WHEN NEGATIVE ADVANTAGES ARE UNRELIABLE: RISK-CALIBRATED NEGATIVE SUPPRESSION FOR GENERATIVE RECOMMENDATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In generative recommendation, reinforcement learning (RL) optimizes recommendation policies using rewards derived from user feedback. Methods such as GRPO compute advantages from the relative rewards of multiple recommendations generated for the same context and suppress those with negative advantages. However, the user feedback available during training is incomplete: recommendations unsupported by observed feedback may still match user interests, so suppressing them risks discouraging otherwise plausible recommendations. To investigate this problem, we measure the fraction of negative-advantage recommendations whose advantages become positive when evaluated against subsequent feedback. Across three recommendation domains, 5.2%–13.1% of recommendations with negative observed advantages receive positive advantages under validation feedback, demonstrating the recurrence of these sign conflicts. This motivates selectively attenuating feedback-sensitive negative updates while retaining useful negative learning signals. The challenge is to determine which updates to attenuate using information available during training. We find that sign conflicts are more likely among negative-advantage recommendations with better observed retrieval ranks, providing an observable cue available during training. Building on this insight, we propose Risk-Calibrated Negative Suppression (RCNS), which reuses this ordering to allocate continuous attenuation without reversing advantage signs. Using held-out validation feedback, we calibrate a shared domain-level attenuation parameter to maximize coverage of sign-conflicting advantage mass subject to an empirical risk budget on attenuating non-conflicting negative mass. The calibrated parameter remains fixed during policy training and requires no subsequent feedback for newly generated recommendations. Offline experiments show that RCNS achieves greater held-out advantage-weighted coverage of sign-conflicting updates than uniform attenuation under the same validation risk budget. In the main downstream comparison, RCNS improves mean HR@5 across five independent seeds over GRPO by 5.0%–19.7% across three domains.

1 INTRODUCTION

Recommender systems seek to identify items that match users' latent interests from large catalogs, using their observed behavior and context (Hidasi et al., 2016; Cai et al., 2023). Generative recommenders formulate this task as conditional sequence generation: models produce item identifiers or textual recommendations from user context, integrating user behavior and item semantics within a unified generation framework (Geng et al., 2023; Rajput et al., 2023; Zheng et al., 2024; Wang et al., 2024a). Yet learning to reproduce historical interactions does not directly optimize recommendation quality. To better align generation with recommendation objectives, recent work introduces reinforcement learning (RL), evaluating generated outputs with recommendation-specific rewards and using these rewards to optimize the policy (Lin et al., 2026; Kong et al., 2025).

Group Relative Policy Optimization (GRPO; Shao et al., 2024) provides a direct mechanism for such reward-driven training. For each user context, it compares the rewards of a group of generated recommendations and computes their advantages from within-group relative performance. In advantage-based policy optimization, positive advantages favor reinforcing the corresponding recom-

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.

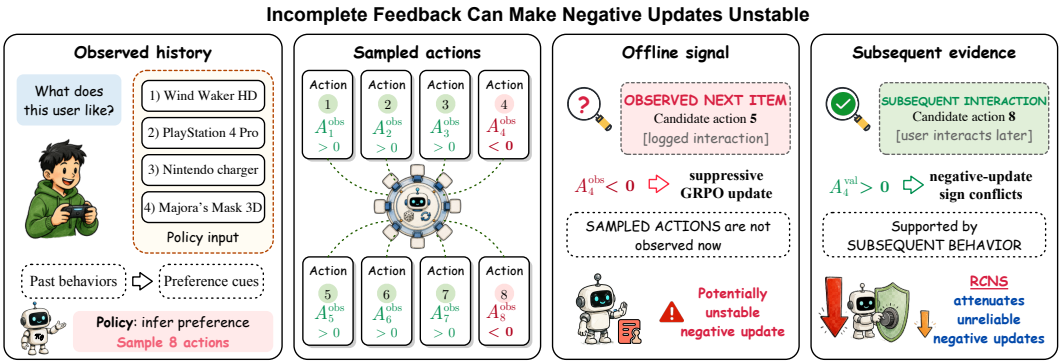


Figure 1: Feedback-sensitive negative updates under incomplete feedback. Fixed recommendations and rollout groups can receive opposing advantage signs under observed and subsequent feedback.

mendations, whereas negative advantages favor suppressing them (Schulman et al., 2017). Rewards therefore do more than evaluate recommendation quality: they determine which recommendations the policy should generate less often. For such suppression to support learning user interests, the observed rewards must provide sufficiently reliable evidence of relative recommendation quality.

Recommendation feedback, however, cannot fully provide this evidence. Interaction logs record only the choices users make under an existing exposure process. An unobserved interaction may indicate disinterest, but may also reflect a lack of opportunity for exposure or engagement (Liang et al., 2016; Schnabel et al., 2016). Even observed feedback can be noisy and cannot be treated as complete, accurate preference labels (Li et al., 2024). A plausible recommendation may therefore receive a below-average reward within its group and a negative advantage because it receives limited support from the available feedback. Incomplete observations can thus translate into suppression of potentially relevant recommendations. This motivates our study of *negative-update reliability*: does the evidence for suppressing a recommendation continue to hold across feedback sets?

To investigate this question, we hold recommendations sampled from a generative recommender and their rollout-group assignments fixed, evaluate them separately against observed and subsequent held-out feedback, and recompute their advantages, as illustrated in Figure 1. When an advantage changes from negative to positive, a suppressive signal becomes reinforcing. We call this a *sign conflict* and use it to diagnose the sensitivity of negative updates to the available feedback. Across three recommendation domains, 5.2%–13.1% of recommendations with negative observed advantages exhibit sign conflicts under validation feedback. Similar conflicts also appear across different rollout backbones, indicating that this instability recurs across domains and models.

This finding motivates reconsidering how negative learning signals are used, with feedback-sensitive negative updates selectively attenuated while useful suppression is retained. The central challenge is not only how much suppression to reduce overall, but how to allocate that reduction among recommendations. However, identifying sign conflicts requires another feedback set, and policy training cannot directly obtain subsequent feedback for newly generated recommendations. Turning the offline diagnostic into a training-time update rule therefore requires a cue that is both associated with negative-update stability and available when the update is computed.

Our further analysis identifies such a cue in the observed retrieval ranking used to compute rewards. Among negative-advantage recommendations, those with better retrieval ranks exhibit more sign conflicts and a larger conflicting share of absolute advantage mass. This is compatible with their negative advantages: group-relative advantages describe performance within each rollout group, whereas this ordering also supports cross-group comparisons. In our retrieval-based setting, a recommendation can fall below its own group's mean reward while still ranking better than recommendations in other groups. Moreover, the association between ordering and sign conflicts persists in pooled comparisons after approximately matching observed advantage magnitudes, suggesting that proximity to zero alone does not explain the observed pattern. These results suggest that retrieval rank helps distinguish recommendations with the same negative advantage sign but different feedback sensitivity, providing a training-time basis for selective attenuation.

Building on this finding, we propose *Risk-Calibrated Negative Suppression*, a reliability-aware policy optimization method. RCNS connects the selection of attenuation targets with attenuation strength by using the observed retrieval ordering to determine which negative updates should receive stronger attenuation and a validation risk budget to control how much attenuation is applied. Specifically, it compares negative-advantage recommendations across rollout groups within a batch, assigning higher priorities to better-ranked recommendations and applying continuous attenuation through a bounded function that is monotone in priority. The rule only reduces negative-advantage magnitudes, without reversing signs or changing non-negative advantages.

RCNS calibrates attenuation strength using validation feedback to identify sign conflicts and an empirical risk budget to constrain attenuation of non-conflicting negative-advantage mass. Subject to this budget, it selects one shared parameter per domain to maximize coverage of sign-conflicting advantage mass. The calibrated parameter is then fixed during policy training, which requires only the current observed advantages and retrieval ranks, not subsequent feedback for newly generated recommendations.

We evaluate RCNS at two levels: offline attenuation allocation and downstream recommendation quality. Under the same validation risk budget, RCNS covers more sign-conflicting advantage mass on held-out data than uniform attenuation. In the main Qwen3-1.7B experiments across three domains, RCNS improves mean HR@5 across five independent seeds by 5.0%–19.7% relative to GRPO. Together, these results support incorporating negative-update reliability into policy optimization under incomplete feedback.

Our main contributions are:

- We identify a negative-update reliability problem in generative recommendation by fixing recommendations and rollout groups, revealing recurring advantage sign conflicts across domains and rollout backbones under incomplete feedback.
- We uncover a systematic relationship between negative-update reliability and the retrieval ordering underlying reward computation: better-ranked negative-advantage recommendations are more likely to exhibit sign conflicts, providing an observable training-time cue for selective attenuation.
- We propose RCNS, a risk-calibrated policy optimization method that combines ranking-based attenuation priorities with continuous non-reversing suppression and validation-based risk control, and evaluate its effectiveness through both offline allocation and downstream training experiments.

2 Related Work

Generative recommendation and policy optimization. Language models support recommendation through text-to-text modeling, instruction tuning, prompted ranking, and collaborative-information integration (Geng et al., 2023; Bao et al., 2023; Hou et al., 2024; Zhang et al., 2025a). More recent generative retrieval methods use discrete item identifiers and learned tokenization to integrate semantic and behavioral information into the generation process (Rajput et al., 2023; Zheng et al., 2024; Wang et al., 2024b;a; Qu et al., 2025; Liu et al., 2025). Policy optimization in recommendation includes off-policy learning and long-term slate optimization (Chen et al., 2019; Ie et al., 2019), while recent generative recommenders increasingly use preference alignment and RL post-training (Zhou et al., 2025; Lin et al., 2026; Kong et al., 2025). In general-purpose language-model optimization, GRPO computes group-relative advantages without a separate value model (Shao et al., 2024). Hysteretic Policy Optimization (HPO) reduces the influence of negative advantages through a fixed or batch-adaptive shared coefficient and modified normalization (Sana et al., 2026). RCNS instead uses ranking-based priorities for recommendation-level allocation and a validation risk budget for calibration, retaining GRPO's advantage computation and token aggregation.

