INTERNALIZING RUBRIC GUIDANCE FOR RESEARCH AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

A research rubric specifies what an answer should cover, but using it only as a scalar reward leaves its instructional content unused. We study how to turn this content into better research agents that act without rubric hints. We introduce rubric-guided self-distillation (RSD): criterion-level feedback selects weak queries, the current policy generates new tool trajectories with targeted rubric guidance, and a reward threshold filters the resulting demonstrations. The agent then learns these trajectories under the original, unhinted prompt in supervised phases interleaved with fresh-policy reinforcement learning. No separate demonstration model is required. On Qwen3-8B, RSD improves over rubric-based GRPO by 14.23, 7.61, and 2.84 score points on ResearchQA, DeepResearch Bench, and DeepResearch Bench II, with paired 95% task-bootstrap confidence intervals excluding zero on all three benchmarks. The method also improves LSB from 26.04 to 28.79. At 14B and 32B, it outperforms both starting models and size-matched GRPO on all four research benchmarks, with RQA gains over GRPO of 15.07 and 14.47 points. Strategy comparisons highlight rubric-guided demonstration learning as a major contributor to the gains, and RSD achieves the highest scores on all four research benchmarks among the evaluated 8B configurations. These results show that rubrics can provide not only rewards but also demonstrations that improve research performance without rubric hints at inference.

1 INTRODUCTION

A research agent must decide what to search for, connect evidence across sources, and produce an answer that addresses a user's question. Success rarely reduces to an exact-match label. Rubric-based reinforcement learning (RL) addresses this difficulty by evaluating task-specific criteria and combining their verdicts into a reward (Gunjal et al., 2025; Mei et al., 2026; Shao et al., 2025). Yet the rubric contains more information than its aggregate score. Two answers can receive the same reward while missing entirely different requirements: a comparison, a quantitative argument, or a crucial qualification. Scalar feedback ranks sampled behaviors; the criterion text can also help construct behaviors that were not sampled.

This distinction suggests using the rubric during generation. The difficulty is that assistance changes the input: a policy that succeeds with detailed coverage requirements need not succeed when given only the user's question. Rubric scaffolding and criterion-conditioned distillation address this mismatch in different ways (Zhou et al., 2025; Xia et al., 2026). For research agents, the object to transfer is a complete sequence of searches, observations, and answers. The practical question is therefore: *how can training-only rubric information become supervision for an agent that acts without it?*

Our approach separates two uses of the same rubric. In the RL branch, the agent receives the original query and explores without hints; the rubric judge supplies group-relative rewards. In the demonstration branch, criterion-level feedback identifies weak queries, and the same policy retries them with targeted rubric guidance. High-scoring trajectories become supervised targets, but the hint is removed from their input before learning. This directly trains the policy to reproduce useful research actions under the original prompt.

We call this procedure *rubric-guided self-distillation* (RSD). Accepted demonstrations accumulate during an RL block and are learned in a dedicated, multi-epoch supervised fine-tuning (SFT) phase.

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.

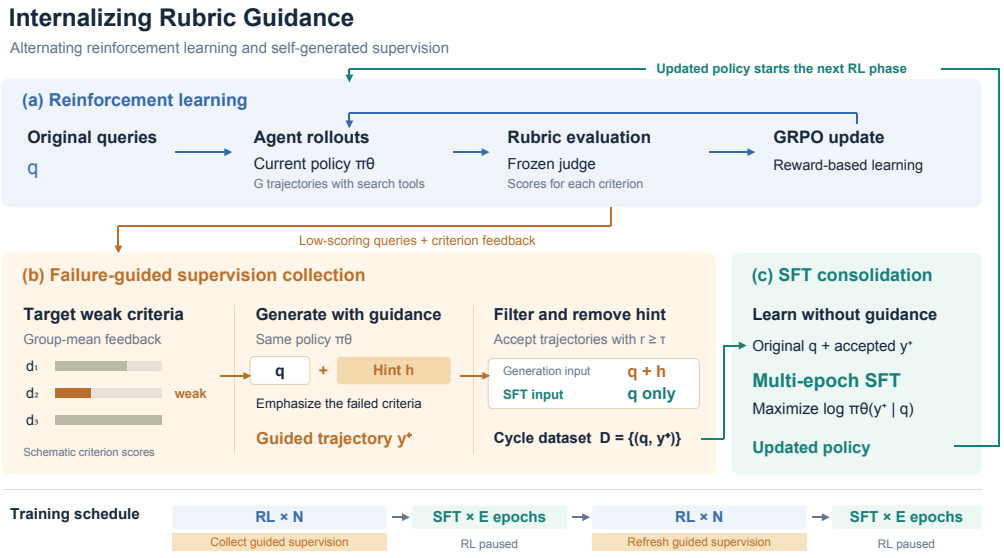

Figure 1: **Use rubric guidance to generate supervision, then learn without the hint.** (a) Fresh, unhinted trajectories drive GRPO. (b) Criterion feedback guides query selection and assisted resampling; accepted trajectories enter a cycle dataset with their original prompt restored. (c) Dedicated SFT phases learn assistant actions and synchronize the updated policy before RL resumes. All symbols follow Section 2; N and E denote RL-block length and SFT epochs. Evaluation uses original queries without rubric hints.

The updated policy then resumes unhinted RL and supplies the next cycle's demonstrations. Figure 1 illustrates the two data paths and their shared policy. A separate demonstration model is unnecessary: the rubric supplies privileged training information, while an independent judge still evaluates the generated answers.

Our design is motivated by exploratory runs in which hinted generations score 0.086–0.244 higher than fresh generations, without comparable improvement in subsequent unhinted behavior (Appendix C.1). This gap between assisted generation and unhinted behavior motivates a dedicated learning step. During dedicated SFT phases, the policy learns to reproduce accepted research trajectories from the original prompt, without rubric hints.

Building on self-training, context distillation, and interleaved RL–SFT (Zelikman et al., 2022; Askell et al., 2021; Ma et al., 2026; Wang et al., 2026), RSD repeatedly uses criterion-level feedback to guide the current policy toward better research trajectories. It filters these trajectories by reward and learns them under the original prompt in dedicated SFT phases. Section 4 compares RSD with the closely related CriPO and ReLIFT methods.

