LET THE WEAK JUDGE: ANNOTATION-FREE PROCESS REWARDS FOR MULTIMODAL E-COMMERCE RELEVANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Outcome-only reinforcement learning can improve final-label accuracy without making discriminative multimodal reasoning evidence-grounded. In tasks with a small label space, GRPO can exploit lexical shortcuts and fabricate *forced associations*: rhetorically plausible links that reach a target label without a supporting cross-modal evidence chain. Process reward models (PRMs) address this problem in principle, but usually require costly step-level annotations and can mis-score policy traces outside their training distribution. We propose **WISE-GRPO**, an annotation-free process reward based on *evidence recoverability*. At each intermediate step, we mask the policy's upfront answer and ask an independent weak-model committee to continue from the reasoning prefix. The prefix is rewarded only when the evidence it contains lets the committee recover the ground-truth label. An answer-first policy and a think-first evaluator prevent answer leakage, while progressive weights emphasize the integrative decisions near the end of the task template. Across four e-commerce benchmarks and two multimodal backbones, WISE-GRPO achieves the best accuracy among the evaluated methods in all eight settings, improving over outcome-only GRPO by $+1.5$ to $+4.8$ accuracy points. Analyses show that the reward tracks evidence rather than surface form and reduces forced-association behavior, without step-level human annotation.

1 INTRODUCTION

Outcome-only reinforcement learning does not require evidence-grounded reasoning. When GRPO (Shao et al., 2024) optimizes a multimodal policy against a sparse label reward on discriminative tasks with a small label space, a policy can reach the correct class by exploiting lexical shortcuts rather than resolving the relevant cross-modal evidence. More problematically, it can fabricate *forced associations*: hedging rhetoric that bridges an unsupported logical gap to the target label. Across the four e-commerce benchmarks studied here, up to 64.3% of late-training outcome-reward-only responses contain such rhetoric.

This failure exposes a gap between label correctness and evidence communicability. A reasoning trace should not receive credit merely because it looks coherent or happens to end in the correct label. It should receive credit when the evidence accumulated so far is sufficient for an independent reader to make the correct decision. Existing process reward models (PRMs) approach this goal by scoring individual steps (Lightman et al., 2023), but step-level human labels for query–product reasoning are expensive and quickly become stale as catalogue attributes and business rules evolve (Khalifa et al., 2025; Chen et al., 2026). A static learned PRM can also reward fluent but evidence-lacking traces after the policy shifts away from its training distribution (Zhang et al., 2025b).

We therefore ask a different question: *does a reasoning prefix make the correct label recoverable?* A prefix that organizes the necessary cross-modal evidence should enable a reader without access to the policy's decision to recover the ground-truth label. This operationalizes process supervision as a verifiable prefix-continuation test: a prefix receives credit only when it enables an independent evaluator to recover the label.

We instantiate this principle with **WISE-GRPO** (**W**eak-model **I**nformed **S**tep-wise **E**valuation GRPO). The policy outputs an answer-first trace; at every intermediate prefix, its answer is masked

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.

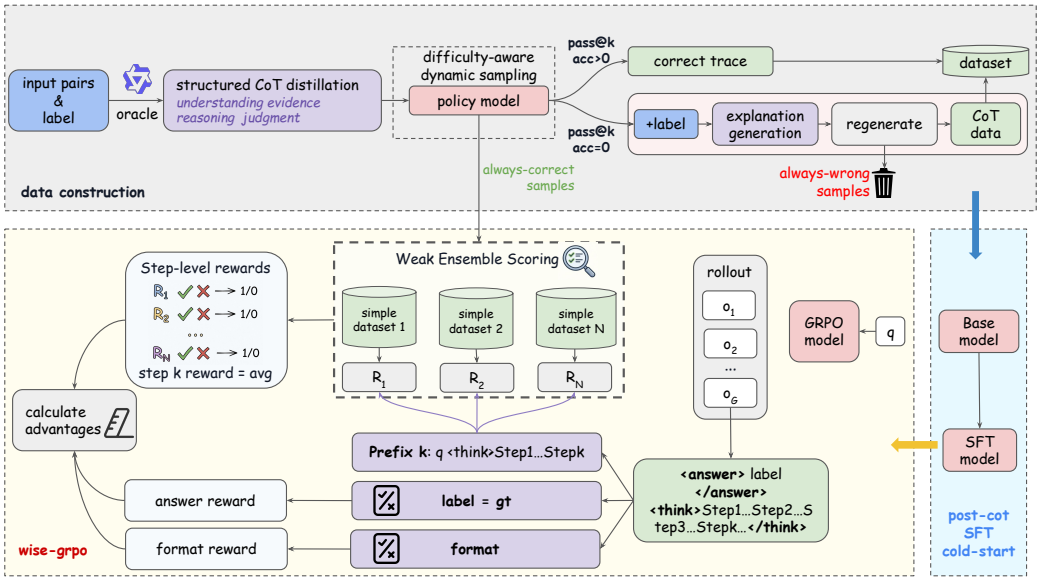


Figure 1: **Overview.** *Top:* oracle-guided structured CoT distillation partitions data by pass@N difficulty; correct traces seed SFT, always-correct instances train the weak committee. *Bottom:* WISE-GRPO rewards each step by how well a weak-model committee, given only the prefix, recovers the ground-truth label.

and a committee of evidence-sensitive weak evaluators independently continues the reasoning in a think-first format. The process reward is the committee's recovery accuracy on the ground-truth label. The asymmetric formats prevent answer leakage. Easy-data training and committee construction provide evaluators with headroom while preserving sensitivity to prefix evidence, and progressive step weights emphasize the later integrative decisions of the fixed reasoning template.

Our contributions. First, we formulate annotation-free process supervision as *evidence recoverability*: a prefix is rewarded only when it lets an independent evaluator recover the correct discrete label. Second, we realize this criterion through an answer-masked asymmetric continuation protocol and an evidence-sensitive weak-model committee, yielding WISE-GRPO. Third, across four e-commerce benchmarks and two multimodal backbones, WISE-GRPO improves accuracy over outcome-reward and supervised-PRM baselines. Controlled probes further show that the recovery reward responds to evidence content rather than surface form, while training reduces forced-association behavior.

2 Related Work

Process rewards for LLM RL. Outcome reward models (ORMs) provide sparse feedback that limits multi-step credit assignment, motivating a line of work on dense signals (Lightman et al., 2023; Wang et al., 2023). Two threads are particularly relevant. Step-level supervised PRMs collect per-step labels either from humans or from Monte-Carlo rollouts, then train a scorer used as an RL reward or a re-ranker (Luo et al., 2024). This line reaches strong results in math reasoning but inherits the cost and coverage of step-level data. A second thread parameterizes the reward as progress under a distinct evaluator policy, arguing that a step should be credited when it changes the likelihood of eventual success (Setlur et al., 2024). Concurrent work also shows that generative verifiers trained with next-token prediction (Zhang et al., 2025a) and implicit PRMs derived from outcome-only data (Yuan et al., 2024) can approach supervised-PRM performance without dense annotation. Our answer-recovery test sits in the same family of annotation-free process rewards, and instantiates it for a setting these works do not address: discriminative multimodal classification with a small label space, where the correctness signal is categorical rather than a scalar success probability.

LLM-based e-commerce relevance. Query-product relevance has moved from lexical matching (Devlin et al., 2019) to reasoning-driven LLM pipelines. LREF combines SFT, CoT, and DPO (Tang et al., 2025; Rafailov et al., 2024). TaoSR1 applies GRPO with difficulty-aware sampling (Dong et al., 2026) but keeps outcome-only rewards. TaoSR-SHE (Jiao et al., 2026) and

TaoSR-AGRL (Yang et al., 2026) introduce step-level supervision, but both rely on hand-crafted criteria or a separately trained PRM. WISE-GRPO differs in that step rewards are derived from the answer-recovery behavior of a weak-model committee, requiring neither hand-designed criteria nor step-level human labels.

3 Evidence Recoverability as Process Supervision

3.1 The Answer-Recovery Test

A reasoning prefix is worth crediting only if it lets an independent reader reach the correct conclusion. We instantiate this criterion by extracting the first k steps of a policy trajectory, masking the policy's upfront answer, and passing the prefix to a committee of weak models that must resume reasoning and produce their own label. A step earns reward only when the committee recovers the ground-truth label from the prefix alone. The subjective question "is this step correct?" is replaced by the verifiable question "does this prefix carry enough evidence for a capacity-limited reader to finish the task?".