Learning from incomplete and noisy feedback. Confidence-weighted and exposure-aware methods address missing recommendation feedback (Hu et al., 2008; Liang et al., 2016), while propensity-based and doubly robust approaches address selection bias (Schnabel et al., 2016; Wang et al., 2019; Saito et al., 2020), with extensions jointly modeling nonrandom missingness and outcome noise (Li et al., 2024). Other work reduces unreliable supervision through loss filtering, memorization-guided

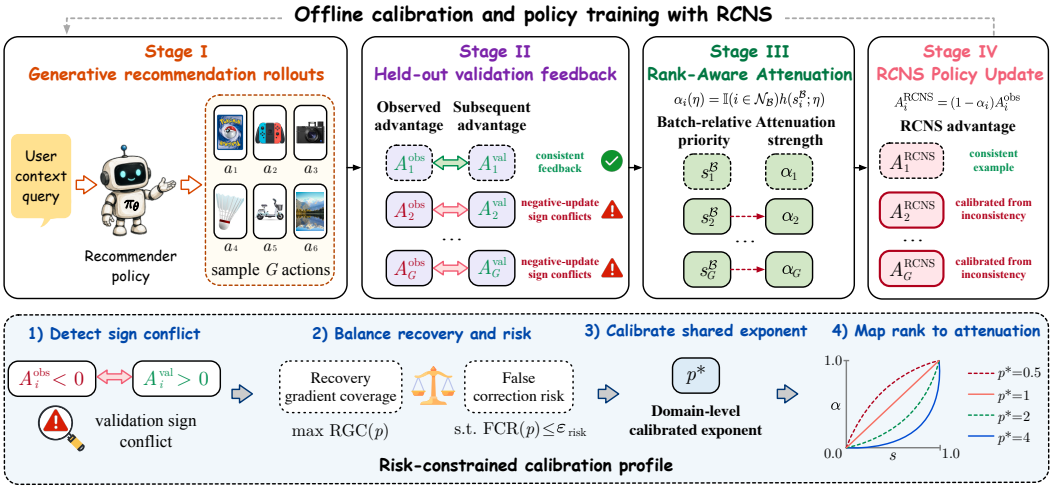


Figure 2: RCNS with the default power profile. Validation feedback on fixed rollouts calibrates a domain-level exponent; training applies it to observed ranks and advantages of new rollouts.

learning, and meta-guided update selection (Wang et al., 2021; Gao et al., 2022; Lin et al., 2024), while robust and disentangled negative sampling reduce the risk of sampling latent positives as negatives (Ding et al., 2020; Lai et al., 2023). RCNS targets update-level reliability: a generated recommendation may receive a negative group-relative advantage without an explicit negative-item label. We diagnose this instability through negative-to-positive advantage conflicts across feedback sets and calibrate continuous, non-reversing attenuation under a validation risk budget.

3 Problem Formulation

Reward-driven policy optimization under incomplete feedback. Given a user context x, a recommendation policy π_θ generates a rollout group $\mathcal{G}_x = \{a_i\}_{i=1}^{G}$, where each action a_i is a generated recommendation. Rewards are computed from available user feedback, and GRPO converts them into group-relative advantages. In our setting, rewards are derived from retrieval ranks. Let $Q(a_i; \mathcal{F})$ denote the best (smallest) rank of a target item in feedback set $\mathcal{F}$ under the retrieval ranking induced by a_i. The observed rank is $q_i^{\text{obs}} = Q(a_i; \mathcal{F}_x^{\text{obs}})$, the scalar reward returned by the reward implementation is r_i^{obs}, and the advantage is $A_i^{\text{obs}} = (r_i^{\text{obs}} - \mu_x^{\text{obs}})/(\sigma_x^{\text{obs}} + \delta)$. Here, μ_x^{obs} and σ_x^{obs} are the group reward mean and sample standard deviation, and $\delta > 0$ ensures numerical stability. The reward function f decreases over its active rank range and is zero beyond a cutoff; misses and ranks beyond the cutoff therefore share the same reward, as specified in Appendix B.

A negative advantage indicates a reward below the group mean, not necessarily a negative user preference. In the GRPO surrogate, positive advantages favor reinforcing generated recommendations, whereas negative advantages favor suppressing them. Under incomplete feedback, a plausible recommendation may receive a low reward because available observations do not cover the relevant user interest. We therefore examine whether its suppressive learning signal remains consistent when the feedback used for evaluation changes.

Negative-update reliability and its observable diagnostic. We assess negative-update reliability by holding generated recommendations and rollout-group membership fixed while replacing the evaluation feedback. For subsequent feedback $\mathcal{F}_x^{\text{sub}}$, we obtain the scalar reward r_i^{sub} from the same reward implementation and recompute the group-normalized advantage A_i^{sub}.

Specifically, we define a negative-to-positive sign conflict by $y_i^{\text{SC}} = \mathbb{I}(A_i^{\text{obs}} < 0 \wedge A_i^{\text{sub}} > 0)$. A conflict means that the same recommendation receives a suppressive signal under observed feedback and a reinforcing signal under subsequent feedback. Sign conflicts thus measure cross-feedback instability in negative-update signals. This motivates a policy update rule that can selectively reduce unreliable negative suppression while preserving useful negative learning signals.

4 RISK-CALIBRATED NEGATIVE SUPPRESSION

As illustrated in Figure 2, RCNS combines ranking-based priorities, non-reversing attenuation that respects those priorities, and empirical validation-risk calibration. Together, they determine which negative updates receive greater attenuation and how extensively to attenuate them. The calibrated rule is then integrated into GRPO.

4.1 BATCH-RELATIVE ATTENUATION PRIORITY

RCNS reuses the ordering underlying reward computation to compare negative updates across rollout groups. GRPO centers and scales rewards independently for each context, so its advantages describe within-group standing and need not preserve this cross-group ordering. The association between this ordering and sign conflicts in RQ2 motivates using it to prioritize attenuation.

For a rollout batch $\mathcal{B}$ containing complete groups, let $\mathcal{N}_{\mathcal{B}} = \{i \in \mathcal{B} : A_i^{\text{obs}} < 0,\ q_i^{\text{obs}} \text{ defined}\}$ and $n_{\mathcal{B}} = |\mathcal{N}_{\mathcal{B}}|$. We order these recommendations by ascending complete-catalog retrieval rank, using average positions for ties. For $n_{\mathcal{B}} > 1$, recommendation $i \in \mathcal{N}_{\mathcal{B}}$ receives priority $s_i^{\mathcal{B}} = 1 - [\text{rank}_{\mathcal{B}}(q_i^{\text{obs}}) - 1]/(n_{\mathcal{B}} - 1)$; recommendations outside $\mathcal{N}_{\mathcal{B}}$ receive priority zero. Offline diagnostics use the same reward computation as policy training; rank conventions are specified in Appendix B. This reverse percentile lies in $[0, 1]$ and assigns greater priority to better-ranked negative-advantage recommendations. This priority construction retains ordinal distinctions that scalar reward truncation can erase; it introduces no additional source of behavioral supervision.

4.2 CONTINUOUS NON-REVERSING ATTENUATION

Attenuation should respect priority while preserving useful negative learning signals. For a chosen continuous family $h(s; \eta)$ with scalar parameter η in a compact set $\mathcal{H}$, we require $0 \le h(s; \eta) \le 1$ and monotonicity in priority s: higher-priority negatives receive no weaker attenuation. Define

$$\alpha_i(\eta) = \mathbb{I}(i \in \mathcal{N}_{\mathcal{B}}) h(s_i^{\mathcal{B}}; \eta), \qquad A_i^{\text{RCNS}}(\eta) = [1 - \alpha_i(\eta)] A_i^{\text{obs}}. \tag{1}$$

Here, α_i denotes the attenuated fraction and $1 - \alpha_i$ the retained advantage weight. Bounded attenuation keeps negative advantages non-positive and leaves non-negative advantages unchanged. In the offline comparisons, recommendations without a valid extracted title have no diagnostic rank: every allocation rule assigns them $\alpha_i = 0$ and excludes them from priority construction, while retaining their advantages in the metric denominators. Calibration selects the profile parameter.

4.3 VALIDATION-CALIBRATED EMPIRICAL RISK

Calibration balances coverage of sign conflicts against attenuation of the remaining negative advantages. We form calibration batches from complete rollout groups and fix their assignment throughout parameter selection. Conflict labels use held-out validation feedback; test feedback does not enter calibration. Across these batches, let $\mathcal{N} = \{i : A_i^{\text{obs}} < 0\}$, with $y_i^{\text{SC}} = \mathbb{I}(A_i^{\text{sub}} > 0)$ and $w_i = |A_i^{\text{obs}}|$ for $i \in \mathcal{N}$. These weights quantify advantage mass, which serves as a proxy for update magnitude.

Recovery Gradient Coverage (RGC) is the fraction of sign-conflicting advantage mass attenuated; False Correction Risk (FCR) is the fraction of non-conflicting negative-advantage mass attenuated. For nonzero denominators,

$$\text{RGC}(\eta) = \frac{\sum_{i \in \mathcal{N}} \alpha_i(\eta) w_i y_i^{\text{SC}}}{\sum_{i \in \mathcal{N}} w_i y_i^{\text{SC}}}, \qquad \text{FCR}(\eta) = \frac{\sum_{i \in \mathcal{N}} \alpha_i(\eta) w_i (1 - y_i^{\text{SC}})}{\sum_{i \in \mathcal{N}} w_i (1 - y_i^{\text{SC}})}. \tag{2}$$

Here, correction means attenuation. Non-conflicts include both $A_i^{\text{sub}} < 0$ and $A_i^{\text{sub}} = 0$; FCR constrains the attenuated fraction of their total negative-advantage mass.

Given a validation risk budget ϵ_{risk} (abbreviated as ϵ in the results), we select a parameter within the chosen family:

$$\eta^\star \in \arg\max_{\eta \in \mathcal{H}} \text{RGC}_{\text{val}}(\eta) \quad \text{subject to} \quad \text{FCR}_{\text{val}}(\eta) \le \epsilon_{\text{risk}}. \tag{3}$$

If no parameter is feasible, we apply no attenuation and retain the original GRPO advantages. The budget constrains empirical FCR on the fixed validation rollouts.

4.4 INTEGRATION WITH GRPO

After calibration, $\eta^\star$ remains fixed throughout policy training. For each new rollout batch, RCNS applies the priority construction in Section 4.1 before splitting the batch into PPO minibatches. These priorities and attenuated advantages are reused during minibatch optimization. We substitute $A_i^{\mathrm{RCNS}}(\eta^\star)$ into the clipped GRPO policy surrogate:

$$\mathcal{L}_{\mathrm{RCNS}} = -\mathbb{E}_{i,t}\left[\min\left(\rho_{i,t}A_i^{\mathrm{RCNS}}, \mathrm{clip}(\rho_{i,t}, 1-\epsilon_{\mathrm{clip}}, 1+\epsilon_{\mathrm{clip}})A_i^{\mathrm{RCNS}}\right)\right], \tag{4}$$

where $\rho_{i,t} = \pi_\theta(a_{i,t} \mid x, a_{i,<t}) / \pi_{\theta_{\mathrm{old}}}(a_{i,t} \mid x, a_{i,<t})$ is the token-level importance ratio and ϵ_{clip} is the PPO clipping threshold, distinct from the calibration budget. Sequence masks, token aggregation, reference-policy regularization, and all other GRPO components remain unchanged.

Implementation details. Unless otherwise stated, RCNS uses the power profile $h(s;p) = s^p$, with $p \in [p_{\min}, p_{\max}]$ and $0 < p_{\min} < p_{\max}$, instantiating η as a single domain-level exponent. For fixed calibration inputs, increasing p concentrates attenuation on high-priority negatives and makes RGC and FCR non-increasing. We select the smallest feasible exponent; Appendix D gives the derivation and search details, and Appendix D.2 compares alternative profiles.