We evaluate RSD on unhinted research performance, the role of demonstration learning, and transfer across model sizes. At 8B, RSD gains 14.23, 7.61, and 2.84 points over GRPO on ResearchQA, DeepResearch Bench, and DeepResearch Bench II, with paired 95% task-bootstrap intervals excluding zero on all three benchmarks. LSB improves from 26.04 to 28.79. Strategy comparisons highlight the value of learning complete rubric-guided research trajectories under the original prompt. At 14B and 32B, RSD improves over both the respective starting models and size-matched GRPO on all four research benchmarks; RQA gains over GRPO are 15.07 and 14.47 points.

A unified transfer analysis explains how target quality, imitation under the original input, and query coverage interact in the learning cycle (Section 2; Appendix H).

2 RUBRIC-GUIDED SELF-DISTILLATION

2.1 LEARNING WITH AND WITHOUT RUBRIC ASSISTANCE

Let q denote the original system and user messages, $C_q = \{d_k\}_{k=1}^{K_q}$ a task-specific rubric, and y a multi-turn trajectory containing assistant actions and tool observations. A fixed, independent judge assigns binary criterion scores $s_k(q, y) \in \{0, 1\}$ to the final answer. The rubric score is

$$r_{\text{rub}}(q, y) = \frac{1}{K_q} \sum_{k=1}^{K_q} s_k(q, y). \tag{1}$$

The scalar training reward r equals r_{rub}, halved for truncated trajectories.

For each fresh query, we sample G trajectories without hints and compute group-relative advantages,

$$A_i = \frac{r_i - \overline{r}}{\text{std}(r_1, \ldots, r_G) + \epsilon}. \tag{2}$$

The RL branch uses the clipped GRPO objective (Shao et al., 2024), with a KL penalty to the reference policy and entropy regularization. Fresh, unhinted trajectories supply the RL updates; assisted demonstrations supply the supervised targets.

A hint h changes generation from $\pi_\theta(y \mid q)$ to $\pi_\theta(y \mid q, h)$. RSD samples demonstrations from the assisted distribution but optimizes their likelihood under the original messages. This matches the deployment prompt, while teacher forcing still conditions on demonstration prefixes.

2.2 FROM CRITERION FEEDBACK TO ACCEPTED TRAJECTORIES

Query selection. For each trajectory in a fresh group, define

$$p_i = \frac{1}{K_q} \sum_k s_{ik} - \lambda \sum_k \mathbb{I}[s_{ik} < \tau_f], \qquad p_q = \min_i p_i. \tag{3}$$

We enqueue a query when $p_q < \tau_q$ and prioritize smaller values. The experiments use $\lambda = 0.1$ and $\tau_f = \tau_q = 0.8$. The queue stores queries rather than individual answers and has capacity 100. The unnormalized failure count gives additional priority to queries with more unmet criteria.

Hint construction. The group-level profile $\overline{s}_k = G^{-1} \sum_i s_{ik}$ includes every answer in the group, not just its worst answer. We build a coverage-oriented hint from this profile and the original rubric text. Criteria with $\overline{s}_k < 0.7$ receive additional emphasis; available criterion details are included without grading-level headings. The hint does not print the numeric scores. It is appended to the original user message. This path uses a deterministic rubric-to-hint transformation rather than an additional model that diagnoses the failure. The rubric may contain task-specific knowledge, so the assisted generator has privileged training information even though its parameters match the learner's.

Resampling and acceptance. Each RL round targets eight queued queries in addition to 16 fresh queries. The fresh branch generates eight trajectories per query, while the assisted branch generates two. The current rollout policy generates each assisted trajectory in the same tool environment. We keep trajectories with scalar reward at least $\tau = 0.8$. Trajectories below the threshold are discarded. A query has a replay limit of one in these runs. Accepted demonstrations y^+ accumulate in a cycle-local dataset $D = \{(q, y^+)\}$; the queue need not supply its full target when insufficient eligible queries are available.

2.3 LEARNING TRAJECTORIES UNDER THE ORIGINAL PROMPT

For an accepted pair (q, y^+), let $y^{+,-h}$ denote the trajectory with the added hint removed from the input messages. We retain all subsequent assistant actions and tool observations. The supervised objective is

$$\mathcal{L}_{\text{SFT}}(\theta; D) = -\frac{1}{|D|} \sum_{(q,y^+) \in D} \frac{1}{|\mathcal{A}(y^+)|} \sum_{t \in \mathcal{A}(y^+)} \log \pi_\theta(a_t \mid q, y^{+,-h}_{<t}), \tag{4}$$

Rubric-guided self-distillation (RSD)

Input: policy θ, queries and rubrics, phase boundaries $\mathcal{T}$, queue Q; $D \leftarrow \varnothing$.

1. For each RL round t, sample fresh queries and G unhinted trajectories per query.
2. Score criteria and update θ with fresh-only GRPO, Eq. 2.
3. Aggregate group feedback and enqueue eligible weak queries using Eq. 3.
4. Resample available queued queries under rubric hints; accept trajectories with $r \geq \tau$ and append their original-prompt versions to D.
5. If $t \in \mathcal{T}$: finish the cycle's in-flight collection; freeze D; initialize an independent SFT optimizer; minimize Eq. 4; synchronize updated parameters; clear D.

Output: the policy evaluated on original messages, without rubric hints.

Steps 1–4 describe data dependencies, not a strict synchronous ordering; queue work overlaps rollout. $\mathcal{T} = \{19, 39\}$ at 8B and $\{19, 39, 56\}$ at 14B/32B.

Algorithm 1: The demonstration branch supplies supervised targets without changing the conditioning of fresh RL updates. SFT is skipped when the accepted set is empty.

where $\mathcal{A}(y^+)$ contains supervised assistant tokens. System, user, and tool-observation tokens provide context but receive no direct supervised loss. Tool calls are assistant actions and are included in the target. Distributed training computes this loss over the supervised assistant tokens.