Preventing answer leakage through asymmetric continuation. We deliberately place the policy and the committee on *opposite* sides of the answer-reasoning order. The policy generates in a **Post-CoT** form, <answer> $y^\star$ </answer> <think> t </think>: emitting the label first anchors the decision and prevents mid-trace drift from corrupting the final prediction. The committee generates in a **Pre-CoT** form, <think> t </think> <answer> $\hat{y}$ </answer>: without access to the policy's answer, each evaluator must produce its own reasoning from the truncated prefix and commit to a label at the end. This asymmetry is what makes recovery accuracy meaningful: the two agents optimize different objectives yet share a label space, so the committee's label is an independent judgment of the prefix's evidence content.

3.2 Constructing an Evidence-Sensitive Evaluator Committee

For committee construction we collect the trivial-instance pool $\mathcal{D}_{\text{easy}}$, partition it into M disjoint subsets, and train M weak evaluators independently via SFT from the same base as the policy. Disjoint training subsets induce complementary errors across evaluators, which is what majority voting later exploits. Training only on easy data also prevents the evaluators from acquiring the policy's harder-case reasoning patterns; a committee that shares the policy's biases would give free credit for structural familiarity rather than for evidence.

3.3 Aggregating Recovery across Reasoning Steps

A recoverable prefix is useful only when the evaluator has enough headroom to rely on the prefix rather than on its own priors. Section 4 validates this condition empirically. During RL rollout we extract the reasoning trace t from the policy's Post-CoT output, parse it into K steps $\{s_1, \ldots, s_K\}$, and for each k form a Pre-CoT prefix $p_k = [x,\ \texttt{<think>},\ s_1, \ldots, s_k]$ that is dispatched to the committee. Each evaluator RM_j resumes reasoning from p_k and produces a label $\hat{y}_{k,j}$. The step-level process reward is the committee's empirical recovery accuracy,

$$R_k(c) \;=\; \frac{1}{M}\sum_{j=1}^{M} \mathbf{1}[\hat{y}_{k,j} = y^\star]\,. \tag{1}$$

The total trajectory reward combines outcome, format, and progressive process signals,

$$R(c) \;=\; \lambda_{\text{acc}} R_{\text{acc}} \;+\; \lambda_{\text{fmt}} R_{\text{fmt}} \;+\; \lambda_{\text{proc}} \sum_{k=1}^{K} \alpha_k R_k(c), \tag{2}$$

with $R_{\text{acc}} = \mathbf{1}[\hat{y} = y^\star]$, format reward $R_{\text{fmt}} = \frac{1}{2}(R_{\text{tag}} + R_{\text{step}})$, and non-decreasing step coefficients $0 < \alpha_1 \le \cdots \le \alpha_K$. Rather than weighting all prefix-level rewards uniformly, we align their contributions with the functional roles of the structured reasoning template. Early steps identify the intent and extract local evidence, whereas later steps consolidate evidence across modalities and apply the task-specific decision rule that determines the label. We therefore use non-decreasing coefficients $0 < \alpha_1 \le \cdots \le \alpha_K$ to place more optimization mass on these integrative decisions, rather than

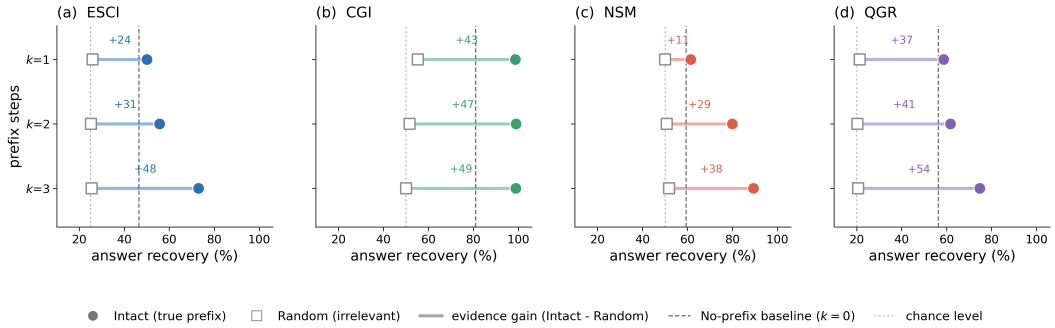


Figure 2: Committee answer recovery increases with prefix length k, and the increase comes from evidence itself. Each panel corresponds to one task (ESCI, CGI, NSM, QGR); the y axis shows prefix steps $k = 1, 2, 3$ and the x axis shows answer recovery (%). Filled circles mark Intact (true-prefix) recovery, open squares mark Random (irrelevant-prefix) recovery, and the connecting line labels the evidence gain (Intact minus Random). The black dashed line marks the No-prefix baseline ($k = 0$); the gray dotted line marks chance level.

claiming that every recovery curve must be strictly monotone. Table 4 validates this weighting choice directly across all eight settings.

We optimize with GRPO. For a group of G samples $\{c_i\}_{i=1}^{G} \sim \pi_{\text{old}}(\cdot \mid x)$, the group-normalized advantage $A_i = (R(c_i) - \mu_R)/\sigma_R$ and the clipped surrogate $\mathcal{L}_{\text{GRPO}} = -\mathbb{E}[\min(\rho_i A_i,\ \text{clip}(\rho_i, 1 - \varepsilon, 1 + \varepsilon)A_i)]$, where $\rho_i = \pi_\theta(c_i \mid x)/\pi_{\text{old}}(c_i \mid x)$.

4 Why Answer Recovery Measures Evidence

The method rests on the hypothesis that a weak-model committee recovers the correct label from a prefix only when the prefix contains genuine evidence, rather than fluency, lexical overlap, or label priors. We test whether the reward tracks evidence content, remains informative across evaluator strengths, and separates correct from erroneous reasoning more sharply than a supervised PRM. All probes use pre-conclusion prefixes, so recovery accuracy reflects accumulated evidence rather than the final decision.

4.1 Recovery Tracks Evidence, Not Surface Form

Recovering the answer from a prefix does not by itself show that the prefix carries evidence: a fluent, on-topic, or template-conforming prefix could make the label easier to guess. To separate genuine evidence from surface form, we probe the committee under three prefix conditions on a class-balanced set of 4,000 examples per task.

- **Intact** contains the first k steps of a policy-generated trajectory **that reaches the correct final answer**, so larger k supplies more evidence.
- **Random** replaces the prefix with a fluent, template-matched trace drawn from an unrelated example, preserving style and format but carrying no relevant evidence.
- **No-prefix** provides only the original input and serves as the committee's input-only baseline.

Figure 2 isolates the effect of evidence from the effect of surface form. Random stays near the task-specific chance level at every k, so a fluent but content-irrelevant prefix adds nothing. Intact instead rises with k, confirming that the gain tracks how much task-relevant content the prefix carries, not how fluent it looks. On tasks where the reasoning chain is short and already easy to solve, this rise saturates quickly; a chain that already reaches the correct answer has little headroom left to reveal, so a plateau is the expected signature of a solved task, not a failure of the probe. The Intact-Random gap isolates the prefix's own contribution, and the No-prefix baseline marks the accuracy attainable from the input alone. Committee recovery therefore tracks evidence, not surface form, and because the process reward during training is computed on these prefixes, this probe characterizes the very signal that drives optimization.