5 EXPERIMENTS

Our experiments follow the motivation for RCNS and address five questions: how prevalent sign conflicts are, whether they exhibit a usable training-time allocation signal, whether RCNS improves attenuation allocation under a fixed risk budget, whether these gains translate into better downstream recommendation performance, and how RCNS behaves under alternative allocation rules and calibration budgets.

5.1 EXPERIMENTAL SETUP

We evaluate three Amazon Reviews'23 domains (Hou et al., 2026): CDs and Vinyl (C&V), Video Games (VG), and Movies and TV (M&TV), using chronological training, validation, and test splits. Our main policy is Qwen3-1.7B (Yang et al., 2025), with Qwen3-Embedding-0.6B (Zhang et al., 2025b) as the retrieval encoder. Training feedback defines observed rewards; validation calibrates $p^\star$; test feedback evaluates reliability transfer and recommendation quality.

GRPO is the primary reference, alongside sequential and generative recommenders and the fixed and adaptive HPO variants. Offline diagnostics hold recommendations and their groups ($G = 8$ per context) fixed across feedback evaluations; allocation experiments report RGC and FCR. Downstream results report HR and NDCG at cutoffs 5 and 10 as means and sample standard deviations across five independent seeds. Appendices A and B provide data, baseline, and implementation details.

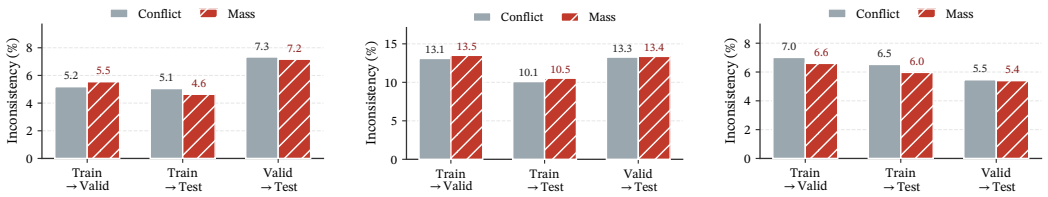
5.2 RQ1: HOW PREVALENT ARE NEGATIVE-UPDATE SIGN CONFLICTS?

Figure 3 shows that sign conflicts recur across all three domains and both held-out feedback windows. We evaluate recommendations with negative observed advantages separately against validation and test feedback, reporting their sign-conflict rate and conflicting share of absolute advantage mass. An auxiliary comparison evaluates recommendations with negative validation advantages against test feedback. Each comparison retains complete rollout groups with the required feedback coverage.

The comparable recommendation-level and mass-based rates show that conflicts involve appreciable advantage mass, rather than only near-zero advantages. The cross-backbone analysis in Appendix F also shows this phenomenon for Qwen3-1.7B, Llama-3.2-3B-Instruct, and Qwen3-4B in every domain under a shared offline evaluation protocol. This recurrence motivates selective attenuation and raises the next question: can training-time evaluation signals distinguish negative updates more likely to conflict?

5.3 RQ2: DOES RANKING INFORMATION REVEAL RELIABILITY STRUCTURE?

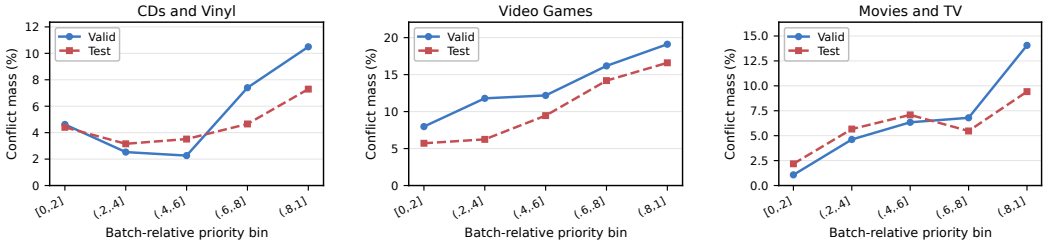
Better-ranked negative-advantage recommendations have a larger conflicting fraction of their advantage mass. Figure 4 partitions recommendations with negative observed advantages into five intervals



(a) CDs and Vinyl (b) Video Games (c) Movies and TV

Figure 3: Negative-to-positive conflicts across feedback windows. Conflict rate and mass are normalized by the count and total absolute advantage, respectively, of recommendations with negative advantages under the reference feedback.

of batch-relative priority $s_i^{\mathcal{B}}$. The highest-priority interval has a larger conflicting fraction than the lowest in every domain under both validation and test feedback.



(a) CDs and Vinyl (b) Video Games (c) Movies and TV

Figure 4: Conflicting fraction of negative observed advantage mass within each batch-relative priority bin, evaluated separately with validation and test feedback.

This association could reflect greater proximity to zero, which makes an advantage easier to reverse. Within each diagnostic batch, we pair negative-advantage recommendations from the top and bottom priority quartiles one-to-one on $|A_i^{\text{obs}}|$, with a maximum difference of 0.02.

As shown in Figure 5, the pooled priority contrast remains positive after matching: high-priority recommendations have a 9.1 percentage-point higher test conflict rate (95% CI: $[4.5, 14.7]$). The estimated difference is positive in every domain. The pooled difference also remains positive when considering the continuous subsequent-feedback advantage shift rather than only a binary sign change. These pooled comparisons support this ordering as an allocation cue beyond proximity to zero. Appendix E reports domain estimates, matching balance, and sensitivity analyses.

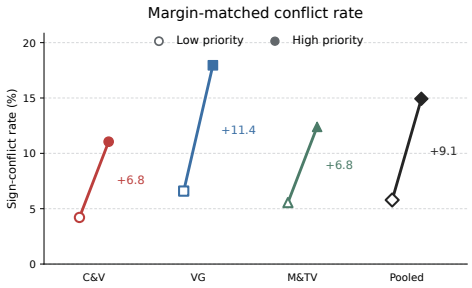
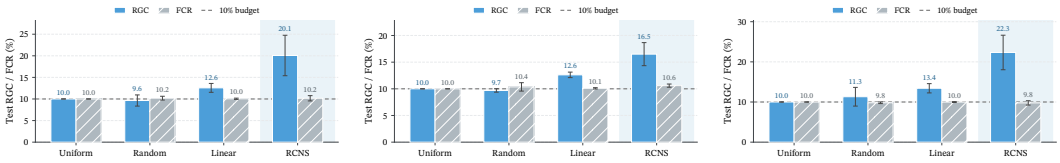


Figure 5: Test conflict rates after within-batch matching of top/bottom priority quartiles on $|A^{\text{obs}}|$ (caliper 0.02).

5.4 RQ3: Does RCNS Improve Allocation under a Shared Risk Budget?

RCNS increases test RGC over uniform attenuation in all three domains, as shown in Figure 6. We compare it with uniform, random, and batch-relative linear attenuation under a shared 10% validation risk budget. All rules are calibrated on validation feedback and evaluated on test feedback.

Uniform attenuation calibrates a constant, while the linear control calibrates global strength. The random control calibrates concentration and global strength. RCNS calibrates the exponent of its default batch-relative power profile.



(a) CDs and Vinyl (b) Video Games (c) Movies and TV

Figure 6: Offline allocation comparison at a 10% validation risk budget. Bars show mean test RGC and FCR across five independent rollout seeds; error bars indicate sample standard deviations. The dashed line marks the validation risk budget.

RCNS achieves greater mean recovery coverage than uniform and batch-relative linear attenuation in every domain, with mean test FCR near the validation budget. These results support ranking-based selective attenuation under a shared validation risk constraint.

5.5 RQ4: Does Reliability-Aware Attenuation Improve Policy Optimization?

We next examine whether selective attenuation translates into better recommendation quality after policy training. Table 1 reports GRPO and RCNS alongside sequential and generative recommenders and the two HPO variants.

Table 1: Recommendation performance ($\times 10^3$; mean $\pm$ sample SD across five independent seeds). GRPO and RCNS use Qwen3-1.7B. Relative gains are computed before display rounding. Bold and underlined values indicate the best and second-best results for each dataset and metric.

Dataset	Metric	Sequential Recommenders			Generative Recommenders		Policy Optimization			
		GRU4Rec	Caser	SASRec	TIGER	LETTER	GRPO	HPO Fixed	HPO Adaptive	RCNS
C&V	HR@5	2.70±0.22	0.90±0.07	0.19±0.06	1.64±0.24	1.20±0.17	<u>2.80</u>±0.96	2.55±0.27	2.28±0.39	**3.06**±1.17
	NDCG@5	<u>2.01</u>±0.16	0.57±0.06	0.15±0.04	1.05±0.23	0.68±0.14	1.87±0.58	1.73±0.19	1.55±0.21	**2.06**±0.74
	HR@10	3.45±0.17	1.76±0.23	0.42±0.07	3.04±0.22	2.12±0.16	<u>4.64</u>±1.27	4.03±0.37	3.73±0.44	**5.03**±1.38
	NDCG@10	2.25±0.13	0.84±0.04	0.22±0.01	1.49±0.15	0.98±0.12	<u>2.46</u>±0.68	2.21±0.22	2.01±0.21	**2.70**±0.81
VG	HR@5	6.12±0.38	5.85±0.71	0.09±0.04	8.56±0.50	6.33±1.15	<u>11.46</u>±0.61	10.00±0.35	9.40±0.45	**12.04**±1.20
	NDCG@5	3.33±0.22	2.99±0.51	0.05±0.01	5.30±0.31	3.98±0.95	<u>7.66</u>±0.47	6.77±0.35	6.48±0.24	**7.94**±0.82
	HR@10	11.68±0.32	11.55±1.46	2.88±0.11	14.22±0.83	10.71±0.77	<u>17.89</u>±0.73	14.83±0.54	14.52±0.46	**18.30**±1.74
	NDCG@10	5.12±0.18	4.81±0.73	0.97±0.03	7.12±0.32	5.38±0.84	<u>9.72</u>±0.50	8.32±0.37	8.14±0.18	**9.94**±0.98
M&TV	HR@5	4.12±0.89	1.66±0.13	0.70±0.39	3.26±0.39	2.08±0.35	8.30±0.69	7.28±1.01	<u>8.83</u>±2.05	**9.93**±2.04
	NDCG@5	3.16±0.86	1.04±0.11	0.54±0.36	2.10±0.31	1.33±0.23	5.53±0.64	5.30±0.84	<u>6.47</u>±1.76	**6.61**±1.47
	HR@10	5.46±0.77	3.04±0.38	0.83±0.34	5.77±0.72	3.21±0.51	11.76±1.59	10.87±1.49	<u>12.82</u>±2.98	**14.49**±2.58
	NDCG@10	3.58±0.79	1.48±0.19	0.58±0.35	2.91±0.30	1.70±0.28	6.64±0.91	6.45±0.96	<u>7.74</u>±2.04	**8.06**±1.62

RCNS has higher means across five independent seeds than GRPO on all four reported metrics in every domain, with relative HR@5 gains of 5.0%–19.7%. Appendix G reports paired differences and their uncertainty across the five training seeds. The gains also extend to Qwen3-4B: the results in Appendix C show that RCNS improves the mean of all four metrics in every domain within this backbone. This provides downstream evidence beyond the main policy size, complementing the cross-backbone offline diagnostic in RQ1.