What is transferred. The supervised target includes both information gathering and answer construction. At a search action, the learner predicts the demonstrated tool call from the original query and preceding interaction; at an answer action, it predicts the response using the observations already present in that interaction. Tool results are therefore evidence available to the learner, rather than text it is asked to generate. Removing the hint changes the initial conditioning while retaining this action–observation structure. The learning signal asks the policy to recover useful research decisions from the query and interaction history alone. At evaluation, those decisions must produce a new trajectory through the live tool environment.

After a block of $N = 20$ RL rounds, the system freezes the accumulated demonstrations and performs up to $E = 8$ SFT epochs. SFT has its own optimizer and learning-rate schedule; it does not add a loss term to the RL optimizer step. After consolidation, the updated model parameters are synchronized to the rollout engine and the RL optimizer's master weights. The RL optimizer state persists; the SFT optimizer is recreated at each consolidation phase. Consumed demonstrations are cleared before collecting the next cycle.

The cycle boundary also defines the lifetime of a demonstration. Collection can overlap learner updates, so accepted trajectories may come from several policy versions within the same block. Before SFT, the system completes in-flight collection and freezes the cycle dataset. This provides a stable target set for repeated optimization. Clearing that set after consolidation makes subsequent supervision depend on the resumed policy's newly observed weaknesses and newly generated trajectories. Thus the recurring cycle couples fresh exploration with an updated source of supervised targets.

The 8B run consolidates after zero-indexed rounds 19 and 39, then continues RL through round 56. The 14B/32B runs also consolidate at round 56. Algorithm 1 specifies these schedules; Appendix A gives phase sizes and update counts.

2.4 A UNIFIED VIEW: TARGET QUALITY AND REALIZED TRANSFER

We distinguish the quality of assisted targets from the quality of the policy learned from them. For a bounded reward $R(q, y) \in [0, 1]$, define

$$J_\mu(v) = \mathbb{E}_{q\sim\mu,\, y\sim v(\cdot|q)} R(q, y). \tag{5}$$

Here v is a trajectory distribution in the evaluation environment. The analysis fixes the accepted target distribution for one learning cycle, matching the frozen dataset used during consolidation.

Let ν denote the accepted demonstration distribution, p the learned unhinted policy, p_0 its reference policy, S the selected-query distribution, and μ the evaluation distribution. For a common bounded

reward and environment, Appendix H derives

$$J_\mu(p) - J_\mu(p_0) \geq \underbrace{J_S(\nu) - J_S(p_0)}_{\text{demonstration advantage}} - \underbrace{\sqrt{\mathbb{E}_S \,\mathrm{KL}(\nu\|p)/2}}_{\text{imitation gap}} - \underbrace{2\,\mathrm{TV}(\mu, S)}_{\text{query shift}}. \tag{6}$$

Why the three terms arise. On a selected query, the reward difference between the target and student is bounded by their total variation distance. Pinsker's inequality bounds this distance by $\sqrt{\mathrm{KL}(\nu\|p)/2}$; averaging over selected queries gives the imitation penalty in Eq. 6. Moving the policy improvement from S to μ contributes twice the query-distribution total variation because the improvement lies in $[-1, 1]$. This yields the bound by first comparing the student with its target and then comparing training-query coverage with evaluation-query coverage.

Implications for the training cycle. A high-scoring assisted trajectory supplies a useful target when its advantage survives both penalties. This separates the roles of collection and consolidation: rubric guidance and acceptance determine what behavior is available to imitate, while original-prompt SFT determines how that behavior is learned. Query selection additionally determines where supervision is concentrated. Repeating the cycle changes the selected targets as the policy changes, whereas each individual SFT phase optimizes a fixed collection. The empirical comparisons assess these choices through the resulting unhinted policies.

The bound connects the learning cycle to unhinted transfer: guidance supplies advantageous targets, consolidation reduces imitation error, and query coverage determines where the learned behavior applies. Appendix H derives the bound, relates it to token-normalized SFT, and extends it to differences in reward and environment.

3 Experiments

3.1 Setup and reporting protocol

Training. We study Qwen3 models at 8B, 14B, and 32B (Yang et al., 2025). The training collection contains 926 retained queries from an original 999-query corpus, paired with GPT-5-generated research rubrics from the data pipeline of Mei et al. (2026). Each full RL run supplies 57 rounds of 16 fresh query groups with eight trajectories per group: 912 query-group visits and 7,296 fresh trajectories. Training uses a local embedding retrieval service over 11,976 cached search records; evaluation uses live search. Appendix A gives optimization and generation settings, and Appendix B describes data provenance.

Comparisons. The primary reference is rubric-based GRPO trained on fresh, unhinted trajectories. We also evaluate task adaptations of CriPO (Xia et al., 2026), APTER (Wang et al., 2026), and ReLIFT (Ma et al., 2026). These in-framework adaptations share the research-agent task environment. Appendix C describes the baseline configurations; APTER and ReLIFT request external DeepSeek-R1 demonstrations. ReLIFT-32 denotes the adaptation with a 32-demonstration SFT trigger. We separately compare joint RL–SFT with online replay (Fused-online) or a frozen demonstration bank (Fused-bank), SFT-only, and random query selection. All results use the generic agent evaluation interface without rubric hints.

Evaluation. The research suite consists of ResearchQA (RQA; 400 sampled queries), DeepResearch Bench (DRB; all 100 tasks), DeepResearch Bench II (DRB-II; all 132 tasks), and LocalSearchBench (LSB; 400 sampled queries) (Yifei et al., 2025; Du et al., 2025; Li et al., 2026; He et al., 2025). We report the rubric-coverage, RACE overall, DRB-II overall, and LSB composite scores, respectively. The secondary evaluation uses 400 sampled GPQA-main questions (Rein et al., 2023). This multiple-choice evaluation uses GPQA-main through the same agent harness. Generation is greedy, with thinking disabled and a 2,048-token limit per model call. Task-specific evaluation rubrics are available to the judge, not to the acting policy.

We average per-task judge scores over all input tasks, assigning empty generations zero. All scores use our common agent harness. We compare research performance on RQA, DRB, DRB-II, and LSB, with GPQA as a secondary evaluation. Appendix D details the scoring procedures.