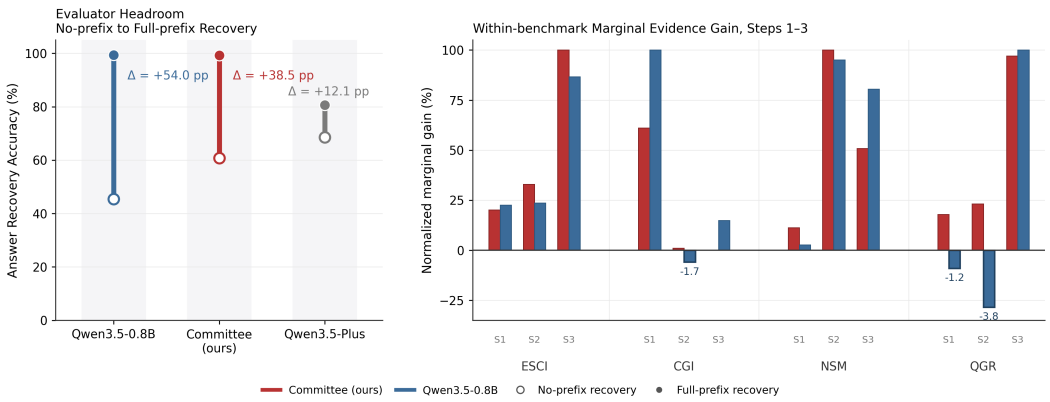


Figure 3: **Evaluator scale alone does not determine useful prefix sensitivity. (a)** No-prefix to full-prefix recovery, averaged across four tasks. **(b)** Marginal recovery gains over the first three reasoning steps. Qwen3.5-Plus leaves little headroom, while Qwen3.5-0.8B exhibits negative partial-prefix gains on CGI and QGR; the committee provides more stable intermediate signals.

4.2 Useful Evaluators Need Both Headroom and Stable Prefix Sensitivity

For answer recovery to serve as a useful process reward, the evaluator must respond to evidence in the prefix rather than solve the task from its own priors. This creates a capacity trade-off. A strong evaluator may recover the label without the prefix, whereas an insufficiently adapted weak evaluator may respond unreliably to partial evidence. To examine this trade-off, we compare the flagship Qwen3.5-Plus model, the task-agnostic Qwen3.5-0.8B model, and our committee.

What operating region makes recovery informative? For an evaluator with no-prefix recovery p_0, full-prefix recovery p_K, and prefix-k recovery $\hat{p}_k$, the available headroom is $p_K - p_0$, and the marginal signal at step k is $g_k = \hat{p}_k - \hat{p}_{k-1}$. A strong evaluator leaves little headroom because it can recover labels from its own priors. An insufficiently adapted weak evaluator may retain headroom but produce erratic or negative g_k on partial prefixes. An informative evaluator should therefore combine substantial headroom with stable intermediate signals.

Figure 3 illustrates this trade-off. Qwen3.5-Plus leaves limited headroom, whereas Qwen3.5-0.8B exhibits negative partial-prefix gains on CGI and QGR. The committee retains substantial headroom while providing more stable intermediate signals; its CGI trajectory saturates after the first prefix step. We construct the committee by adapting members on disjoint easy-data subsets. Standalone accuracy, diversity, and tie statistics for the $M{=}3$ committee are reported in App. A.

4.3 Why the Recovery Reward Beats a Supervised PRM

A process reward is useful for GRPO only if it assigns systematically different signals to correct and erroneous reasoning within a rollout group. We therefore evaluate whether WISE-RM, the committee recovery reward R_k from §3.3, separates correct CoTs from plausible but erroneous ones more sharply than a static step-level supervised reward model $\mathrm{RM}_k^{\mathrm{sup}}$. We conduct the comparison on a controlled diagnostic set containing $20{,}000$ human-verified error-free CoTs and $80{,}000$ erroneous CoTs. The erroneous samples cover four flaw types: Cognitive Bias, Criterion Scaling, Logical Inconsistency, and Factual Hallucination.

This separation is operationally relevant for GRPO, which learns from relative reward differences within each rollout group. A wider gap between error-free and erroneous CoTs therefore yields a sharper learning signal. The diagnostic shows that answer recovery provides stronger within-group separation than the static supervised reward model, without attributing the difference to any single source of reward-model error.

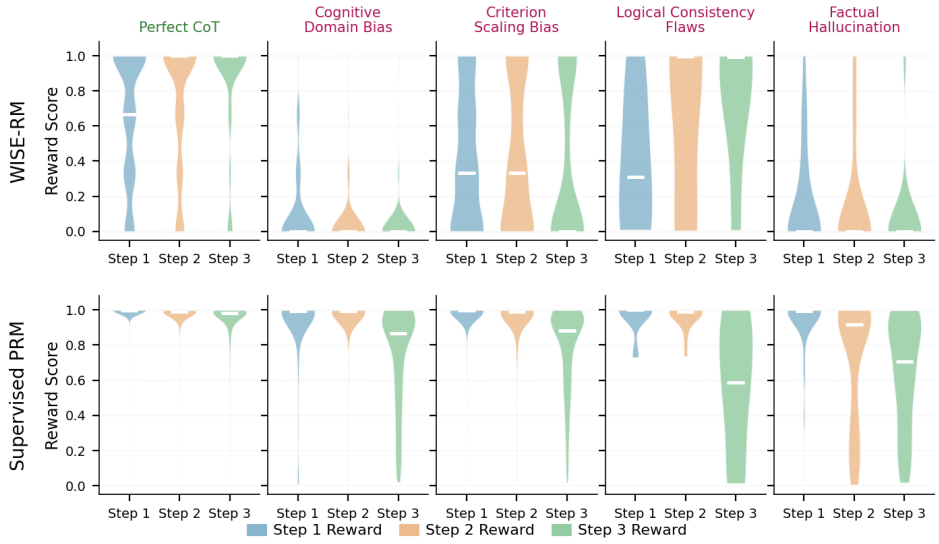


Figure 4: **WISE-RM separates Perfect from Error CoTs more sharply than a supervised PRM.** Reward distributions on 20K Perfect and 80K Error CoTs (four flaw types). The supervised PRM's high-score mode persists across all error types; WISE-RM shifts downward to the 0.0–0.4 range on errors, providing a wider relative-advantage signal under GRPO.

5 Experiments

5.1 Setup

Datasets and metrics. We evaluate on Amazon ESCI (Reddy et al., 2022), a public four-way query-product benchmark, and three industrial multimodal benchmarks from a large social-commerce platform: QGR, NSM, and CGI (App. D). We report accuracy and per-class F1.

Training data and protocol. All learned methods start from the same cold-start SFT checkpoint and draw on the same ∼90k RL data pool. We construct training data by routing oracle-generated K-step CoT traces according to *pass@N*: trivial instances train the committee, frontier instances initialize the policy, and intractable instances are regenerated with oracle-provided factual hints. PC-DPO and MR-GRPO also derive their supervision from this pool, so performance differences primarily reflect the training objective. We report single-seed results.

Baselines. We compare WISE-GRPO with SFT, PC-DPO, outcome-reward GRPO (OR-GRPO), and GRPO with a static supervised reward model (MR-GRPO).

5.2 Main Results

Table 1 compares WISE-GRPO with the baselines across four benchmarks and two multimodal backbones (Bai et al., 2025b;a). WISE-GRPO achieves the highest accuracy in all eight backbone–benchmark settings without step-level human annotation.

OR-GRPO yields mixed and generally modest gains over SFT, indicating that final-answer correctness alone does not reliably shape the intermediate reasoning behavior targeted here. MR-GRPO consistently underperforms OR-GRPO, suggesting that dense supervision from a static reward model is insufficient when the reward cannot sharply separate correct from plausible but erroneous CoTs. The diagnostic in §4.3 provides supporting evidence for this interpretation.