5.6 RQ5: Ablations and Sensitivity Analyses

5.6.1 Allocation Controls

Table 2 compares allocation rules on Movies and TV and CDs and Vinyl at validation risk budgets of 10% and 20%. Uniform attenuation removes targeting, random allocation replaces ranking-based priorities with random priorities, and linear allocation retains the ranking-based priorities without nonlinear concentration.

At the 10% budget, RCNS achieves the highest mean among the allocation variants on all four metrics in both domains, while uniform attenuation underperforms GRPO. At 20%, RCNS leads all four metrics on Movies and TV; on CDs and Vinyl, Linear leads HR@5 and NDCG@5, while RCNS leads HR@10 and NDCG@10.

Table 2: Downstream allocation ablations ($\times 10^3$; mean $\pm$ sample SD across five independent seeds). Bold and underlined values indicate the best and second-best allocation variants for each dataset, metric, and budget.

			$\epsilon = 0.10$				$\epsilon = 0.20$			
Dataset	Metric	GRPO	Uniform	Random	Linear	RCNS	Uniform	Random	Linear	RCNS
M&TV	HR@5	8.30±0.69	6.99±0.48	<u>9.11</u>±0.54	8.45±0.91	**9.93**±2.04	6.99±1.07	8.31±0.39	<u>8.33</u>±0.15	**8.55**±1.49
	NDCG@5	5.53±0.64	4.49±0.23	<u>6.18</u>±0.62	5.70±0.94	**6.61**±1.47	4.59±0.78	<u>5.65</u>±0.33	5.42±0.27	**5.78**±1.12
	HR@10	11.76±1.59	10.03±0.52	<u>14.10</u>±0.87	13.30±1.31	**14.49**±2.58	11.49±1.41	12.58±0.58	<u>13.02</u>±0.43	**13.13**±2.36
	NDCG@10	6.64±0.91	5.46±0.25	<u>7.77</u>±0.74	7.25±1.07	**8.06**±1.62	6.02±0.87	<u>7.00</u>±0.40	6.92±0.28	**7.23**±1.40
C&V	HR@5	2.80±0.96	2.18±0.33	<u>2.81</u>±0.61	2.55±0.31	**3.06**±1.17	2.51±0.14	2.64±0.09	**2.70**±0.22	<u>2.68</u>±0.68
	NDCG@5	1.87±0.58	1.45±0.24	<u>1.84</u>±0.41	1.74±0.27	**2.06**±0.74	1.75±0.11	1.76±0.10	**1.84**±0.16	<u>1.79</u>±0.46
	HR@10	4.64±1.27	3.76±0.54	<u>4.67</u>±0.86	4.22±0.54	**5.03**±1.38	4.16±0.41	<u>4.54</u>±0.26	4.10±0.28	**4.56**±0.98
	NDCG@10	2.46±0.68	1.97±0.30	<u>2.45</u>±0.50	2.29±0.30	**2.70**±0.81	2.29±0.18	<u>2.37</u>±0.15	2.30±0.17	**2.40**±0.55

5.6.2 Calibration Transfer and Robustness

In the fixed diagnostic sweep reported in Appendix D, test FCR differs from the main 10% validation risk budget by at most 0.69 percentage points, without test-time parameter search. Across five independent rollout seeds, the sample standard deviation of the calibrated exponent ranges from 0.36 to 0.56, as reported in Appendix D.1. These repetitions capture rollout and calibration variability.

With the pretrained-calibrated parameter held fixed, mean test FCR remains within one percentage point of the 10% budget after both the first and second training passes across all three domains, as reported in Appendix D.3.

Across the broader budget sweep reported in Appendix D, larger budgets select smaller exponents and increase both test recovery and test FCR. This reveals the coverage–risk trade-off on held-out feedback.

5.6.3 Downstream Sensitivity to the Risk Budget

Figure 7 reports results for risk budgets from 5% to 35% in five-percentage-point increments on Movies and TV, with five independent seeds per setting.

Downstream performance varies non-monotonically with the risk budget, with the strongest results concentrated around 5%–10%. The 10% budget performs best on both HR metrics, while 5% gives the highest NDCG@5 and a nearly identical NDCG@10.

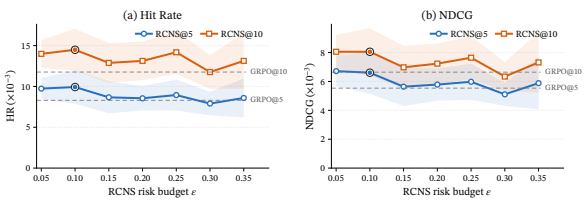


Figure 7: Risk-budget sensitivity on Movies and TV ($\times 10^3$). Curves show five-seed means with ± 1 sample SD bands; dashed lines indicate GRPO.

Together with the offline analysis, these results show that greater recovery coverage does not necessarily translate into better downstream performance. The risk budget therefore balances the attenuation of feedback-sensitive negative updates with the preservation of useful negative learning signals.

6 Conclusion

Under incomplete user feedback, negative group-relative advantages do not always provide stable support for suppressing recommendations. Our diagnostics reveal recurring sign conflicts across domains and show that the ordering underlying reward computation helps distinguish feedback-sensitive negative updates. RCNS reuses this ordering to allocate non-reversing attenuation under an empirical validation risk budget. Offline experiments show greater held-out coverage of sign-conflicting advantage mass than uniform attenuation under the same budget, while downstream experiments yield consistent gains over GRPO across all evaluated domains and metrics. These findings support reliability-aware attenuation as an effective approach to policy optimization for generative recommendation under incomplete feedback.

AI USE STATEMENT

Generative AI assisted experimental-design feedback, implementation and analysis code, diagnostic interpretation, manuscript editing, reference formatting, and figure creation or refinement. The authors checked AI-assisted code and analyses against the data and verified claims, citations, numerical results, and figures against source artifacts. AI did not generate synthetic evaluation data or replace author judgment in result selection or reporting. The authors take responsibility for all paper content, including AI-assisted text, code, and artifacts.

REPRODUCIBILITY STATEMENT

Appendices A and B document data construction, temporal splits, models, optimization, rewards, and downstream evaluation on complete test sets across five independent seeds. Appendices D, E, and F detail calibration, matched analyses with sensitivity checks, and cross-backbone diagnostics.

REFERENCES

Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. TALLRec: An effective and efficient tuning framework to align large language model with recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*, RecSys '23, pp. 1007–1014, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400702419. doi: 10.1145/3604915.3608857. URL https://doi.org/10.1145/3604915.3608857.

Xuheng Cai, Chao Huang, Lianghao Xia, and Xubin Ren. LightGCL: Simple yet effective graph contrastive learning for recommendation, 2023. URL https://arxiv.org/abs/2302.08191.

Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed H. Chi. Top-K off-policy correction for a REINFORCE recommender system. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, pp. 456–464, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359405. doi: 10.1145/3289600.3290999. URL https://doi.org/10.1145/3289600.3290999.

Jingtao Ding, Yuhan Quan, Quanming Yao, Yong Li, and Depeng Jin. Simplify and robustify negative sampling for implicit collaborative filtering. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Yunjun Gao, Yuntao Du, Yujia Hu, Lu Chen, Xinjun Zhu, Ziquan Fang, and Baihua Zheng. Self-guided learning to denoise for robust recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, pp. 1412–1422, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450387323. doi: 10.1145/3477495.3532059. URL https://doi.org/10.1145/3477495.3532059.

Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. Recommendation as language processing (RLP): A unified pretrain, personalized prompt & predict paradigm (P5), 2023. URL https://arxiv.org/abs/2203.13366.

Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. Session-based recommendations with recurrent neural networks, 2016. URL https://arxiv.org/abs/1511.06939.

Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. Large language models are zero-shot rankers for recommender systems. In *Advances in Information Retrieval: 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24–28, 2024, Proceedings, Part II*, pp. 364–381, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-56059-0. doi: 10.1007/978-3-031-56060-6_24. URL https://doi.org/10.1007/978-3-031-56060-6_24.

Yupeng Hou, Jiacheng Li, Xiangjun Fu, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. Bridging language and items for retrieval and recommendation: Benchmarking LLMs as semantic encoders, 2026. URL https://arxiv.org/abs/2403.03952.

Yifan Hu, Yehuda Koren, and Chris Volinsky. Collaborative filtering for implicit feedback datasets. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, ICDM '08, pp. 263–272, USA, 2008. IEEE Computer Society. ISBN 9780769535029. doi: 10.1109/ICDM.2008.22. URL https://doi.org/10.1109/ICDM.2008.22.

Eugene Ie, Vihan Jain, Jing Wang, Sanmit Narvekar, Ritesh Agarwal, Rui Wu, Heng-Tze Cheng, Tushar Chandra, and Craig Boutilier. SLATEQ: a tractable decomposition for reinforcement learning with recommendation sets. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, IJCAI'19, pp. 2592–2599. AAAI Press, 2019. ISBN 9780999241141.

Wang-Cheng Kang and Julian McAuley. Self-attentive sequential recommendation, 2018. URL https://arxiv.org/abs/1808.09781.

Xiaoyu Kong, Leheng Sheng, Junfei Tan, Yuxin Chen, Jiancan Wu, An Zhang, Xiang Wang, and Xiangnan He. MiniOneRec: An open-source framework for scaling generative recommendation, 2025. URL https://arxiv.org/abs/2510.24431.

Riwei Lai, Li Chen, Yuhan Zhao, Rui Chen, and Qilong Han. Disentangled negative sampling for collaborative filtering. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, WSDM '23, pp. 96–104, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394079. doi: 10.1145/3539597.3570419. URL https://doi.org/10.1145/3539597.3570419.

Haoxuan Li, Chunyuan Zheng, Wenjie Wang, Hao Wang, Fuli Feng, and Xiao-Hua Zhou. Debiased recommendation with noisy feedback. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pp. 1576–1586, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704901. doi: 10.1145/3637528.3671915. URL https://doi.org/10.1145/3637528.3671915.

Dawen Liang, Laurent Charlin, James McInerney, and David M. Blei. Modeling user exposure in recommendation. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pp. 951–961, Republic and Canton of Geneva, CHE, 2016. International World Wide Web Conferences Steering Committee. ISBN 9781450341431. doi: 10.1145/2872427.2883090. URL https://doi.org/10.1145/2872427.2883090.