Table 1: **Research gains without rubric hints at inference.** Scores are on a $\times 100$ scale (higher is better); empty answers score zero. GPQA is a secondary agent-harness evaluation. All trained rows use 57 fresh RL rounds, but total compute and supervised exposure differ. ReLIFT-32 executes two supervised updates during training. Bold denotes column maxima among these configurations.

Method	RQA	DRB	DRB-II	LSB	GPQA
Qwen3-8B	44.83	21.36	13.40	25.87	25.50
GRPO	45.12	21.95	14.60	26.04	31.00
CriPO	45.23	23.17	14.54	25.81	26.00
APTER	45.06	21.58	15.23	26.10	26.50
ReLIFT-32	45.27	21.93	14.27	25.91	27.50
RSD-8B	**59.35**	**29.56**	**17.45**	**28.79**	**34.25**

Table 2: **Rubric-guided demonstrations drive strong unhinted performance.** SFT-only learns rubric-guided demonstrations without RL; random selection changes the query-selection rule. Fused-bank matches the 8B RSD demonstration budget of 379 trajectories with eight exposures each, using joint RL–SFT updates. Scores use the same all-input aggregation as Table 1; Appendix A specifies evaluation checkpoints.

Method	RQA	DRB	DRB-II	LSB	GPQA
GRPO	45.12	21.95	14.60	26.04	31.00
Fused-online	45.67	22.18	11.10	25.83	28.25
Fused-bank	48.41	23.10	13.94	25.84	29.00
SFT-only	56.49	27.02	16.19	27.43	30.25
Random selection	59.08	29.09	15.50	27.83	32.25
RSD-8B	**59.35**	**29.56**	**17.45**	**28.79**	**34.25**

3.2 Does rubric guidance improve unhinted performance?

RSD improves over GRPO on all four research benchmarks (Table 1). RQA increases from 45.12 to 59.35, DRB from 21.95 to 29.56, DRB-II from 14.60 to 17.45, and LSB from 26.04 to 28.79.

We quantify task-sampling uncertainty using 10,000 paired bootstrap resamples of task-level scores. These intervals measure task-sampling uncertainty, conditional on the trained checkpoints, judge outputs, and search results. Appendix E reports intervals for the other comparisons.

Magnitude and uncertainty. The paired 95% intervals for the gains over GRPO are [12.83, 15.61] on RQA, [5.81, 9.46] on DRB, and [1.13, 4.66] on DRB-II. All three lower endpoints are positive. The absolute gains differ across benchmarks because their scoring procedures and task sets differ; we interpret each within its own benchmark. Pairing preserves which tasks are easy or difficult for both checkpoints, and zero-scored empty answers remain in the resampling population. These comparisons directly evaluate the intended transfer: both policies act on the original queries, while rubric information is confined to training supervision and evaluation judging.

RSD also exceeds the evaluated CriPO, APTER, and ReLIFT-32 adaptations on all four research benchmarks. APTER executes ten SFT phases using 346 selected demonstrations. ReLIFT-32 performs one SFT update on 32 demonstrations at each of rounds 25 and 52. Appendix C details these configurations and the 64-demonstration trigger variant.

3.3 Which training strategies account for the gains?

Joint RL–SFT with online replay. Fused-online jointly optimizes $\mathcal{L}_{\mathrm{GRPO}} + 1.0\,\mathcal{L}_{\mathrm{SFT}}$ using online replay (Table 2). It underperforms RSD on RQA, DRB, DRB-II, and LSB. The fixed-bank comparison below further examines joint optimization with repeated supervised exposure.

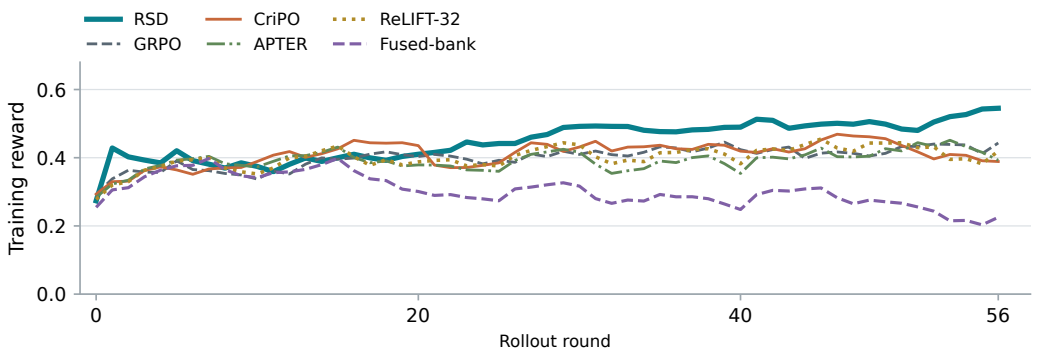

Figure 2: **Training reward at 8B.** Lines show trailing means over up to five consecutive rounds for GRPO, CriPO, APTER, ReLIFT-32, Fused-bank, and RSD. Each curve represents one training run.

Dedicated phases outperform the fixed-bank fused configuration. Fused-bank improves over pure GRPO by 3.29 RQA points (paired bootstrap 95% CI [2.19, 4.37]). Compared with Fused-bank, RSD gains 10.94 points on RQA, 6.46 on DRB, and 3.50 on DRB-II (Table 2; differences use unrounded scores). Fused-bank and RSD use the same 379 demonstrations with eight exposures per trajectory. Fused-bank distributes this supervision across 114 joint RL–SFT updates, using a persistent Adam optimizer at learning rate 5×10^{-7}. In RSD, demonstrations are collected within each learning cycle and learned in dedicated SFT phases with an independent optimizer and schedule. The comparison therefore contrasts joint and phased learning under a matched demonstration-exposure budget. Appendix C gives the optimization settings.

Demonstration learning captures substantial gains. SFT-only reaches 56.49 on RQA and 27.02 on DRB, substantially above GRPO. RSD has higher point estimates on all four research benchmarks and GPQA. The RQA gain is 2.86 points with a paired 95% task interval of [0.77, 4.92]; intervals for DRB and DRB-II span zero (Appendix E). Together, these results identify rubric-guided demonstration learning as a strong contributor, with an additional RQA gain for RSD under the reported collection schedules.