5.3 Ablation Studies

We jointly ablate the CoT SFT cold start, reasoning order, and step-weighting scheme. Table 3 shows that the two-stage pipeline consistently outperforms either component in isolation across all tasks and both backbones. CoT SFT supplies the policy with a structured reasoning prior before RL, while

Method	QGR						ESCI					CGI				NSM			
	$F1_0$	$F1_1$	$F1_2$	$F1_3$	$F1_4$	Acc.	F1(I)	F1(C)	F1(S)	F1(E)	Acc.	P	R	F1	Acc.	P	R	F1	Acc.
Qwen2.5-VL-3B-Instruct																			
SFT	77.61	74.64	**75.84**	72.30	81.58	76.52	70.21	**55.74**	67.72	85.93	77.94	82.84	83.52	83.18	83.11	87.18	92.75	89.88	84.09
PC-DPO	67.43	63.49	62.22	60.74	75.90	66.28	59.48	19.74	57.33	81.97	70.65	77.42	75.95	76.68	76.90	82.12	88.93	85.39	76.81
OR-GRPO	73.68	70.18	70.54	67.23	79.98	72.54	66.74	42.91	62.55	85.27	75.31	81.53	81.49	81.51	81.51	87.75	93.87	90.71	85.34
MR-GRPO	72.91	69.45	69.78	66.53	79.12	71.78	65.50	42.91	60.83	84.51	74.17	80.74	80.70	80.72	80.72	87.40	93.21	90.21	84.59
WISE-GRPO	**79.16**	**76.19**	74.32	**73.71**	**82.83**	**77.38**	**71.80**	50.56	**68.82**	**87.15**	**78.95**	**83.79**	**84.52**	**84.15**	**84.08**	**89.53**	**95.19**	**92.27**	**87.85**
Qwen3-VL-4B-Instruct																			
SFT	79.88	78.46	79.23	76.12	83.99	79.62	75.12	85.78	67.15	74.64	76.34	86.13	85.97	86.05	86.06	87.69	93.41	90.46	84.99
PC-DPO	72.60	71.46	71.91	69.33	76.40	72.40	67.93	76.81	60.80	67.51	68.70	81.03	78.85	79.93	80.20	82.47	89.59	85.89	77.56
OR-GRPO	80.66	79.25	80.02	76.92	84.86	80.42	76.21	86.27	68.15	75.73	77.24	87.43	87.16	87.30	87.32	88.36	94.53	91.34	86.35
MR-GRPO	79.45	78.05	78.82	75.76	83.53	79.20	75.08	85.00	67.12	74.61	76.06	86.77	86.50	86.63	86.66	88.02	93.87	90.85	85.59
WISE-GRPO	**82.26**	**80.79**	**81.55**	**78.46**	**86.51**	**82.00**	**77.86**	**87.79**	**69.71**	**77.37**	**78.84**	**88.89**	**88.61**	**88.75**	**88.77**	**89.59**	**95.85**	**92.62**	**88.35**

Table 1: Performance comparison of WISE-GRPO and baseline methods across four benchmarks. Results are grouped by benchmark and metric, with methods listed on the left. For QGR, $F1_k$ denotes the F1 score of class k. For CGI and NSM, P/R/F1 are reported on the positive classes, i.e., identical-pair and same-SPU classes, respectively. The best result for each metric under each backbone is shown in bold.

Paradigm	QGR						ESCI					CGI				NSM			
	$F1_0$	$F1_1$	$F1_2$	$F1_3$	$F1_4$	Acc.	F1(I)	F1(C)	F1(S)	F1(E)	Acc.	P	R	F1	Acc.	P	R	F1	Acc.
Qwen2.5-VL-3B-Instruct																			
Pre-CoT	78.97	**76.24**	**77.49**	**76.63**	71.22	75.90	**79.09**	**75.89**	**74.98**	78.64	77.20	80.38	80.70	80.54	80.50	89.23	90.58	89.90	84.49
Post-CoT	**79.16**	76.19	74.32	73.71	**82.83**	**77.38**	71.80	50.56	68.82	**87.15**	**78.95**	**83.79**	**84.52**	**84.15**	**84.08**	**89.53**	**95.19**	**92.27**	**87.85**
Qwen3-VL-4B-Instruct																			
Pre-CoT	**82.48**	80.58	**82.00**	**80.65**	77.33	80.50	**79.43**	81.96	**73.30**	76.61	77.80	86.31	86.76	86.53	86.50	**89.97**	91.57	90.76	85.79
Post-CoT	82.26	**80.79**	81.55	78.46	**86.51**	**82.00**	77.86	**87.79**	69.71	**77.37**	**78.84**	**88.89**	**88.61**	**88.75**	**88.77**	89.59	**95.85**	**92.62**	**88.35**

Table 2: Comparison between reason-first Pre-CoT and answer-first Post-CoT. Post-CoT is the default paradigm in WISE-GRPO. Metric definitions follow Table 3; bold indicates the best result per metric and backbone.

subsequent optimization further improves the decision boundaries learned from this initialization. The advantage is particularly clear on the multi-class QGR and ESCI benchmarks, where the policy must distinguish fine-grained relevance levels rather than make a binary match decision.

Table 2 compares the two output orders and shows that answer-first Post-CoT achieves higher overall accuracy than Pre-CoT in all eight backbone–benchmark settings. This result supports the use of an upfront answer as a stable decision anchor, while still allowing the reasoning trace to expose the evidence needed for prefix recovery. Finally, Table 4 shows that progressive weighting improves over uniform weighting in all eight settings, with an average gain of 1.48 accuracy points. The largest gains occur on QGR and ESCI, consistent with the motivation for assigning greater weight to later integrative steps.

6 What Evidence-Recovery Rewards Change

6.1 Gains Concentrate Beyond Lexical Shortcuts

If WISE-GRPO's gains come from cross-modal reasoning rather than surface matching, they should concentrate where surface matching fails. Using the Jaccard overlap $J(A, B) = |A \cap B|/|A \cup B|$ between the query and product text as a proxy for textual-shortcut strength, we partition each test set into five quantile bins of J and compute the accuracy gap Δ between WISE-GRPO and OR-GRPO within each bin.

Figure 5 supports this prediction. WISE-GRPO improves over OR-GRPO throughout the low- and mid-overlap bins, where lexical cues are weak or insufficient for reliable discrimination. The gains are most pronounced in the low-to-mid overlap regime, where lexical evidence is present but does not determine the label. By contrast, performance becomes slightly worse in the highest-overlap bin,

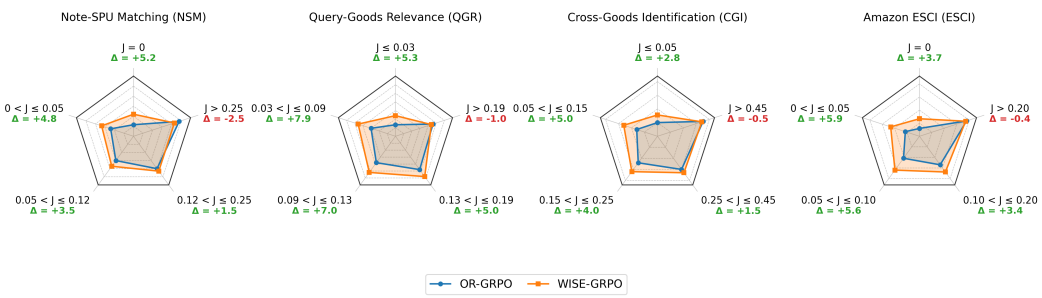


Figure 5: **Gains of WISE-GRPO over OR-GRPO concentrate where lexical shortcuts fail.** Accuracy delta (Δ) across quantile bins of query-product Jaccard overlap J. In the lowest-overlap bin gains range from $+2.8\%$ (CGI) to $+5.3\%$ (QGR), peaking at $+7.9\%$ (QGR) in low-to-mid overlap bins; in the highest-overlap bins the delta turns slightly negative—step-level supervision provides little additional value when lexical cues alone suffice.

Method	QGR						ESCI					CGI				NSM			
	$F1_0$	$F1_1$	$F1_2$	$F1_3$	$F1_4$	Acc.	F1(I)	F1(C)	F1(S)	F1(E)	Acc.	P	R	F1	Acc.	P	R	F1	Acc.
Qwen2.5-VL-3B-Instruct																			
SFT	67.90	59.66	57.90	57.67	70.86	62.84	63.71	38.83	60.61	79.36	65.26	72.97	73.69	73.33	73.20	80.93	87.22	83.96	74.60
RL	73.00	65.01	62.15	63.55	75.80	67.96	71.37	49.78	67.66	85.13	72.71	76.46	77.29	76.87	76.75	85.31	90.32	87.74	80.77
SFT+RL	**79.16**	**76.19**	**74.32**	**73.71**	**82.83**	**77.38**	**71.80**	**50.56**	**68.82**	**87.15**	**78.95**	**83.79**	**84.52**	**84.15**	**84.08**	**89.53**	**95.19**	**92.27**	**87.85**
Qwen3-VL-4B-Instruct																			
SFT	73.78	65.48	62.66	64.26	77.13	68.66	67.16	62.88	58.47	69.36	64.46	77.43	78.20	77.81	77.70	81.95	89.13	85.39	76.76
RL	79.05	71.73	69.58	70.38	80.68	74.32	74.71	69.94	67.25	76.84	72.20	81.49	82.30	81.89	81.80	85.96	91.17	88.49	81.93
SFT+RL	**82.26**	**80.79**	**81.55**	**78.46**	**86.51**	**82.00**	**77.86**	**87.79**	**69.71**	**77.37**	**78.84**	**88.89**	**88.61**	**88.75**	**88.77**	**89.59**	**95.85**	**92.62**	**88.35**

Table 3: Ablation study on the two-stage training pipeline. "RL" denotes RL-only training without the CoT SFT cold start, "SFT" denotes CoT SFT-only training, and "SFT+RL" denotes the complete pipeline. For QGR, $F1_k$ denotes class-k F1. For ESCI, I, C, S, and E denote Irrelevant, Complementary, Substitute, and Exact, respectively. For CGI and NSM, P/R/F1 are reported on the positive classes. Bold indicates the best result per metric and backbone.

where lexical matching alone often makes the label readily identifiable and leaves little room for step-level supervision to help.