Guanyu Lin, Chen Gao, Yu Zheng, Yinfeng Li, Jianxin Chang, Yanan Niu, Yang Song, Kun Gai, Zhiheng Li, Depeng Jin, and Yong Li. Inverse learning with extremely sparse feedback for recommendation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM '24, pp. 396–404, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703713. doi: 10.1145/3616855.3635797. URL https://doi.org/10.1145/3616855.3635797.

Jiacheng Lin, Tian Wang, and Kun Qian. Rec-R1: Bridging generative large language models and user-centric recommendation systems via reinforcement learning, 2026. URL https://arxiv.org/abs/2503.24289.

Enze Liu, Bowen Zheng, Cheng Ling, Lantao Hu, Han Li, and Wayne Xin Zhao. Generative recommender with end-to-end learnable item tokenization. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, pp. 729–739, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400715921. doi: 10.1145/3726302.3729989. URL https://doi.org/10.1145/3726302.3729989.

Haohao Qu, Wenqi Fan, Zihuai Zhao, and Qing Li. TokenRec: Learning to tokenize ID for LLM-based generative recommendations. *IEEE Transactions on Knowledge and Data Engineering*, 37 (10):6216–6231, 2025. doi: 10.1109/TKDE.2025.3599265.

Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Keshavan, Trung Vu, Lukasz Heidt, Lichan Hong, Yi Tay, Vinh Q. Tran, Jonah Samost, Maciej Kula, Ed H. Chi, and Maheswaran

Sathiamoorthy. Recommender systems with generative retrieval. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Yuta Saito, Suguru Yaginuma, Yuta Nishino, Hayato Sakata, and Kazuhide Nakata. Unbiased recommender learning from missing-not-at-random implicit feedback. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, WSDM '20, pp. 501–509, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368223. doi: 10.1145/3336191.3371783. URL https://doi.org/10.1145/3336191.3371783.

Mohamed Sana, Nicola Piovesan, Antonio De Domenico, Fadhel Ayed, and Haozhe Zhang. HPO: Hysteretic policy optimization for stable and efficient training under sparse-reward regime, 2026. URL https://arxiv.org/abs/2605.30201.

Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: debiasing learning and evaluation. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pp. 1670–1679. JMLR.org, 2016.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Jiaxi Tang and Ke Wang. Personalized top-N sequential recommendation via convolutional sequence embedding. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, pp. 565–573, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355810. doi: 10.1145/3159652.3159656. URL https://doi.org/10.1145/3159652.3159656.

Wenjie Wang, Fuli Feng, Xiangnan He, Liqiang Nie, and Tat-Seng Chua. Denoising implicit feedback for recommendation. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, WSDM '21, pp. 373–381, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450382977. doi: 10.1145/3437963.3441800. URL https://doi.org/10.1145/3437963.3441800.

Wenjie Wang, Honghui Bao, Xinyu Lin, Jizhi Zhang, Yongqi Li, Fuli Feng, See-Kiong Ng, and Tat-Seng Chua. Learnable item tokenization for generative recommendation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, pp. 2400–2409, New York, NY, USA, 2024a. Association for Computing Machinery. ISBN 9798400704369. doi: 10.1145/3627673.3679569. URL https://doi.org/10.1145/3627673.3679569.

Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. Doubly robust joint learning for recommendation on data missing not at random. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6638–6647. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/wang19n.html.

Ye Wang, Jiahao Xun, Minjie Hong, Jieming Zhu, Tao Jin, Wang Lin, Haoyuan Li, Linjun Li, Yan Xia, Zhou Zhao, and Zhenhua Dong. EAGER: Two-stream generative recommender with behavior-semantic collaboration. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pp. 3245–3254, New York, NY, USA, 2024b. Association for Computing Machinery. ISBN 9798400704901. doi: 10.1145/3637528.3671775. URL https://doi.org/10.1145/3637528.3671775.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang,

Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Yang Zhang, Fuli Feng, Jizhi Zhang, Keqin Bao, Qifan Wang, and Xiangnan He. CoLLM: Integrating collaborative embeddings into large language models for recommendation. *IEEE Trans. on Knowl. and Data Eng.*, 37(5):2329–2340, May 2025a. ISSN 1041-4347. doi: 10.1109/TKDE.2025.3540912. URL https://doi.org/10.1109/TKDE.2025.3540912.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models, 2025b. URL https://arxiv.org/abs/2506.05176.

Bowen Zheng, Yupeng Hou, Hongyu Lu, Yu Chen, Wayne Xin Zhao, Ming Chen, and Ji-Rong Wen. Adapting large language models by integrating collaborative semantics for recommendation. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pp. 1435–1448, 2024. doi: 10.1109/ICDE60146.2024.00118.

Guorui Zhou, Jiaxin Deng, Jinghao Zhang, Kuo Cai, Lejian Ren, Qiang Luo, Qianqian Wang, Qigen Hu, Rui Huang, Shiyao Wang, Weifeng Ding, Wuchao Li, Xinchen Luo, Xingmei Wang, Zexuan Cheng, Zixing Zhang, Bin Zhang, Boxuan Wang, Chaoyi Ma, Chengru Song, Chenhui Wang, Di Wang, Dongxue Meng, Fan Yang, Fangyu Zhang, Feng Jiang, Fuxing Zhang, Gang Wang, Guowang Zhang, Han Li, Hengrui Hu, Hezheng Lin, Hongtao Cheng, Hongyang Cao, Huanjie Wang, Jiaming Huang, Jiapeng Chen, Jiaqiang Liu, Jinghui Jia, Kun Gai, Lantao Hu, Liang Zeng, Liao Yu, Qiang Wang, Qidong Zhou, Shengzhe Wang, Shihui He, Shuang Yang, Shujie Yang, Sui Huang, Tao Wu, Tiantian He, Tingting Gao, Wei Yuan, Xiao Liang, Xiaoxiao Xu, Xugang Liu, Yan Wang, Yi Wang, Yiwu Liu, Yue Song, Yufei Zhang, Yunfan Wu, Yunfeng Zhao, and Zhanyu Liu. OneRec technical report, 2025. URL https://arxiv.org/abs/2506.13695.

A Dataset Statistics and Preprocessing

We use Amazon Reviews'23 5-core histories, retain interactions from January 1, 2021, and divide each domain into contiguous chronological training, validation, and test periods. Users have at least five historical interactions; malformed or missing titles are removed, titles are truncated to 200 characters, and inputs retain the 40 most recent items. The resulting prediction records number 19,809/22,974/15,942 for CDs and Vinyl, 14,297/18,139/22,838 for Video Games, and 19,960/15,201/15,693 for Movies and TV (train/validation/test); these counts are distinct from the historical interactions supplied as context. Policy training uses one most-recent training context per eligible user: 8,012, 7,763, and 8,993, respectively. Separate fixed-rollout diagnostic subsets contain 3,710, 3,301, and 3,179 contexts, producing 29,680, 26,408, and 25,432 actions at $G = 8$. Policy training, calibration, and offline diagnostics follow these temporal splits.

Table 3: Dataset statistics after preprocessing.

Dataset	#Users	#Items	#Interactions
CDs and Vinyl	17,012	24,839	58,725
Video Games	20,898	10,533	55,274
Movies and TV	18,284	24,501	50,854

Sensitivity to the subsequent-feedback horizon. We test sensitivity to the amount of subsequent evidence by progressively extending the post-cutoff feedback horizon. For each fraction, we retain the earliest corresponding fraction of chronologically ordered interaction records in the combined validation–test split, including all records tied at the boundary timestamp; validation and test are neither independently nor randomly subsampled. We then recompute subsequent-feedback advantages for the same fixed rollouts and report the fraction of recommendations with negative observed advantages that become positive. This protocol differs from RQ1: RQ1 treats validation and test as separate feedback windows, whereas this analysis uses chronological prefixes of their union. Because future-evidence coverage expands with the prefix, eligibility is recomputed at each horizon, and each rate is conditional on recommendations with negative observed advantages in complete rollout groups with the required feedback coverage.

Table 4: Sign-conflict rates across chronological subsequent-feedback horizons.

Held-out record fraction	C&V	VG	M&TV	Pooled
10%	3.95%	11.30%	6.24%	8.50%
25%	3.73%	12.30%	5.46%	8.85%
50%	5.06%	13.16%	7.02%	10.04%
75%	5.68%	13.39%	7.34%	10.26%
100%	6.07%	13.10%	8.02%	10.29%

Sign conflicts persist across all evaluated horizons, with the pooled rate increasing from 8.50% to 10.29%. These comparisons characterize sensitivity to expanding feedback windows, with feedback coverage and eligible recommendations varying across horizons.

A.1 Feedback-Set Cardinality Controls

Why evaluate a subsequent-feedback window? Our diagnostic examines whether suppression based on observed feedback remains aligned with feedback revealed over a subsequent interval. We therefore evaluate the fixed recommendations against the feedback items in the corresponding validation or test split, rather than only the first subsequent item. A single interaction provides limited evidence about user interests: a recommendation unsupported by the first subsequent interaction may receive support from interactions occurring later in the interval. Using the subsequent-feedback window allows the diagnostic to capture these later-supported recommendations. We complement this window-based evaluation with cardinality controls that distinguish the single-target setting from evaluation against multiple subsequent targets.

Control protocol. The observed feedback contains the final training target. For each user, we chronologically order the distinct items in the relevant held-out split and retain the first k items as

subsequent feedback. We keep the generated recommendations, rollout groups, observed rewards and advantages, and observed-rank priorities unchanged. For each subsequent-feedback set, we apply the same best-target retrieval reward and recompute the group-relative advantages. A conflict occurs when $A_i^{\text{obs}} < 0$ and $A_i^{(k)} > 0$, where $A_i^{(k)}$ denotes the advantage computed from the first k subsequent items. Validation and test feedback are evaluated separately.

Complete single-target evaluation. For $k = 1$, we include every user in the fixed-rollout diagnostic pool with at least one item in the relevant held-out split. This evaluation does not require four subsequent items. Both the observed and subsequent feedback sets therefore contain exactly one target. Table 5 compares this setting with the full feedback window using the same users and observed-negative denominators. Although restricting feedback to one item lowers the measured conflict rates, conflicts remain present in all three domains with both validation and test feedback.

Table 5: Single-target and full-window conflict rates. N^- is the number of observed negative-advantage recommendations. Rates are percentages; each row uses the same users and denominator for both feedback definitions.

Feedback	Domain	Users	N^-	$k = 1$	Full window
Validation	C&V	3,090	2,005	2.59	5.19
	VG	2,576	4,380	8.15	13.11
	M&TV	2,119	1,682	4.04	7.02
	Pooled	7,785	8,067	5.91	9.87
Test	C&V	1,881	1,227	3.26	5.05
	VG	1,691	2,921	5.51	10.10
	M&TV	2,061	1,532	3.85	6.53
	Pooled	5,633	5,680	4.58	8.05

Priority association under single-target feedback. We further test whether the association between observed-rank priority and sign conflict remains under $k = 1$. Within each domain and diagnostic batch, we greedily match recommendations in the top and bottom priority quartiles one-to-one by observed advantage magnitude, using the caliper $\left||A_i^{\text{obs}}| - |A_j^{\text{obs}}|\right| \leq 0.02$. Confidence intervals use 5,000 batch-cluster bootstrap resamples. Among matched recommendations, the pooled high-minus-low conflict-rate difference is 10.34 percentage points with validation feedback and 6.48 percentage points with test feedback; both confidence intervals are above zero (Table 6). Thus, the pooled priority association remains present when both feedback sets contain a single target.