Demonstration learning with random query selection. Random selection obtains 59.08 on RQA and 29.09 on DRB. RSD achieves the highest scores across the four research benchmarks and GPQA in Table 2. These results show that rubric-guided demonstration learning is effective with a simple query-selection rule.

Reading the strategies together. Together, these comparisons suggest that learning from rubric-guided demonstrations accounts for much of the improvement. The evaluated variants improve our understanding of how query selection and the optimization schedule shape this learning process.

3.4 How does learning evolve across rounds and model sizes?

Figure 2 shows training rewards for six configurations. Appendix G adds strategy, scale, and exploratory curves. The curves track reward evolution over changing training queries; external evaluations measure the resulting unhinted performance.

At 14B and 32B, RSD outperforms both the starting models and size-matched GRPO on RQA, DRB, DRB-II, LSB, and GPQA (Table 3). Relative to GRPO, RQA/DRB/DRB-II gains are 15.07/9.72/1.37 points at 14B and 14.47/8.64/3.78 at 32B; GPQA gains are 6.00 and 3.00 points. This consistent cross-scale advantage supports using RSD beyond 8B. The strongest scale depends on the task: 14B leads on RQA and DRB, while 32B has the highest GPQA accuracy.

At 8B, GPQA rises from 31.00 to 34.25 relative to GRPO. This secondary evaluation complements the research suite by testing multiple-choice answers through the same agent interface.

Table 3: **Gains over size-matched GRPO across Qwen3 scales.** All models are evaluated without rubric hints; empty outputs count as zero. GRPO-14B/32B use checkpoint 53; 8B GRPO and RSD use round 56. Larger RSD runs include a final SFT phase at round 56. Bold denotes column maxima.

Model	ResearchQA	DRB	DRB-II	LSB	GPQA
Qwen3-8B	44.83	21.36	13.40	25.87	25.50
GRPO	45.12	21.95	14.60	26.04	31.00
RSD-8B	59.35	29.56	**17.45**	28.79	34.25
Qwen3-14B	47.19	22.66	15.38	25.89	30.00
GRPO-14B	47.92	23.00	15.60	25.68	33.75
RSD-14B	**62.99**	**32.71**	16.97	**29.02**	39.75
Qwen3-32B	46.42	23.35	14.65	26.86	38.50
GRPO-32B	47.12	21.95	13.27	26.44	39.50
RSD-32B	61.58	30.59	17.05	27.74	**42.50**

3.5 Training schedules and computational cost

At 8B, two SFT phases learn 379 trajectories in total, which also supply the demonstration bank for Fused-bank. The 14B and 32B runs add a final phase after round 56. Appendix A, Table 6, reports phase sizes and executed update counts.

Allocation-based training cost is 69.2 GPU-hours for RSD-8B versus 48.3 for GRPO, a difference of 20.9 GPU-hours, or approximately 43%. These figures cover the allocated GPU tasks and exclude external judge and search costs (Appendix F). Fresh-branch generation produces approximately 20.45M response tokens for RSD and 14.09M for GRPO.

3.6 Case study: connecting components to system operation

Table 4 compares saved GRPO and RSD-8B trajectories on a ResearchQA question, with no rubric hints given to either agent. Both call search twice. GRPO describes the components and basic heat-transfer process; RSD explicitly searches for thermal storage and control strategies, then explains hot-water priority, preheating, and backup heating. The evaluator assigns higher coverage to component interactions and coordinated heat delivery. This example illustrates how more specific research actions accompany a more integrated answer under the original prompt.

4 Related Work

Rubrics as supervision. Rubrics as Rewards extends reward-based training beyond tasks with exact verification (Gunjal et al., 2025). Deep Research as Rubric constructs evaluation criteria through research and evidence synthesis (Mei et al., 2026); DR Tulu evolves rubrics for deep-research training (Shao et al., 2025). Our work assumes an existing task-specific rubric and studies how its feedback is consumed by the learner.

Guided exploration and criterion-level distillation. RuscaRL supplies rubric-derived guidance during rollout and gradually reduces it (Zhou et al., 2025). CriPO directly studies the mismatch between rubric-guided generation and unassisted inference, and distinguishes unexplored criteria from criteria whose signals are suppressed by scalar advantages (Xia et al., 2026). Where CriPO uses localized self-distillation and token-level advantage modification, RSD supervises complete research trajectories, including tool calls and answer construction. It generates and filters assisted trajectories, restores the original prompt, and learns the retained assistant actions in dedicated supervised phases.

Self-training and context distillation. STaR iteratively fine-tunes on successful self-generated rationales, using the correct answer to help generate rationales for previously failed questions (Zelikman et al., 2022). Context distillation transfers behavior elicited by a prompt into model parameters

Query: What are the main components of a solar combisystem, and how does it function to provide both space heating and domestic hot water?

GRPO	**Score: 65.63**	**RSD-8B**	**Score: 75.00**
Search. "main components of a solar combisystem and how it functions for space heating and domestic hot water"		**Search.** "solar combisystem components function architecture thermal storage control strategies"	
Answer. Lists collectors, a storage tank, a circulation pump, and controls.		**Answer.** Connects collectors and a buffer tank through an antifreeze loop.	
Describes heating a fluid, transferring heat to the tank, and supplying hot water or space heating.		Adds hot-water priority, domestic-water preheating, and backup boilers or heat pumps; discusses storage, sizing, and freeze protection.	
Criterion coverage. Component interaction: *Moderately*. Coordinated heat delivery: *Moderately*.		**Criterion coverage.** Component interaction: *Mostly*. Coordinated heat delivery: *Mostly*.	

Table 4: **From component listing to integrated system explanation.** Saved 8B trajectories on ResearchQA (index 20, zero-based). Searches are verbatim; answers are summarized. Orange highlights general heat transfer; green highlights operational detail. Two of eight criterion judgments are shown; scores average all eight. Both agents search twice without rubric hints.

(Askell et al., 2021). RSD applies these principles to rubric-guided, multi-turn research trajectories, learning search actions and answer construction under the original prompt.