6.2 Suppressing Forced Associations

The climb of step-level rewards during training (Figure 6) suggests the policy's reasoning is converging toward the committee's expected evidence organization. To test whether this reshapes inference-time behavior, we quantify *forced association*—the reward-hacking behavior in which the policy uses hedging rhetoric to bridge a logical gap to the target label without evidence support. Let R_t be the response set at training checkpoint t and $A(r)$ the set of forced-association phrases extracted from response r by an LLM-as-a-Judge (Li et al., 2024). We report two complementary metrics:

$$\mathrm{FAR}_t = \frac{\sum_{r \in R_t} \mathbf{1}[|A(r)| > 0]}{|R_t|}, \qquad \mathrm{FAD}_t = \frac{1000 \cdot \sum_{r \in R_t} |A(r)|}{\sum_{r \in R_t} |r|}. \tag{3}$$

Both metrics decline substantially for WISE-GRPO as training progresses, while the OR-GRPO baseline remains persistently high (App. C). The policy-entropy trajectories tell a compatible story: the baseline's entropy decreases while its FAR stays elevated—convergence to a shortcut-reliant strategy; WISE-GRPO shows a simultaneous decline in both entropy and hallucination frequency—convergence toward evidence-grounded reasoning (Jin et al., 2026).

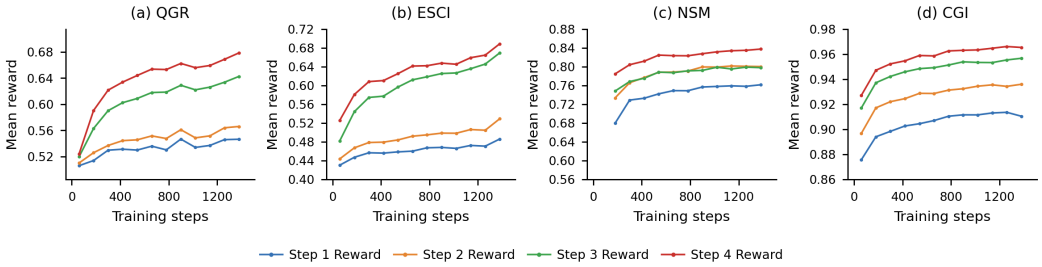


Figure 6: **Step-level rewards climb continuously during WISE-GRPO training on all four benchmarks**, indicating gradual alignment of the policy's reasoning with the evidence organization recognized by the committee.

Weighting	QGR						ESCI					CGI				NSM			
	$F1_0$	$F1_1$	$F1_2$	$F1_3$	$F1_4$	Acc.	F1(I)	F1(C)	F1(S)	F1(E)	Acc.	P	R	F1	Acc.	P	R	F1	Acc.
Qwen2.5-VL-3B-Instruct																			
Uniform	77.37	75.90	72.63	**74.59**	75.98	75.30	70.33	**54.18**	67.15	85.42	77.10	82.48	83.79	83.13	83.00	89.41	93.48	91.40	86.60
Progressive	**79.16**	**76.19**	**74.32**	73.71	**82.83**	**77.38**	**71.80**	50.56	**68.82**	**87.15**	**78.95**	**83.79**	**84.52**	**84.15**	**84.08**	**89.53**	**95.19**	**92.27**	**87.85**
Qwen3-VL-4B-Instruct																			
Uniform	80.40	79.07	81.39	**79.04**	82.08	80.38	74.11	85.88	**71.08**	76.81	76.90	87.53	88.40	87.96	87.90	89.54	94.20	91.81	87.20
Progressive	**82.26**	**80.79**	**81.55**	78.46	**86.51**	**82.00**	**77.86**	**87.79**	69.71	**77.37**	**78.84**	**88.89**	**88.61**	**88.75**	**88.77**	**89.59**	**95.85**	**92.62**	**88.35**

Table 4: Ablation study of step-wise reward weighting. Uniform assigns equal weights across steps, whereas Progressive uses the proposed schedule. Metric definitions follow Table 3; boldface marks the best result for each metric and backbone.

7 Discussion and Conclusion

We introduced WISE-GRPO, an annotation-free process-reward framework for multimodal e-commerce relevance. Its central idea is to treat process supervision as evidence recovery: a reasoning prefix receives credit when it allows an independent evaluator, without access to the policy answer, to recover the correct label. This criterion turns the otherwise subjective question of whether a reasoning step is good into a concrete continuation test. Across four benchmarks and two multimodal backbones, WISE-GRPO achieves the highest accuracy among the evaluated methods and consistently improves over outcome-only GRPO without human step-level process annotation.

Our results suggest that useful process rewards need not be learned as static step classifiers. Instead, they can be induced from whether evidence accumulated in a prefix changes an evaluator's ability to complete the decision. The intact-prefix and random-prefix controls support this interpretation: recovery improves with task-relevant evidence, but not with fluent, template-matched text from unrelated examples. The resulting reward therefore targets a property that outcome-only rewards leave unspecified, namely whether the intermediate reasoning makes the final decision independently recoverable.

The evaluator analysis further highlights that scale alone is not sufficient for this role. Strong evaluators can solve an instance from their own priors and leave little room for prefix evidence, while a task-agnostic weak model can respond unreliably to partial evidence. The easy-data committee occupies a more useful operating region: it retains headroom for evidence while producing stable intermediate recovery signals. This finding suggests a broader design principle for verifier-based RL: evaluator capability should be matched to the information that a partial trajectory is expected to reveal, rather than simply maximized.

WISE-GRPO adds training-time verification over reasoning prefixes, while committee members are queried concurrently and inference uses only the policy model. The framework is especially natural for discriminative tasks with a small, verifiable label space, but its underlying perspective is broader. Future work can develop richer recovery criteria for open-ended generation and adapt evaluator committees to new domains with different decision policies.

AI USE STATEMENT

In this work, we used generative AI models to construct synthetic structured reasoning traces for training and to support the implementation and analysis of our experiments. We also used generative AI tools for language editing and for parts of the experimental code. All AI-assisted outputs, including generated training data, code, text, and figures, were reviewed and validated by the authors. The authors take full responsibility for the final content of this work, including all claims, experimental results, and artifacts produced with the aid of generative AI.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Xinquan Chen, Chongying Yue, Bangwei Liu, Xuhong Wang, Yingchun Wang, and Chaochao Lu. Vrprm: Process reward modeling via visual reasoning, 2026. URL https://arxiv.org/abs/2508.03556.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

Chenhe Dong, Shaowei Yao, Pengkun Jiao, Jianhui Yang, Yiming Jin, Zerui Huang, Xiaojiang Zhou, Dan Ou, Haihong Tang, and Bo Zheng. Taosr1: The thinking model for e-commerce relevance search, 2026. URL https://arxiv.org/abs/2508.12365.

Pengkun Jiao, Yiming Jin, Jianhui Yang, Chenhe Dong, Zerui Huang, Shaowei Yao, Xiaojiang Zhou, Dan Ou, and Haihong Tang. She: Stepwise hybrid examination reinforcement learning framework for e-commerce search relevance, 2026. URL https://arxiv.org/abs/2510.07972.

Renren Jin, Pengzhi Gao, Yuqi Ren, Zhuowen Han, Tongxuan Zhang, Wuwei Huang, Wei Liu, Jian Luan, and Deyi Xiong. Revisiting entropy in reinforcement learning for large reasoning models, 2026. URL https://arxiv.org/abs/2511.05993.