Table 6: Advantage-matched conflict rates (%) under complete $k = 1$ evaluation. Δ is the high-minus-low difference (percentage points), with a 95% batch-cluster bootstrap confidence interval.

Feedback	Domain	Pairs	High	Low	Δ	95% CI
Validation	C&V	329	5.17	1.82	3.34	[0.60, 6.80]
	VG	672	16.52	1.64	14.88	[10.17, 19.78]
	M&TV	275	7.64	0.00	7.64	[3.39, 12.38]
	Pooled	1,276	11.68	1.33	10.34	[7.14, 13.45]
Test	C&V	190	7.89	2.11	5.79	[−0.58, 13.07]
	VG	440	10.68	2.95	7.73	[2.77, 13.90]
	M&TV	234	7.26	2.56	4.70	[0.00, 9.80]
	Pooled	864	9.14	2.66	6.48	[3.15, 10.07]

Raw-reward control under single-target feedback. We further examine whether the priority association extends from advantage signs to recommendation-level raw rewards. We reuse the matched high- and low-priority negative recommendations from the complete $k = 1$ test evaluation and compare their recorded rewards directly. Both observed and subsequent feedback contain one target, and the same observed-advantage matching is retained; the outcome uses raw rewards rather than group-relative advantages.

For each recommendation, we record whether its raw retrieval reward increases under the first chronological target in the test split, $r_i^{(1)} > r_i^{\text{obs}}$. High-priority negatives have a higher probability of raw-reward improvement in all three domains (Table 7). In the pooled sample, this probability is 11.34% for high-priority recommendations and 3.01% for low-priority recommendations, a difference of 8.33 percentage points with a 95% batch-cluster bootstrap confidence interval of $[4.84, 12.57]$. Intervals use 5,000 resamples, stratified by domain for the pooled analysis.

Table 7: Raw-reward improvement under single-target test feedback. High and Low are percentages; Δ is their difference in percentage points with a 95% batch-cluster bootstrap confidence interval.

Domain	Pairs	High	Low	Δ [95% CI]
C&V	190	9.47	2.11	7.37 [0.00, 16.57]
VG	440	12.95	3.64	9.32 [3.76, 16.21]
M&TV	234	9.83	2.56	7.26 [2.28, 12.40]
Pooled	864	11.34	3.01	8.33 [4.84, 12.57]

As a complementary continuous comparison, we compute the raw-reward change, $\Delta r_i = r_i^{(1)} - r_i^{\text{obs}}$. The pooled high-minus-low difference in mean reward change is $+0.003208$, with a 95% batch-cluster bootstrap confidence interval of $[-0.000662, 0.007387]$.

Together, these results extend the priority association beyond advantage-sign comparisons: under single-target feedback and matched observed-advantage magnitudes, high-priority negatives are more likely to receive higher subsequent raw rewards.

Fixed-cohort cardinality comparison. To compare $k \in \{1, 2, 4\}$ without changing user composition, we separately select users with at least four distinct items in the validation or test split. We then evaluate the first one, two, or four items for those same users. Batch assignments and priorities are retained from the corresponding complete-cohort diagnostic, and the matched pairs are fixed across k. This restricted cohort is distinct from the complete $k = 1$ evaluation above.

In this fixed-cohort comparison, conflict rates increase from $k = 1$ to $k = 4$ in all six domain–split combinations (Table 8). The result shows how the measured conflict rate changes as additional chronological feedback items are included, with user composition and observed-negative denominators held fixed. Table 9 reports the pooled matched differences and confidence intervals for validation and test feedback across the three cardinalities.

Table 8: Conflict rates (%) for the same users across $k = 1, 2, 4$. Each user has at least four distinct items in the relevant split; users and observed-negative denominators remain fixed across k.

Feedback	Domain	Users	N^-	$k = 1$	$k = 2$	$k = 4$
Validation	C&V	792	462	2.38	3.25	4.76
	VG	423	792	7.32	15.03	23.23
	M&TV	399	354	5.37	8.47	10.17
	Pooled	1,614	1,608	5.47	10.20	15.05
Test	C&V	392	238	4.20	5.46	6.30
	VG	264	422	7.11	13.51	20.14
	M&TV	388	330	0.91	2.42	7.27
	Pooled	1,044	990	4.34	7.88	12.53

Table 9: Pooled high-minus-low differences (percentage points) and 95% confidence intervals on the fixed cohort. Matched pairs are fixed across k.

Feedback	Pairs	$k = 1$	$k = 2$	$k = 4$
Validation	171	12.28 [4.76, 21.11]	12.87 [3.51, 22.86]	16.96 [6.83, 27.59]
Test	101	4.95 [−1.15, 11.30]	5.94 [−5.21, 16.82]	0.99 [−14.00, 16.35]

Summary. These controls show that sign conflicts and the pooled priority association persist when observed and subsequent feedback each contain one target. The subsequent-feedback window serves

a complementary purpose: it evaluates current suppression against interests expressed throughout the held-out interval, including feedback that arrives after the first interaction. We therefore retain the window-based diagnostic as the main evaluation and use the cardinality controls to characterize its dependence on the available subsequent feedback.

A.2 Sign Conflicts across Reward Functions

Our default reward emphasizes gains at higher retrieval positions to support top-K recommendation. To examine whether sign conflicts persist across reward designs, we compare the default nonlinear cutoff reward with linear cutoff and full-catalog linear rewards. We hold the generated recommendations, rollout groups, and feedback sets fixed across conditions and recompute group-relative advantages for each reward. For full-catalog target rank q, catalog size N, and cutoff $K = 1000$, the default reward is $[\exp(2(1 - \log q/\log K)) - 1]/(\exp(2) - 1)$ for $q \le K$ and zero otherwise. The two alternatives are $\max\{0, (K - q)/(K - 1)\}$ and $(N - q)/(N - 1)$, respectively. Each condition reports the fraction of its observed-negative recommendations whose advantages become positive under validation or test feedback.

Table 10 shows that sign conflicts occur under all three reward functions in all three domains. They persist after removing the nonlinear transformation and after additionally removing the rank cutoff. Thus, cross-feedback instability in negative updates is not specific to the default reward design, further motivating the study of negative-update reliability under incomplete feedback.

Table 10: Sign-conflict rates (%) across reward functions. Values are means ± sample SD across five independent rollout seeds.

Domain	Reward	Validation	Test
C&V	Nonlinear cutoff	5.38 ± 0.51	5.27 ± 0.52
	Linear cutoff	6.20 ± 0.67	5.67 ± 0.82
	Full-catalog linear	49.51 ± 0.47	49.19 ± 0.57
VG	Nonlinear cutoff	12.64 ± 0.42	11.29 ± 0.73
	Linear cutoff	14.34 ± 0.38	12.60 ± 0.95
	Full-catalog linear	39.53 ± 0.50	40.15 ± 0.61
M&TV	Nonlinear cutoff	5.93 ± 0.89	6.70 ± 0.90
	Linear cutoff	6.39 ± 0.79	7.21 ± 0.86
	Full-catalog linear	47.29 ± 0.65	47.74 ± 0.37

B Reproducibility Details

Models and inputs. For the main downstream experiments, policy training is initialized from pretrained Qwen3-1.7B, and the retrieval encoder is Qwen3-Embedding-0.6B. Each prompt contains at most 40 recent item titles and requests one next-item recommendation. Policy training uses a maximum prompt length of 2,048 tokens and a maximum response length of 1,024 tokens. We perform full-parameter training in bfloat16 mixed precision.

Baselines and metrics. GRPO (Shao et al., 2024) is the primary policy-optimization reference. Additional baselines include sequential recommenders GRU4Rec (Hidasi et al., 2016), Caser (Tang & Wang, 2018), and SASRec (Kang & McAuley, 2018); generative recommenders TIGER (Rajput et al., 2023) and LETTER (Wang et al., 2024a); and fixed and adaptive HPO (Sana et al., 2026). RQ3 specifies the offline allocation controls. Offline diagnostics recompute advantages under replacement feedback on fixed recommendations and rollout groups, separating feedback changes from generation changes. RGC, FCR, and attenuated-mass precision weight recommendations by $|A_i^{\mathrm{obs}}|$. Downstream metrics are Hit Rate (HR) and normalized discounted cumulative gain (NDCG) at cutoffs 5 and 10.

Optimization and rollout sampling. GRPO and RCNS are trained for two passes over the policy-training contexts with AdamW, learning rate 8×10^{-7}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 0.01, a constant learning-rate schedule, no warmup, PPO clipping parameter $\epsilon_{\mathrm{clip}} = 0.2$, gradient clipping

at 1.0, KL coefficient 0.001, and entropy coefficient 0. Each context produces $G = 8$ rollouts with temperature 1.0, top-$p = 1$, and no top-k truncation. The global rollout batch contains 128 contexts with eight sampled actions each (1,024 actions). RCNS computes priorities once over the negative actions in this full batch, before it is split into PPO minibatches of 32 contexts. The resulting priorities and attenuated advantages are reused without recomputation within those minibatches. The per-GPU microbatch size is 16, and one PPO epoch is applied per rollout batch. GRPO and RCNS use identical initialization and optimization for each of the five independent seeds (42–46); RCNS changes only the attenuation applied to negative advantages.

HPO baselines. We implement HPO-Fixed and HPO-Adaptive (Sana et al., 2026) within our verl-based training framework. Both variants use group-mean-centered rewards without group-standard-deviation normalization, retain the clipped policy objective, and multiply negative-advantage surrogate terms by a shared coefficient c, leaving nonnegative terms unchanged. HPO-Fixed uses $c = 0.6$. HPO-Adaptive uses $c = \mathrm{clip}(f_+/(f_- + 10^{-8}), 0.4, 1)$, where f_+ and f_- are the fractions of positive- and negative-advantage recommendations among those with nonzero advantages in the PPO minibatch. These statistics are aggregated across data-parallel workers before microbatching. Both variants normalize the summed token-level surrogate by $N\bar{L}$, where N is the number of valid responses and $\bar{L}$ their mean length in that minibatch.

Both HPO variants start from pretrained Qwen3-1.7B and train for two passes over the policy-training contexts. Each context produces eight rollouts with temperature 1.0 and top-$p = 1.0$. The global rollout batch contains 128 contexts, and PPO minibatches contain 32 contexts, with one PPO epoch per rollout batch. We use learning rate 8×10^{-7}, clipping threshold 0.2, and KL coefficient 0.001.

Priority construction. Calibration and policy training use the same priority construction. If a calibration batch contains only one eligible negative recommendation, we assign it zero priority; none of the evaluated diagnostic batches contains such a singleton.

