yingqi Qu, Jing Liu, Muyun Yang, Bing Xu, and Tiejun Zhao. An empirical study of llm-as-a-judge for llm evaluation: Fine-tuned judge model is not a general substitute for gpt-4. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 5880–5895, 2025. doi: 10.18653/v1/2025.findings-acl.306. URL https://aclanthology.org/2025.findings-acl.306/.

Maximilian Idahl and Zahra Ahmadi. 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)*, pp. 550–562. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-demo.44. URL https://aclanthology.org/2025.naacl-demo.44/.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 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*, pp. 1647–1661, 2018. doi: 10.18653/v1/N18-1149. URL https://aclanthology.org/N18-1149/.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. OCR-free document understanding transformer. In *Computer Vision – ECCV 2022*, pp. 498–517, 2022. doi: 10.1007/978-3-031-19815-1_29. URL https://arxiv.org/abs/2111.15664.

Tomasz Korbak, Hady Elsahar, Germán Kruszewski, and Marc Dymetman. On reinforcement learning and distribution matching for fine-tuning language models with no catastrophic forgetting. In *Advances in Neural Information Processing Systems*, volume 35, pp. 16203–16220, 2022. doi: 10.52202/068431-1179. URL https://doi.org/10.52202/068431-1179.

Kenton Lee, Mandar Joshi, Iulia Turc, Hexiang Hu, Fangyu Liu, Julian Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2Struct: Screenshot parsing as pretraining for visual language understanding. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 18893–18912, 2023. URL https://proceedings.mlr.press/v202/lee23g.html.

Qwen Team. Qwen3.5-9B model card. Hugging Face model card, 2026. URL https://huggingface.co/Qwen/Qwen3.5-9B. Accessed 2026-09-25.

Ayako Sato, Hwichan Kim, Zhousi Chen, Masato Mita, and Mamoru Komachi. Exploring the effects of alignment on numerical bias in large language models. *arXiv preprint arXiv:2601.16444*, 2026. URL https://arxiv.org/abs/2601.16444v2. Version 2; accepted at AIBSD 2026, a workshop at AAAI 2026.

Janvijay Singh, Austin Xu, Yilun Zhou, Yefan Zhou, Dilek Hakkani-Tür, and Shafiq Joty. On the shelf life of fine-tuned llm-judges: Future-proofing, backward-compatibility, and question generalization. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2509.23542v2.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. Large language models are inconsistent and biased evaluators. *arXiv preprint arXiv:2405.01724*, 2024. URL https://arxiv.org/abs/2405.01724.

Sihang Zeng, Kai Tian, Kaiyan Zhang, Yuru Wang, Junqi Gao, Runze Liu, Sa Yang, Jingxuan Li, Xinwei Long, Jiaheng Ma, Biqing Qi, and Bowen Zhou. Reviewrl: Towards automated scientific review with rl. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 16931–16943, 2025. doi: 10.18653/v1/2025.emnlp-main.857. URL https://aclanthology.org/2025.emnlp-main.857/.

Pujun Zheng, Jiacheng Yao, Jinquan Zheng, Chenyang Gu, Guoxiu He, Jiawei Liu, Yong Huang, Tianrui Guo, and Wei Lu. From isolated scoring to collaborative ranking: A comparison-native framework for llm-based paper evaluation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 23863–23883, 2026. doi: 10.18653/v1/2026.findings-acl.1195. URL https://aclanthology.org/2026.findings-acl.1195/.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. 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*, pp. 29330–29355. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1420. URL https://aclanthology.org/2025.acl-long.1420/.

A Dataset Statistics

The domain corpus is used only as a shifted training source. Table 4 reports the source sizes and Rating means needed to interpret the direction of the observed drift. The generated means exceed the human-source mean in every domain, but this descriptive difference is not a causal identification strategy.

Table 4: Training-source statistics. Page counts are measured after PDF rendering; DeepReview text length is reported only for the available text-bearing variant.

Dataset/domain	Train n	Pages mean/median	Text chars mean	Rating mean	Label source
Arxiv-CS	938	18.33 / 17	64.8k	7.09	Generated
Arxiv-Math	949	23.75 / 25	59.1k	7.38	Generated
Arxiv-Bio	941	21.16 / 21	62.3k	6.73	Generated
Arxiv-Finance	929	23.59 / 27	63.3k	6.42	Generated
Arxiv-Physics	939	17.83 / 16	57.2k	7.03	Generated
Arxiv-Materials/Chem.	944	18.14 / 16	54.5k	7.19	Generated
DeepReview PDF-only	40,107	21.88 / 21	39.8k	5.21	Human

B Training Details

Table 5 records the main SFT and controlled-GRPO settings. The four supporting domain checkpoints use the same PDF-only task, but they are not presented as matched reruns. This distinction is retained throughout the paper because it determines which comparisons support a controlled claim.

Table 5: Main SFT and controlled GRPO hyperparameters.

Setting	Image SFT	Controlled domain GRPO
Base model	Qwen3.5-9B	Selected Image SFT
Hardware strategy	FSDP2, 8 GPUs	FSDP2 actor/reference, vLLM rollout
Epochs	1	3
Learning rate	5×10^{-6}	1×10^{-6}
Global/train batch	16	16
Responses per prompt	–	4
Max prompt/sequence	32,768	24,000
Max response	included above	2,048
KL coefficient	–	0.01
Page cap	32	32

The SFT optimizer uses weight decay 0.1, betas $(0.9, 0.95)$, gradient clipping at 1.0, cosine decay, 1% warmup, no-padding batches, gradient checkpointing, and Ulysses sequence parallelism size 4. The controlled GRPO runs validate before training and every 10 steps. Actor and reference workers use parameter and optimizer offloading; rollout uses tensor parallel size 2, chunked prefill, prefix caching, and GPU memory utilization 0.4.