Interleaved fine-tuning and learning from feedback. ReLIFT interleaves RL with demonstrations for the hardest questions (Ma et al., 2026). APTER uses persistent rubric-defined deficiencies to trigger targeted supervised updates (Wang et al., 2026). RSD uses the current policy itself to generate demonstrations, with targeted guidance from the existing task rubric. Accepted demonstrations preserve the complete tool interaction and are learned under the original, unhinted prompt in dedicated SFT phases interleaved with fresh-policy RL.

5 Limitations and Broader Impact

Experimental scope. Task-bootstrap intervals quantify task-sampling uncertainty for the evaluated checkpoints. Training schedules and optimization settings are specified in Section 3 and Appendix C.

Evaluation scope. Automatic rubric judges and live search measure research-answer quality under the reported harness; human factuality assessment and a comprehensive semantic contamination audit remain open. Consequential deployment requires complementary evidence and citation checking.

6 Conclusion

RSD turns criterion-level rubric feedback into supervision for research agents: the current policy generates guided trajectories, reward filtering selects demonstrations, and dedicated supervised phases learn their actions under the original prompt. The resulting agents outperform size-matched GRPO on RQA, DRB, DRB-II, and LSB at 8B, 14B, and 32B without rubric hints at inference. Strategy comparisons identify demonstration learning as a strong contributor, while the transfer analysis connects target quality to imitation under the deployed input and query coverage. Together, these findings establish a practical use of rubrics beyond scalar rewards: constructing research trajectories that improve unassisted performance.

AI USE STATEMENT

Generative AI assisted manuscript organization, drafting, editing, LaTeX preparation, and the implementation of table and figure generation scripts. The roles of language models in rubric generation, agent training, judging, and demonstration generation are described in Sections 2–3 and Appendix C.

REPRODUCIBILITY STATEMENT

Section 2 specifies the algorithm. Appendices A–E describe training schedules, baseline configurations, data and prompts, evaluation protocols, and uncertainty estimation. The accompanying manuscript sources include scripts for reproducing tables and figures from experiment outputs, machine-readable scores, paired bootstrap results, and source-file hashes.

Code availability. Code is available in the anonymous repository at https://anonymous.4open.science/r/review-artifact-rsd-20260923-0B7F/README.md.

REFERENCES

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. A General Language Assistant as a Laboratory for Alignment. *arXiv preprint arXiv:2112.00861*, 2021. URL https://arxiv.org/abs/2112.00861.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. DeepResearch Bench: A Comprehensive Benchmark for Deep Research Agents. *arXiv preprint arXiv:2506.11763*, 2025. URL https://arxiv.org/abs/2506.11763.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. Rubrics as Rewards: Reinforcement Learning Beyond Verifiable Domains. *arXiv preprint arXiv:2507.17746*, 2025. URL https://arxiv.org/abs/2507.17746.

Hang He, Chuhuai Yue, Chengqi Dong, Mingxue Tian, Hao Chen, Zhenfeng Liu, Jiajun Chai, Xiaohan Wang, Yufei Zhang, Qun Liao, Guojun Yin, Wei Lin, Chengcheng Wan, Haiying Sun, and Ting Su. Localsearchbench: Benchmarking agentic search in real-world local life services. *arXiv preprint arXiv:2512.07436*, 2025. URL https://arxiv.org/abs/2512.07436.

Ruizhe Li, Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. DeepResearch Bench II: Diagnosing Deep Research Agents via Rubrics from Expert Reports. *arXiv preprint arXiv:2601.08536*, 2026. URL https://arxiv.org/abs/2601.08536.

Lu Ma, Hao Liang, Meiyi Qiang, Lexiang Tang, Xiaochen Ma, Zhen Wong, Junbo Niu, Chengyu Shen, Runming He, Yanhao Li, Wentao Zhang, and Bin Cui. Learning What Reinforcement Learning Can't: Interleaved Online Fine-Tuning for Hardest Questions. In *International Conference on Learning Representations*, pp. 80802–80821, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/82c4ad0c3d3d4898a77be5df7b2ede3a-Paper-Conference.pdf.

Wangyi Mei, Zhouhong Gu, Zhenhan Bai, Yin Cai, Lefan Zhang, Zhenxin Ding, Bo Chen, Yan Gao, Yi Wu, Yao Hu, Jiaqing Liang, and Deqing Yang. Deep Research as Rubric for Reinforcement Learning. *arXiv preprint arXiv:2606.01091*, 2026. URL https://arxiv.org/abs/2606.01091.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. *arXiv preprint arXiv:2311.12022*, 2023. URL https://arxiv.org/abs/2311.12022.

Rulin Shao, Akari Asai, Shannon Zejiang Shen, Hamish Ivison, Varsha Kishore, Jingming Zhuo, Xinran Zhao, Molly Park, Samuel G. Finlayson, David Sontag, Tyler Murray, Sewon Min, Pradeep Dasigi, Luca Soldaini, Faeze Brahman, Wen-tau Yih, Tongshuang Wu, Luke Zettlemoyer, Yoon Kim, Hannaneh Hajishirzi, and Pang Wei Koh. DR Tulu: Reinforcement Learning with Evolving Rubrics for Deep Research. *arXiv preprint arXiv:2511.19399*, 2025. URL https://arxiv.org/abs/2511.19399.

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. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Xukai Wang, Liangqi Li, Zhiyue Xu, Jingang Zhou, Xiaoyu Shi, Jiansheng Cai, Bo Zhang, Zhe Li, and Xu-Yao Zhang. APTER: Adaptive Post-Training with Expert-Grounded Rubrics. *arXiv preprint arXiv:2608.14212*, 2026. URL https://arxiv.org/abs/2608.14212.