Reward and RCNS calibration. To evaluate a_i, we extract its recommendation title and encode it with Qwen3-Embedding-0.6B without a retrieval instruction. Offline diagnostics and calibration use the same encoder, catalog, and reward computation as policy training, without masking history items. Complete groups of eight recommendations are evaluated together using the training title-encoding implementation with a 128-token limit and bfloat16 encoder inference, float16 similarities, and CUDA float32 advantage normalization with sample standard deviations and a 10^{-6} denominator stabilizer. The reward catalogs contain 89,243, 25,384, and 114,511 items for CDs and Vinyl, Video Games, and Movies and TV, respectively; these include catalog items beyond those appearing in the filtered interaction records in Table 3. Observed feedback contains the final training target; validation and test feedback contain the user's items in their respective held-out splits and are evaluated separately. For reward computation, let $\hat{q}$ denote the target position returned by the reward computation, with $\hat{q} = 3001$ for a miss in the top-3,000 retrieval. For $1 \le \hat{q} \le K$ with $K = 1000$, let $u(\hat{q}) = 1 - \log(\hat{q})/\log(K)$ and $f(\hat{q}) = [\exp(2u(\hat{q})) - 1]/[\exp(2) - 1]$, with $f(\hat{q}) = 0$ for $\hat{q} > K$. For priority construction, we compute the complete observed-target rank q^{obs} as one plus the number of catalog items with strictly greater similarity, using the same title embeddings and catalog. Group advantages and conflict labels are consequently held fixed across allocation comparisons. For the main downstream comparison in Table 1, we adopt a common validation risk budget of $\epsilon_{\mathrm{risk}} = 0.10$ across all three domains, without a search over budgets. For each domain, we calibrate the exponent on validation feedback using rollouts from pretrained Qwen3-1.7B under the data preprocessing in Appendix A, and keep it fixed throughout policy training.

Calibration rollout generation. Calibration and offline diagnostics use recommendations from pretrained Qwen3-1.7B before RL, following the temporal splits in Appendix A. Generation uses eight recommendations per context, temperature 0.8, top-$p = 0.95$, at most 40 historical titles, and a 2,048-token prompt limit. The fixed diagnostic pool uses seed 42 with a 160-token response limit; the five independent pools use seeds 11, 23, 37, 51, and 73 with a 128-token limit.

RQ3 offline calibration. The RQ3 comparison is repeated across five independent seeds. For each feedback split, we retain complete rollout groups whose users have feedback in that split before forming diagnostic batches. Within each run, recommendations and rollout groups are held fixed across allocation methods and feedback evaluations, with eight actions per group and 128 groups per

diagnostic batch. For RCNS in each run, we select the smallest feasible exponent in $[0.01, 64]$ using 80 bisection steps under $\mathrm{FCR}_{\mathrm{val}} \leq 0.10$. Test feedback is used only for evaluation after calibration. We report arithmetic means and sample standard deviations across the five runs.

Evaluation and random seeds. Final recommendation metrics use the complete held-out test sets listed in Appendix A. Policy inference is deterministic: Qwen3-Embedding-0.6B embeds the generated title for full-catalog ranking, excluding items listed in each evaluation record's history.

Each downstream configuration uses five independent seeds (42–46). All four recommendation metrics report the arithmetic mean and sample standard deviation of the same five runs, with standard deviations computed using denominator $n - 1$. Relative gains compare these means before display rounding. This convention applies to Tables 1 and 2, the Qwen3-4B comparison, and the budget-sensitivity curves; the latter show one-standard-deviation bands.

C Additional Backbone Evaluation: Qwen3-4B

We test whether downstream gains extend to Qwen3-4B by comparing GRPO and RCNS within this backbone. The comparison follows the data splits, reward, rollout-group construction, optimization budget, and full-catalog evaluation protocol in Appendix B, with five independent seeds per method.

Table 11: Qwen3-4B results ($\times 10^3$; mean $\pm$ sample SD across five independent seeds). Relative gains use means before table rounding.

Dataset	Method	HR@5	HR@10	NDCG@5	NDCG@10
CDs and Vinyl	GRPO	$2.55_{\pm 0.59}$	$4.28_{\pm 0.79}$	$1.66_{\pm 0.41}$	$2.21_{\pm 0.47}$
	RCNS, ϵ=0.10	$\mathbf{2.76}_{\pm 0.50}$	$\mathbf{4.54}_{\pm 0.64}$	$\mathbf{1.81}_{\pm 0.43}$	$\mathbf{2.39}_{\pm 0.48}$
	Relative gain	+8.36%	+6.15%	+9.15%	+7.96%
Video Games	GRPO	$9.32_{\pm 1.00}$	$14.98_{\pm 1.99}$	$5.99_{\pm 0.76}$	$7.81_{\pm 1.08}$
	RCNS, ϵ=0.10	$\mathbf{10.43}_{\pm 1.44}$	$\mathbf{16.76}_{\pm 1.82}$	$\mathbf{6.89}_{\pm 1.00}$	$\mathbf{8.91}_{\pm 1.11}$
	Relative gain	+11.93%	+11.87%	+14.97%	+14.07%
Movies and TV	GRPO	$10.77_{\pm 0.94}$	$16.66_{\pm 1.12}$	$6.96_{\pm 0.78}$	$8.86_{\pm 0.83}$
	RCNS, ϵ=0.10	$\mathbf{11.90}_{\pm 1.75}$	$\mathbf{18.12}_{\pm 2.04}$	$\mathbf{7.78}_{\pm 1.27}$	$\mathbf{9.78}_{\pm 1.31}$
	Relative gain	+10.53%	+8.80%	+11.72%	+10.40%

RCNS improves the mean of every reported metric in all three domains (Table 11). Relative HR@5 gains range from 8.36% to 11.93%, with the largest gains on Video Games. Together with the Qwen3-1.7B results, this within-backbone comparison shows downstream gains at both model sizes.

D Risk-Budget Calibration Details

Power-profile calibration. For the default profile $h(s; p) = s^p$, increasing p reduces attenuation for $0 < s < 1$: $p < 1$ spreads attenuation more broadly, $p = 1$ gives a linear profile, and $p > 1$ concentrates it on high-priority negatives. With calibration recommendations, group and batch assignments, priorities, labels, and weights fixed, $s^{p_1} \geq s^{p_2}$ for $p_1 < p_2$. RGC and FCR are therefore non-increasing in p. Over the closed interval $[p_{\min}, p_{\max}]$, continuity ensures that a nonempty feasible set has a smallest exponent, which maximizes coverage:

$$p^\star = \min\{p \in [p_{\min}, p_{\max}] : \mathrm{FCR}_{\mathrm{val}}(p) \leq \epsilon_{\mathrm{risk}}\}. \tag{5}$$

We locate this boundary by one-dimensional search, selecting $p_{\min}$ if all exponents are feasible and applying no attenuation if none is feasible. The power profile fully attenuates priority-one negatives for every admissible exponent. An untied best-ranked negative has this priority when $n_{\mathcal{B}} > 1$; averaging tied positions can make the maximum priority smaller than one.

Table 12 reports the exponent calibrated under each validation risk budget on fixed diagnostic rollouts. Complete groups of eight actions are ordered by group identifier and partitioned into diagnostic batches of 128 groups, retaining the final partial batch. For each budget, we select the smallest feasible $p \in [0.01, 64]$ using 80 bisection steps under $\mathrm{FCR}_{\mathrm{val}}(p) \leq \epsilon_{\mathrm{risk}}$. Test feedback is used only for evaluation after calibration. Larger budgets permit broader attenuation.

Across validation risk budgets from 5% to 35%, larger budgets select smaller exponents and increase both test recovery coverage and test FCR in each domain. The complete sweep therefore shows the expected empirical coverage–risk trade-off without selecting parameters on test feedback.

Table 12: Calibrated exponent $p^\star$ for fixed diagnostic batches.

ϵ	0.05	0.10	0.15	0.20	0.25	0.30	0.35
C&V	20.01	9.31	5.76	4.02	2.99	2.31	1.83
VG	16.68	8.26	5.30	3.78	2.86	2.23	1.78
M&TV	20.18	8.89	5.49	3.85	2.89	2.25	1.79

D.1 Calibration Robustness across Seeds

Calibration and held-out evaluation are repeated across five independent seeds at a 10% validation risk budget. Table 13 reports means and sample standard deviations of test RGC, FCR, and attenuated-mass precision (Precision), all in percentages. For nonzero attenuated mass, $\mathrm{Prec}(p) = \frac{\sum_{i\in\mathcal{N}} \alpha_i(p) w_i y_i^{\mathrm{SC}}}{\sum_{i\in\mathcal{N}} \alpha_i(p) w_i}$ measures its sign-conflicting share, with total attenuated advantage mass as the denominator.

Table 13: Calibration at a 10% validation risk budget across five independent seeds. Test metrics report mean $\pm$ sample SD (%); the last column reports sample SD of $p^\star$.

Dataset	RGC	FCR	Precision	$\mathrm{Std}(p^\star)$
C&V	20.07 ± 4.69	10.16 ± 0.61	9.33 ± 1.94	0.36
VG	16.50 ± 2.16	10.60 ± 0.30	17.08 ± 1.70	0.45
M&TV	22.34 ± 4.30	9.78 ± 0.54	14.01 ± 2.58	0.56

D.2 Alternative Attenuation Profiles

We compare three profiles using the same batch-relative priorities and a 10% validation risk budget. The linear control uses $h(s;c) = cs$, selecting the largest feasible $c \in [0,1]$. The exponential profile uses $h(s;\gamma) = [\exp(\gamma s) - 1]/[\exp(\gamma) - 1]$, with $h(s;0) = s$. We select the smallest feasible $\gamma \in [0, 128]$ by 80 bisection steps, applying no attenuation if none is feasible. Power uses the calibration above. Parameters are selected on validation feedback and evaluated on test feedback across five independent seeds, following Appendix B.

Table 14: Attenuation profiles at a 10% validation risk budget. Entries are mean test RGC / FCR (%).

Profile	C&V	VG	M&TV
Linear	12.57 / 10.05	12.62 / 10.10	13.42 / 9.99
Exponential	19.67 / 10.15	16.30 / 10.57	22.01 / 9.78
Power (default)	20.07 / 10.16	16.50 / 10.60	22.34 / 9.78

Both nonlinear profiles achieve higher test RGC than the linear control, while exponential and power attenuation exhibit similar coverage and FCR (Table 14).

D.3 Calibration Transfer During Policy Training

We evaluate whether the calibration obtained before policy training remains effective as the policy evolves. For each domain, we calibrate the exponent once on a fixed pretrained rollout pool at a validation risk budget of $\epsilon = 0.10$ and denote this initial value by p_0. We then generate rollouts from saved policy checkpoints on the same user contexts and evaluate them using the same reward, priority, and batching rules, while keeping p_0 fixed throughout checkpoint evaluation. The calibrated values are 9.31, 8.26, and 8.89 for C&V, VG, and M&TV, respectively; evaluation uses their unrounded values.