Muhammad Khalifa, Rishabh Agarwal, Lajanugen Logeswaran, Jaekyeom Kim, Hao Peng, Moontae Lee, Honglak Lee, and Lu Wang. Process reward models that think, 2025. URL https://arxiv.org/abs/2504.16828.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. Llms-as-judges: A comprehensive survey on llm-based evaluation methods, 2024. URL https://arxiv.org/abs/2412.05579.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step, 2023. URL https://arxiv.org/abs/2305.20050.

Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Meiqi Guo, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, Jiao Sun, and Abhinav Rastogi. Improve mathematical reasoning in language models by automated process supervision. *arXiv preprint arXiv:2406.06592*, 2024. URL https://arxiv.org/abs/2406.06592.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. URL https://arxiv.org/abs/2305.18290.

Chandan K. Reddy, Lluís Màrquez, Fran Valero, Nikhil Rao, Hugo Zaragoza, Sambaran Bandyopadhyay, Arnab Biswas, Anlu Xing, and Karthik Subbian. Shopping queries dataset: A large-scale esci benchmark for improving product search, 2022. URL https://arxiv.org/abs/2206.06588.

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. Rewarding progress: Scaling automated process verifiers for llm reasoning, 2024. URL https://arxiv.org/abs/2410.08146.

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.

Tian Tang, Zhixing Tian, Zhenyu Zhu, Chenyang Wang, Haiqing Hu, Guoyu Tang, Lin Liu, and Sulong Xu. Lref: A novel llm-based relevance framework for e-commerce search. In *Companion Proceedings of the ACM on Web Conference 2025*, WWW '25, pp. 468–475. ACM, May 2025. doi: 10.1145/3701716.3715246. URL http://dx.doi.org/10.1145/3701716.3715246.

Peiyi Wang, Lei Li, Liang Chen, Dingyue Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. *arXiv preprint arXiv:2312.08935*, 2023. URL https://arxiv.org/abs/2312.08935.

Jianhui Yang, Yiming Jin, Pengkun Jiao, Chenhe Dong, Zerui Huang, Shaowei Yao, Xiaojiang Zhou, Dan Ou, and Haihong Tang. Taosr-agrl: Adaptive guided reinforcement learning framework for e-commerce search relevance, 2026. URL https://arxiv.org/abs/2510.08048.

Lifan Yuan, Wendi Li, Huayu Chen, Ganqu Cui, Ning Ding, Kaiyan Zhang, Bowen Zhou, Zhiyuan Liu, and Hao Peng. Free process rewards without process labels. *arXiv preprint arXiv:2412.01981*, 2024.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. Generative verifiers: Reward modeling as next-token prediction. In *International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=YKBUKp8H3Y.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. The lessons of developing process reward models in mathematical reasoning, 2025b. URL https://arxiv.org/abs/2501.07301.

A Committee Composition and Evaluator Strength

This appendix provides the full diagnostics summarized in §4: standalone accuracy of each committee member, committee-size and diversity statistics, and per-length recovery curves for evaluators of different strength.

Standalone accuracy and committee size. Table 5 reports each weak model's no-prefix accuracy ($M{=}1$) and the majority vote ($M{=}3$). Every member's standalone accuracy is far below its full-prefix ceiling (e.g., $\sim$59% vs. 99.95% on QGR), confirming genuine reliance on prefix evidence. The gain of $M{=}3$ over the single-model average is concentrated in evidence-scarce settings; the core conclusions (random $\approx$ chance, correct $\approx$ ceiling) hold regardless of M. We use $M{=}3$ rather than $M{=}2$ to avoid frequent ties under sparse evidence: no-prefix tie rates under $M{=}2$ reach $33.9/40.3/40.5\%$ on ESCI/NSM/QGR, versus a no-majority rate of only $6.3/8.1\%$ (ESCI/QGR) under $M{=}3$.

Table 5: Standalone (no-prefix) accuracy (%) of the three weak models and their $M{=}3$ majority vote.

Dataset	RM1	RM2	RM3	$M{=}1$ avg	$M{=}3$ vote
ESCI	47.16	46.82	45.87	46.62	48.53
CGI	79.77	78.59	78.59	78.98	78.28
NSM	63.88	57.15	56.90	59.31	60.25
QGR	53.97	55.82	59.09	56.29	59.37

Committee diversity. Table 6 reports multi-rater Fleiss' κ under the no-prefix condition, along with accuracy when the three members agree versus disagree. κ ranges from 0.17 to 0.73, far from identical copies, and the disagreement is systematic: on NSM, one member favors "match" ($\sim$68%) while the other two favor "non-match" ($\sim$69–74%). This complementarity, induced by disjoint easy-data training, is what majority voting exploits.

Table 6: Committee diversity under the no-prefix condition.

Dataset	Fleiss' κ	Acc. when agreeing	Acc. when disagreeing
ESCI	0.507	58.5%	37.6%
CGI	0.728	88.3%	52.1%
NSM	0.172	72.3%	52.3%
QGR	0.488	74.9%	47.5%

Per-length recovery under different evaluator strength. Table 7 gives the per-length curves underlying Figure 3. Qwen3.5-0.8B is non-monotone in prefix length (e.g., $24.8 \to 21.0$ on QGR and $96.9 \to 95.2$ on CGI from $k{=}1$ to $k{=}2$), while our committee responds smoothly to additional evidence; its near-ceiling CGI curve saturates, which is the expected signature of a solved task rather than instability.

Table 7: Recovery accuracy (%) at each prefix length for Qwen3.5-0.8B and our committee.

Task	**Evaluator**	$k{=}1$	$k{=}2$	$k{=}3$	**Full**
ESCI	Qwen3.5-0.8B	29.80	33.90	48.90	99.60
	Committee (ours)	50.03	55.72	73.00	99.56
CGI	Qwen3.5-0.8B	96.90	95.20	99.50	100.00
	Committee (ours)	98.60	98.93	98.89	98.80
NSM	Qwen3.5-0.8B	62.30	79.90	94.80	98.10
	Committee (ours)	61.40	79.91	89.33	98.59
QGR	Qwen3.5-0.8B	24.80	21.00	34.40	99.90
	Committee (ours)	58.66	61.72	74.70	99.95

B Analytical Discussion of Recovery Behavior

This appendix expands the lightweight design characterizations used in the main text. They are design analyses rather than formal guarantees of downstream policy improvement.

Why later steps deserve larger weights. Along a correctly structured trace, later prefixes cannot contain less label-relevant information than their predecessors, since each additional step can only add evidence. This motivates, but does not mathematically force, progressive weighting: our task template places evidence extraction early and business-rule integration late, so later steps are more directly connected to the final judgment. A finite, imperfect committee may still produce locally non-monotone empirical accuracy, as CGI's saturated curve illustrates.

Evaluator operating region (§4.2). The step signal $\Delta_k = \hat{p}_k - p_0$ becomes small when a strong evaluator already solves the input without a prefix. Conversely, a task-agnostic weak evaluator

can have a large endpoint gain but an unstable intermediate curve. The desired operating region is therefore empirical rather than universal: p_0 should leave headroom, p_K should be high, and the intermediate curve should respond smoothly to relevant evidence. Table 7 checks these three properties directly.

Committee variance. Treating each member's step reward as a random variable with equal marginal variance σ^2 and average pairwise correlation $\bar{\rho}$, the variance of the M-member average is $\sigma^2[1 + (M-1)\bar{\rho}]/M$. Independent errors ($\bar{\rho} = 0$) recover the usual σ^2/M rate; positive correlation creates an asymptotic floor. Disjoint training subsets aim to reduce correlation, and the observed $\kappa \in [0.17, 0.73]$ confirms non-identical behavior. Our choice $M = 3$ is motivated mainly by tie avoidance and inference cost, not by a claim of optimality.

C Forced Association Evaluation: Full Details

This section supplements the forced-association evaluation in §6.2, including metric definitions, judge model setting, extraction prompt, and alignment dynamics.