Mingxuan Xia, Yuhang Yang, Chao Ye, Shuai Zhu, Shenzhi Yang, Guangcheng Zhu, Yuhang Zhang, Cheng Peng, Haobo Wang, and Siqing Wang. CriPO: Enhancing Rubric-based RL via Self-Distillation. *arXiv preprint arXiv:2607.18082v3*, 2026. URL https://arxiv.org/abs/2607.18082v3.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Li S. Yifei, Allen Chang, Chaitanya Malaviya, and Mark Yatskar. ResearchQA: Evaluating Scholarly Question Answering at Scale Across 75 Fields with Survey-Mined Questions and Rubrics. *arXiv preprint arXiv:2509.00496*, 2025. URL https://arxiv.org/abs/2509.00496.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping Reasoning With Reasoning. *arXiv preprint arXiv:2203.14465*, 2022. URL https://arxiv.org/abs/2203.14465.

Yang Zhou, Sunzhu Li, Shunyu Liu, Wenkai Fang, Kongcheng Zhang, Jiale Zhao, Jingwen Yang, Yihe Zhou, Jianwei Lv, Tongya Zheng, Hengtong Lu, Wei Chen, Yan Xie, and Mingli Song. Breaking the Exploration Bottleneck: Rubric-Scaffolded Reinforcement Learning for General LLM Reasoning. *arXiv preprint arXiv:2508.16949*, 2025. URL https://arxiv.org/abs/2508.16949.

A Implementation and Training Schedules

A.1 Optimization settings

Table 5 summarizes the training configuration, including rollout generation, query selection, RL optimization, and supervised consolidation.

Table 5: Core training settings for the RSD pipeline. SFT settings describe configured targets; executed update counts are listed in Table 6.

Setting	Value
Fresh groups per RL round / trajectories per group	16 / 8
RL rounds / fresh trajectories	57 / 7,296
RL optimizer / learning rate	Adam / 5×10^{-7}, constant
RL Adam (β_1, β_2) / weight decay	$(0.9, 0.99)$ / 0.1
RL global batch / clipping bounds	32 / $[1 - 0.2, 1 + 0.28]$
Reference KL coefficient / entropy coefficient	0.005 / 10^{-4}
Rollout temperature / response / context limits	1.0 / 8,192 / 20,480 tokens
Replay queue capacity / replay limit per query	100 / 1
Target assisted queries per round / samples per query	8 / 2
Query-priority threshold / acceptance threshold	0.8 / 0.8
Weak-criterion emphasis threshold	0.7
RL rounds between SFT phases	20
Maximum SFT epochs / effective batch	8 / 16
SFT learning rate / configured minimum	10^{-5} / 10^{-6}, cosine
SFT Adam β_2	0.98
SFT early-stop loss / gradient-norm thresholds	0.05 / 0.4

The training reward is the mean of criterion verdicts, halved for truncated trajectories. The 14B/32B runs use `glm-5.3-flash` as the independent rubric judge. DeepSeek-R1 generates demonstrations for the APTER/ReLIFT adaptations. DRB-II uses `glm-5.3-flash` for evaluation.

A.2 Phase boundaries and evaluation checkpoints

The two 8B SFT phases use 111 and 268 accepted trajectories, with 56 and 136 optimizer updates, respectively (Table 6). Figure 2 shows the corresponding fresh-rollout training rewards.

The 8B full RL comparisons use iteration 56. The 14B/32B pure-GRPO evaluations use iteration 53, and their RSD counterparts use round 56. SFT-only is evaluated after its second SFT phase at optimizer step 135, following two collection cycles of 20 rounds each. Random selection is evaluated at iteration 19.

Table 6: **Demonstrations and updates in each SFT phase.** Rounds are zero-indexed. The table reports executed optimizer updates under the configured eight-epoch limit.

Model	Boundary round	Accepted trajectories	Batches/epoch	Optimizer updates
8B	19	111	7	56
8B	39	268	17	136
14B	19	131	9	72
14B	39	241	16	96
14B	56	206	13	91
32B	19	121	8	64
32B	39	259	17	102
32B	56	238	15	75

A.3 Asynchrony and trajectory restoration

The queue is bounded, and generation can overlap learner updates. As a result, the policy version that produces a demonstration need not be the version at the end of the cycle. SFT freezes the col-

lected trajectories and reuses them as supervised targets. At a phase boundary, the system seals the collection partitions, stops or drains speculative work as needed, runs SFT, synchronizes parameters, and resumes rollout. Fresh and replay collection track completion independently.

Restoration replaces the augmented prompt prefix with the original message sequence and retains the multi-turn response tail. Tool observations remain in the causal context; only assistant tokens receive direct supervised loss. These assisted trajectories supply off-policy supervised targets, preserving the assistant actions and their interaction context.

B Data, Prompts, and Tool Environment

The training corpus retains 926 of 999 original queries, each paired with GPT-5-generated research criteria. Redistribution is subject to source-data permissions.

The training retrieval corpus contains 11,976 cached search-query/result records, built from real Tavily results in the data pipeline. Filtering removes duplicate queries and queries shorter than 10 or longer than 500 characters. Retrieval uses a Qwen3-Embedding-0.6B service. Training uses this fixed local retrieval environment. External evaluation uses the configured live search service.

B.1 Acting and hint prompts

The dataset's system message defines a research agent with three tools: `plan`, `tavily_search`, and `summary`. Planning comes first, search supplies observations, and summary terminates the session with the answer. Tool invocations use `<tool_call>` delimiters and JSON arguments. The query-specific rubric is not part of the original user prompt.

Each rubric hint begins with the following instruction:

> To provide a comprehensive answer, thoroughly research and address each of the following aspects of the question. Some aspects are marked as requiring particularly detailed coverage—make sure your search and summary go deep on those.

The hint lists rubric dimensions and marks low-scoring dimensions with "important—ensure detailed coverage." Available criteria blocks appear under "Key requirements for this aspect," with grading-level headers removed. Full criteria provide coverage guidance across dimensions, and the emphasis marker directs additional attention to weak dimensions.

The original system/user messages are restored before SFT.

C Baseline Adaptations and Exploratory Analysis

Pure GRPO and fused training. Pure GRPO uses the rubric reward without replay. Fused-bank combines the fresh GRPO objective with supervised loss at coefficient 1.0, averaging each objective within an update. Its fixed bank contains the 379 trajectories collected for the 8B RSD SFT phases, matching their eight-exposure supervised budget. Eight exposures produce 3,032 trajectory uses and 6,105,768 supervised assistant tokens; training uses 7,296 fresh trajectories and 114 joint updates. Adam is persistent, with learning rate 5×10^{-7}, $(\beta_1, \beta_2) = (0.9, 0.99)$, and no phase reset or early stopping. The demonstration bank remains fixed throughout training. The comparison uses the RSD checkpoint, with its online collection and dedicated optimizer schedule.