Table 15 compares the pretrained reference, the policy after the first full pass over the policy-training contexts, and the final policy after the second pass. With p_0 held fixed, mean test FCR remains within one percentage point of the 10% validation risk budget after both training passes in all three domains.

Recovery coverage varies across domains and training stages, with final-policy mean RGC ranging from 13.57% to 19.72%. These results show that the pretrained calibration remains empirically effective as the policy evolves, without checkpoint-specific recalibration.

Table 15: Calibration transfer with fixed p_0 at $\epsilon = 0.10$. Test metrics are percentages; training-checkpoint results report mean $\pm$ sample SD over five training seeds.

Domain	Policy stage	RGC	FCR
C&V	Pretrained	17.03	9.31
	After first pass	19.25 ± 3.55	10.56 ± 0.51
	Final (2 passes)	16.18 ± 5.46	10.92 ± 1.36
VG	Pretrained	16.05	10.50
	After first pass	14.49 ± 0.60	9.84 ± 1.06
	Final (2 passes)	13.57 ± 2.51	10.62 ± 0.86
M&TV	Pretrained	15.56	9.84
	After first pass	22.97 ± 3.57	9.44 ± 0.20
	Final (2 passes)	19.72 ± 3.27	10.09 ± 0.83

Each domain uses one shared pretrained reference. All training-checkpoint statistics are computed over the same five training seeds. The first training checkpoint is taken after one full pass over the policy-training contexts, and the final checkpoint is taken after the second pass.

E Advantage-Margin Control

To test whether priority predicts sign conflict beyond the mechanical effect of proximity to zero, we compare recommendations with negative observed advantages in the top and bottom priority quartiles. We greedily match them one-to-one within the same dataset and diagnostic batch by $|A^{\mathrm{obs}}|$, imposing the main caliper $\left||A_i^{\mathrm{obs}}| - |A_j^{\mathrm{obs}}|\right| \leq 0.02$. Confidence intervals use 5,000 batch-cluster bootstrap draws, and binary conflict p-values use an exact paired discordance test.

Table 16: Margin-matched test comparisons. Rates are in %; risk differences (RD) and their CIs are in percentage points. RD and $\Delta A = A^{\mathrm{sub}} - A^{\mathrm{obs}}$ contrasts compare high minus low priority.

Dataset	Pairs	High	Low	RD [95% CI]	Paired p-value	ΔA diff. [95% CI]	|SMD|
C&V	190	11.05	4.21	6.84 [1.13, 14.91]	0.011	0.01 [-0.11, 0.14]	5.16×10^{-4}
VG	440	17.95	6.59	11.36 [3.70, 21.39]	9.70×10^{-7}	0.22 [0.09, 0.38]	3.06×10^{-3}
M&TV	234	12.39	5.56	6.84 [0.00, 14.46]	0.017	0.05 [-0.06, 0.14]	3.62×10^{-3}
Pooled	864	14.93	5.79	9.14 [4.54, 14.67]	7.65×10^{-10}	0.13 [0.05, 0.22]	7.88×10^{-4}

Matching achieves close balance in observed advantage magnitude, with a maximum absolute standardized mean difference of 0.0036. Estimated conflict-rate differences are positive in every domain. Table 16 reports the domain-level and pooled contrasts with their confidence intervals.

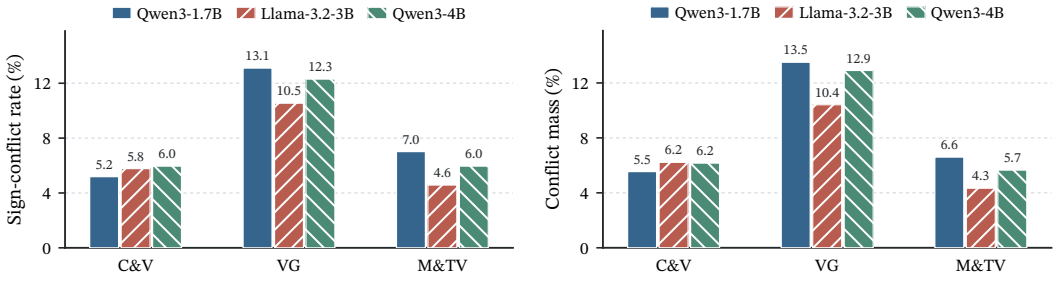
Table 17 varies both the caliper and the fraction defining the priority arms. The pooled conflict-rate difference ranges from 7.49 to 9.14 percentage points, with batch-bootstrap confidence intervals above zero in every setting. The pooled continuous contrast also remains positive throughout. The pooled association persists across the tested priority-arm fractions and calipers.

Table 17: Matching sensitivity on pooled test data. RD is the high-minus-low conflict-rate difference in percentage points.

Arm fraction	Caliper	Pairs	RD [95% CI]	ΔA diff. [95% CI]
0.25	0.01	812	8.99 [4.41, 13.92]	0.12 [0.04, 0.20]
0.25	0.02	864	9.14 [4.54, 14.67]	0.13 [0.05, 0.22]
0.25	0.05	966	8.90 [4.22, 13.95]	0.12 [0.05, 0.20]
$1/3$	0.01	1189	7.49 [3.93, 11.19]	0.10 [0.04, 0.17]
$1/3$	0.02	1245	7.55 [4.02, 11.76]	0.10 [0.04, 0.17]
$1/3$	0.05	1370	7.52 [4.06, 11.28]	0.10 [0.04, 0.16]

F Cross-Backbone Sign Conflicts

We compare Qwen3-1.7B, Llama-3.2-3B-Instruct, and Qwen3-4B under shared contexts, chronological splits, prompts, sampling, item catalogs, Qwen3-Embedding-0.6B evaluation, rank reward, and group normalization. This offline reliability diagnostic is separate from the downstream Qwen3-4B training comparison in Appendix C.



(a) Recommendation-level conflict rate

(b) Advantage-weighted conflict mass

Figure 8: Validation conflicts among recommendations with negative observed advantages: (a) fraction receiving positive validation advantages; (b) share of absolute advantage mass. Llama denotes Llama-3.2-3B-Instruct.

Protocol. Offline sampling uses $G = 8$ actions per context, rollout-sampling seed 42, temperature 0.8, top-$p = 0.95$, and at most 160 response tokens. Prompts include at most 40 recent item titles within 2,048 tokens. The generated recommendation titles are extracted, embedded, and ranked against the fixed catalog without masking history items. Backbones generate separate actions, held fixed within each backbone when replacing the observed training target with validation feedback (the best rank among the user's validation interactions). Rewards follow the rank-based function in Appendix B; advantages use the training implementation with float32 group means and sample standard deviations and a 10^{-6} denominator stabilizer.

Statistics. We retain only complete eight-action rollout groups with validation feedback from the diagnostic subsets in Appendix A. For backbone b and domain d, let $\mathcal{N}_{b,d}$ contain the retained recommendations with negative observed advantages and $c_i = \mathbb{I}(A_i^{\mathrm{val}} > 0)$, where A_i^{val} is the subsequent-feedback advantage computed using validation interactions. We report $\mathrm{ConflictRate}_{b,d} = 100\,\frac{\sum_{i\in\mathcal{N}_{b,d}} c_i}{|\mathcal{N}_{b,d}|}$ and $\mathrm{ConflictMass}_{b,d} = 100\,\frac{\sum_{i\in\mathcal{N}_{b,d}} |A_i^{\mathrm{obs}}| c_i}{\sum_{i\in\mathcal{N}_{b,d}} |A_i^{\mathrm{obs}}|}$.

As shown in Figure 8, sign conflicts occur in all nine backbone–domain settings. Their prevalence varies across models, while Video Games has the highest conflict rate for each backbone. The cross-feedback instability diagnosed in RQ1 therefore extends beyond Qwen3-1.7B under the shared evaluation protocol.

G Paired Uncertainty Analysis

We quantify training-seed uncertainty for the Qwen3-1.7B comparison in Table 1, pairing RCNS ($\epsilon = 0.10$) and GRPO runs with the same seed (42–46). For metric m and seed s, the paired difference is $\Delta_m^{(s)} = M_m^{\mathrm{RCNS},(s)} - M_m^{\mathrm{GRPO},(s)}$, where M is evaluated on the complete test set. Using the per-seed differences reported in Table 19, we resample the five matched pairs with replacement and compute their mean difference, enumerating all $5^5 = 3{,}125$ ordered resamples. The 95% percentile bootstrap interval uses the 2.5th and 97.5th percentiles of this distribution, with linear interpolation.

Table 18: Paired mean differences and 95% bootstrap intervals ($\times 10^3$; positive favors RCNS).

Dataset	Δ HR@5	Δ NDCG@5	Δ HR@10	Δ NDCG@10
C&V	$+0.264$ $[-0.125, 0.640]$	$+0.192$ $[-0.041, 0.425]$	$+0.389$ $[-0.075, 0.841]$	$+0.237$ $[-0.035, 0.465]$
VG	$+0.578$ $[-0.053, 1.243]$	$+0.272$ $[-0.071, 0.803]$	$+0.411$ $[-0.920, 1.681]$	$+0.213$ $[-0.319, 0.926]$
M&TV	$+1.631$ $[-0.344, 3.734]$	$+1.071$ $[-0.536, 2.678]$	$+2.727$ $[-0.204, 6.066]$	$+1.414$ $[-0.489, 3.317]$

Mean paired differences are positive in all twelve dataset–metric comparisons, with the largest absolute gains on Movies and TV. The per-seed differences in Table 19 show the run-to-run variation underlying these mean gains.

Table 19: Per-seed RCNS–GRPO differences ($\times 10^3$), paired by training seed.

Dataset	Seed	Δ HR@5	Δ NDCG@5	Δ HR@10	Δ NDCG@10
C&V	42	$+0.565$	$+0.459$	$+0.439$	$+0.436$
	43	$+0.878$	$+0.512$	$+1.004$	$+0.539$
	44	-0.063	$+0.017$	$+0.063$	$+0.053$
	45	$+0.314$	$+0.182$	$+0.878$	$+0.376$
	46	-0.376	-0.210	-0.439	-0.221
VG	42	$+0.482$	$+0.125$	$+0.657$	$+0.183$
	43	-0.482	-0.184	-1.839	-0.613
	44	$+0.701$	$+0.070$	$+0.963$	$+0.122$
	45	$+1.751$	$+1.291$	$+2.583$	$+1.537$
	46	$+0.438$	$+0.057$	-0.307	-0.166
M&TV	42	$+4.779$	$+3.135$	$+8.220$	$+4.221$
	43	$+3.823$	$+3.263$	$+5.480$	$+3.782$
	44	-1.784	-1.385	-1.147	-1.179
	45	$+0.828$	$+0.592$	$+0.892$	$+0.578$
	46	$+0.510$	-0.250	$+0.191$	-0.333