C.1 Definitions

In RL with sparse outcome rewards, models are susceptible to reward hacking via *forced association*—the behavior where a model, upon encountering attribute conflicts in multimodal evidence, bypasses objective negation or downgrading and uses hedging or compromising rhetoric to bridge a gap to the target label. Let R_t denote the response set at training checkpoint t, $A(r)$ the phrases extracted from r by the judge, and $|r|$ the total word count of r. FAR (breadth) and FAD (density per 1,000 words) are defined in Eq. 4–5.

$$\mathrm{FAR}_t = \frac{\sum_{r \in R_t} \mathbf{1}\big[|A(r)| > 0\big]}{|R_t|} \tag{4}$$

$$\mathrm{FAD}_t = \frac{1000 \cdot \sum_{r \in R_t} |A(r)|}{\sum_{r \in R_t} |r|} \tag{5}$$

C.2 Judge model and evaluation setup

We employ Gemini-3-Flash-Preview as the semantic extraction judge to avoid bias from a judge drawn from the same model family as the policies under evaluation. Temperature is set to 0.0 and outputs restricted to strict JSON. Both compared models generate full reasoning rollouts at selected checkpoints, followed by independent extraction. We validate the judge against human annotation on 200 sampled cases (97% agreement) and confirm the extraction is stable across two different judge LLMs.

C.3 Extraction prompt

Unlike approaches that rely on large models for subjective scoring, we frame the evaluation as an objective text-span matching task by requiring the judge to extract verbatim quotes.

System Prompt for Forced Association Extraction

Please act as a rigorous language analysis expert and read the thinking process of a large language model performing a product matching/relevance analysis task.

Your task is: Extract all core phrases or half-sentences that constitute forced association.

When the model discovers obvious inconsistencies in core features between two entities, it may use compromising, ambiguous, or forcefully rationalizing expressions to arrive at a relevant or matching conclusion. Typical feature words include but are not limited to: *to a certain extent, arguably, roughly matches, has some overlap, not exactly the same but, barely acceptable*, etc.

Extraction Requirements:

1. It must be a verbatim quote from the original text without altering a single word.
2. Do not extract entire long sentences; only extract the core phrase or half-sentence containing the compromising or ambiguous semantics, recommended length is between 5 and 20 characters.
3. If the original text merely describes objective facts or ultimately reaches a mismatched conclusion, do not extract anything.
4. If no such phrases are found, output an empty list `[]`.

Thinking process:
{thinking_content}

Please output strictly valid JSON without any other text or markdown tags, in the format:
{"phrases": ["phrase1", "phrase2"]}

C.4 Alignment dynamics

A decrease in policy entropy does not by itself imply the acquisition of evidence-grounded reasoning. For the baseline, policy entropy smoothly decreases, yet the forced-association rate rises to 64.3% in the late stages of the soft relevance judgment task, indicating that sparse binary rewards drive the model to internalize pseudo-logic as a high-certainty strategy.

WISE-GRPO instead exhibits a phased alignment trajectory (Figure 7). In tasks with soft relevance judgments, the model first probes legitimate decision boundaries, briefly entering a high-entropy exploration phase where the forced-association rate slightly exceeds the baseline. As process rewards accumulate, the model completes logic internalization, leading to a simultaneous cliff-like drop in both forced-association rate and policy entropy (an absolute reduction of 12.1% at checkpoint 1200). In strict binary matching tasks, forced associations immediately trigger reasoning errors in the weak-evaluator committee, so process rewards take effect without an exploration phase, achieving suppression over the baseline throughout training with a relative reduction of 36.7%.

C.5 Qualitative case study

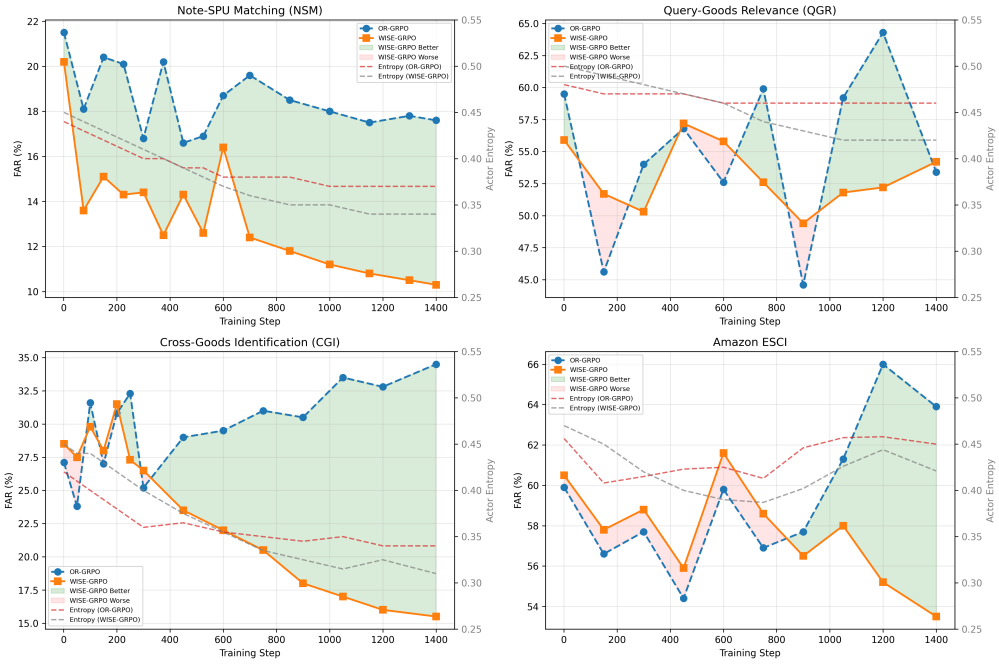
To concretize these dynamics, Table 8 dissects a hard sample that repeatedly triggered forced associations. The baseline answered incorrectly 7 times on this instance, whereas WISE-GRPO consistently predicted correctly. Under the strict binary outcome-reward environment, the baseline overfits superficial textual overlap and applies a compromising concession clause in the final deduction. WISE-GRPO establishes a preemptive negation logic, explicitly stating that the item does not match, is inconsistent, and is not reflected—an uncompromising chain of evidence that stably guides the committee to a correct negative judgment.

D Dataset Details

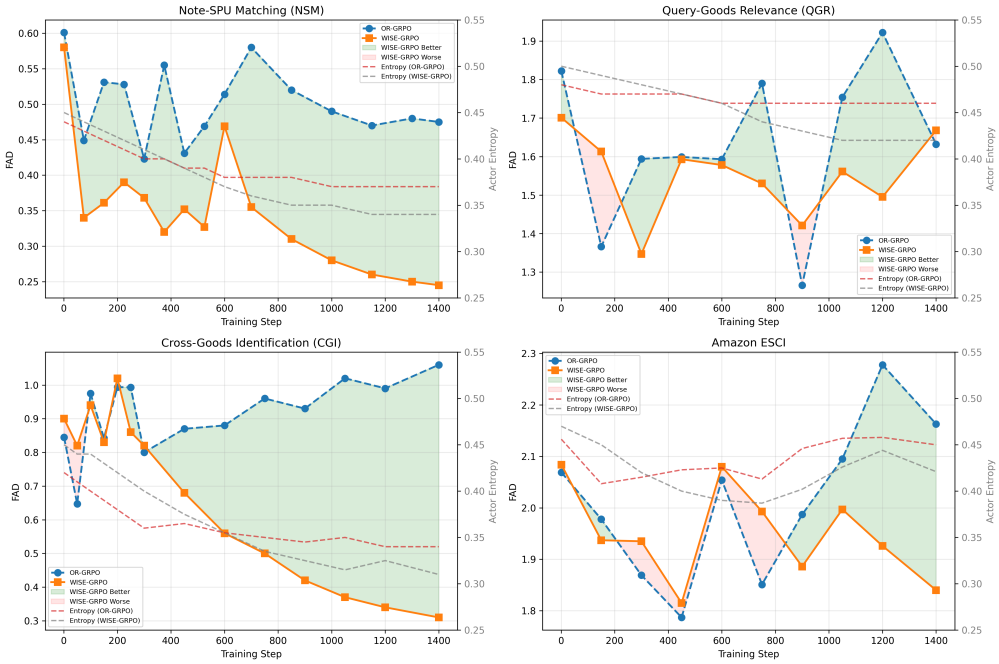
The four benchmarks and one representative example per benchmark are shown in Table 9–10. All take multimodal (image + text) inputs. ESCI is a public benchmark released by Amazon; QGR, NSM, and CGI are proprietary industrial datasets from a large social-commerce platform.

E Prompt Templates

This appendix provides the prompt templates used in the three training stages described in Section 3.3. All prompts were originally written in Chinese and are presented here in English translation. The three prompt types share identical user inputs and differ only in output format: Section E.1 uses answer-conditioned distillation, Section E.2 uses answer-first Post-CoT, and Section E.3 uses think-first Pre-CoT.



(a) Evolution of Forced Association Rate (FAR)



(b) Evolution of Forced Association Density (FAD)

Figure 7: **WISE-GRPO drops forced associations and entropy together; the baseline drops entropy but keeps hallucinating.** FAR and FAD trajectories with policy entropy across four benchmarks.

E.1 Oracle-Guided CoT Distillation Prompts

The oracle receives the correct label and generates a coherent four-step reasoning chain without mentioning the correct label in Steps 1 through 3. We present the QGR prompt as a representative example; for the other three tasks, only the label definitions differ.

Table 8: Case study comparing OR-GRPO and WISE-GRPO on a hard instance. The baseline overfits textual overlap and hedges its final deduction; WISE-GRPO issues an uncompromising negation chain.

Input Context	**Query:** Decorte water-emulsion perilla *(Exact match intent)* **Item:** Decorte perilla makeup primer isolation cream 3-in-1 oil-control moisturizing, brand wwp, 40g **Ground Truth:** Completely Irrelevant (Tier 0)
Baseline: OR-GRPO Predict: Tier 2 ✗	**Step 1:** The query indicates the user wants a water-emulsion product with perilla ingredients... **Step 2:** The title mentions Decorte perilla isolation cream, brand wwp... **Step 3:** The brand Decorte partially matches the title; the product contains perilla...isolation cream belongs to the primer category, which has some correlation... **Step 4:** ***Although the brand name differs slightly, it can partially satisfy the user's needs overall.*** ...Therefore it is moderately relevant.
Ours: WISE-GRPO Predict: Tier 0 ✓	**Step 1:** The user is looking for Decorte brand perilla series products, a water-emulsion set... **Step 2:** The title is Decorte perilla isolation cream, brand wwp... **Step 3:** The product brand wwp **does not match** the Decorte in the query; the product is isolation cream while the user searches for a water-emulsion set, so the categories are **inconsistent**; perilla-related descriptions are **not reflected** in the title... **Step 4:** The product **does not match in brand, category, and function**...Completely irrelevant.

Table 9: Overview of the four benchmarks.

Benchmark	Type	Labels	Key Challenge
ESCI	Public, Amazon	E / S / C / I	Category and functional boundary
QGR	Proprietary	Tier 0–4	Soft multi-level grading
NSM	Proprietary	Match / Non-match	Promotional intent recognition
CGI	Proprietary	Match / Non-match	Core entity attribute comparison

System Prompt

Role: Senior e-commerce relevance expert.

Task: Given a query, a product, and the correct relevance tier, generate a coherent four-step reasoning process. Do not mention the correct tier in Steps 1 through 3.

Label Definitions:

Tier 4: Exact match across category, brand, and attributes.
Tier 3: Highly relevant with minor discrepancies.
Tier 2: Moderately relevant, partially satisfies the query.
Tier 1: Marginally relevant, limited reference value.
Tier 0: Irrelevant.

Output Format:

```
<think>
Step 1: Identify query intent
Step 2: Extract key product attributes
Step 3: Assess relevance
Step 4: Derive the tier with justification
</think>
[Tier label]
```

Table 10: One representative example per benchmark, with images replaced by textual descriptions.

Benchmark	Input Fields	Label
ESCI	**Query:** polka dot skirt **Title:** Vfshow Womens Black and White Polka Dot Print Ruched Ruffle High Waist Pencil Midi Skirt **Brand:** Vfshow **Image:** one product image	Exact
QGR	**Query:** yellow necklace **Title:** CATHY JEWELRY Premium 5-Carat Rectangular Yellow Diamond Necklace **Category:** Necklace **Image:** one product image	Tier 4
NSM	**SPU:** Green Grape Pulp Drink, Local Life category **SPU Image:** one beverage product image **Note Title:** a summer tasting of green grape **Note Body (excerpt):** a chilled Green Grape Pulp Drink became my cool companion in the park... **Note Image:** one note lifestyle image	Match
CGI	**Product A:** Butterfly Star Cat Claw photocard, Collectible Cards category, one photocard image **Product B:** Winter Karina photocard Cat Claw special edition, Collectible Cards category, one photocard image	Non-match

User Prompt

Query: {query}
Product Title: {title}
[Task-specific fields]

The correct answer is [{correct_label}].
Justify in detail why this label applies.

E.2 Policy Model Prompts (Post-CoT)

The policy model uses an answer-first Post-CoT format throughout both the CoT SFT and GRPO stages, so the final prediction acts as an upfront decision anchor.

System Prompt

You are an e-commerce domain expert. Generate your response in the following answer-first format:

<answer> [label] </answer>
<think> [reasoning] </think>

The <think> section should present your step-by-step analysis. The <answer> section should contain the concise final label.

User Prompt

[ESCI]
<image>
Query: {query}
Title: {title} Brand: {brand}
Assign one label: E (Exact) / S (Substitute) / C (Complement) / I (Irrelevant).

[QGR]
<image>
Query: {query}
Title: {title} Category: {category}
SKU: {sku_info}
Assign one tier: Tier 0 / 1 / 2 / 3 / 4.

[NSM]
SPU Image: <image> SPU: {spu_name} Category: {spu_category}
Note Image: <image> Note Title: {note_title}
Note Body: {note_body}
Output Match or Non-match.

[CGI]
Product A: <image> Title: {title_A}
Brand: {brand_A} Category: {category_A}
Product B: <image> Title: {title_B}
Brand: {brand_B} Category: {category_B}
Output Match or Non-match.

E.3 Weak-Model Committee Prompts (Pre-CoT)

The committee shares the same user prompts as the policy model. The only difference is the system prompt, which enforces a think-first output format.

System Prompt

You are an e-commerce domain expert. Generate your response in the following think-first format:

<think> [reasoning] </think>
<answer> [label] </answer>

The <think> section should present your step-by-step analysis. The <answer> section should contain the concise final label.

During GRPO training, for each step k the policy reasoning trace is truncated and sent to the committee as a prefix-continuation prompt:

Prefix-Continuation Format (Step k)

[user prompt]
<think>
Step 1: $\{s_1\}$
...
Step k: $\{s_k\}$

Truncated here. The closing </think> *and* <answer> *tags are withheld, forcing the committee to independently resume reasoning and produce a label.*

F Training and Inference Details

We use DeepSpeed ZeRO-Stage 2 with FlashAttention-2 and bfloat16 mixed-precision for all training stages. For SFT and reward-model (RM) training, the learning rate is 1×10^{-5} with a cosine scheduler and a warmup ratio of 0.1. We train for 2 epochs with a global batch size of 64, using the Liger kernel for efficiency.

For the RL stage (WISE-GRPO), we use vLLM for trajectory rollout, allocating 70% of GPU memory (`gpu_memory_utilization=0.7`) to the inference engine. The learning rate is 2×10^{-6} with a global batch size of 64, a per-device batch size of 2, and gradient accumulation of 4. The rollout top-p is 1.0 and the group size G is 8. WISE-GRPO takes about $3.7\times$ the wall-clock training time of OR-GRPO (113.5h vs. 30.6h) and about $1.2\times$ that of MR-GRPO; since committee members are queried concurrently, the overhead is dominated by the sequential per-step verification rather than committee size.

Table 11: Detailed hyperparameters used in our experiments.

Stage	**Hyperparameter**	**Value**
SFT & RM	Learning rate	1×10^{-5}
	LR scheduler	Cosine
	Warmup ratio	0.1
	Global batch size	64
	Gradient accumulation steps	2
	Training epochs	2
	Cutoff length (tokens)	16,384
RL (WISE-GRPO)	Learning rate	2×10^{-6}
	Global batch size	64
	Per-device batch size	2
	Gradient accumulation steps	4
	Training epochs	1
	Max prompt length (tokens)	8,192
	Max response length (tokens)	4,096
	vLLM GPU utilization	0.7
	Rollout top-p	1.0
	Rollout group size G	8