CriPO. The CriPO adaptation (Xia et al., 2026) uses criterion-level diagnostics and advantage modification within the same research-agent environment. Binary leaf verdicts provide criterion-level feedback, while the trajectory reward is continuous after aggregation. The adaptation includes the unexplored-criterion (UC) loss. We report the final checkpoint after 57 rounds under the common evaluation protocol.

APTER and ReLIFT. Both adaptations use an external DeepSeek-R1 agent to generate complete planning/search/summary trajectories, and the rubric judge accepts a valid teacher demonstration when its aggregate rubric score is at least 0.8. This task-specific threshold differs from the all-leaf-pass rule used to diagnose student failures. The APTER adaptation maps query-specific rubric

leaves to 12 fixed capability categories, aggregates failure rates by category, and triggers targeted teacher requests for persistent weak categories. Its teacher receives privileged information: the target weakness, failed criteria, a representative failure, full rubric criteria, and diagnostics. It uses category failure threshold 0.75, up to eight weak categories, and eight accumulated trigger observations, with a minimum of eight demonstrations and a maximum batch of 64. Its run executes ten SFT phases using 346 selected trajectories.

ReLIFT requests help when none of a query's eight fresh answers passes all criteria. Teachers use the aggregate-score threshold above, and accepted demonstrations are deduplicated by query within the pending pool. ReLIFT-32 starts from the original model and triggers SFT when 32 demonstrations have accumulated. It completes 57 rounds, selecting 32 demonstrations at round 25 and another 32 at round 52; each selection receives one SFT update at learning rate 10^{-6}. The configured batch limit remains 64, but the actual batch at each trigger is 32. The total of 64 counts demonstration uses across the two phases.

The teacher adaptations use one SFT epoch with learning rate 10^{-6}, weight decay 0.01, and Adam $\beta_2 = 0.999$, plus an entropy term of coefficient 10^{-4}. They differ from RSD both in who produces the demonstrations and how strongly demonstrations are optimized. Their acceptance rules and supervised exposure are reported alongside the results.

Table 7: Additional training configurations. Fused-online, also included in Table 2, uses online replay supervision. ReLIFT-64 accumulates 54 demonstrations, below its 64-demonstration SFT trigger. Each row represents a separate training run.

Training configuration	RQA	DRB	DRB-II	LSB	GPQA
Fused-online	**45.67**	22.18	11.10	25.83	28.25
ReLIFT-64	45.46	**22.79**	**14.68**	**26.01**	**30.00**

C.1 Exploratory assistance–internalization diagnostics

We examine three exploratory training strategies: failure replay, separate fresh/hinted groups, and mixed fresh/hinted groups. Table 8 summarizes assisted and unhinted generation under these strategies.

Table 8: Exploratory rollout diagnostics. "Hint reference" measures lexical references to guidance. The last column is the proportion of hinted or mixed groups with rubric-score standard deviation below 10^{-6}.

Exploratory variant	Fresh score	Hinted score	Hint reference	Zero-variance groups
Failure replay	0.403	0.648	82.9%	2.1%
Separate groups	0.385	0.507	91.8%	6.1%
Mixed groups	0.402	0.488	91.6%	0.1%

The failure-replay study reports a subsequent unhinted pre/post difference of 0.042 over 1,456 matched observations. The unit is a matched observation, with repeated queries represented. In the separate-group run, descriptive fitted changes over 50 rounds are 0.259 for hinted scores and 0.041 for fresh scores. Assisted scores, unhinted scores, and references to guidance measure distinct aspects of the trajectories.

These exploratory findings motivate the central evaluation choice: measure learned performance using the original, unhinted input.

D Evaluation Protocol and Score Aggregation

D.1 Scores and generation settings

ResearchQA reports coverage averaged across the query's rubric items and then across tasks. DRB reports the RACE overall score, which evaluates report quality against reference material. FACT citation metrics are outside the reported evaluation. DRB-II reports the evaluator's overall rubric score

together with information-recall, analysis, and presentation dimensions. LSB reports the composite $(c+(u+f+h+s)/40)/2$, where correctness c is binary and completeness u, fluency f, faithfulness h, and safety s are scored on a 0–10 scale. Task scores follow each evaluator's weighting rules.

The checkpoint evaluation configuration specifies temperature 0, top-p 1, top-k 1, seed 42, repetition penalty 1.15, thinking disabled, a maximum of 2,048 generated tokens per call, and a 90-second call timeout. The token limit applies independently to each call within a multi-turn task. Agent answers are extracted from the summary tool output.

D.2 ALL-INPUT AGGREGATION

Let N be the number of input tasks, I the subset with nonempty answers, and s_i the corresponding task score. We report

$$S_{\text{all}} = \frac{100}{N} \sum_{i \in I} s_i = 100 \, \frac{|I|}{N} \, S_{\text{saved}}, \tag{7}$$

where S_{saved} is the mean score over nonempty answers on a 0–1 scale. Empty answers contribute zero to the full-input mean. GPQA accuracy is the number of correct answers divided by N. DRB-II directly averages over all input tasks, including empty answers.

Table 9: Examples of all-input score aggregation. "Scored" counts nonempty answers; the all-input score averages over all N tasks, assigning empty answers zero.

Method	Benchmark	Scored / N	Correct	Saved score	All-input score
GRPO	GPQA	372 / 400	124	33.33	31.00
GRPO-14B	GPQA	381 / 400	135	35.43	33.75
GRPO-14B	DRB	96 / 100	—	23.95	23.00
RSD-8B	GPQA	392 / 400	137	34.95	34.25
Fused-bank	GPQA	360 / 400	116	32.22	29.00
ReLIFT-32	GPQA	370 / 400	110	29.73	27.50
GRPO	DRB	98 / 100	—	22.40	21.95
RSD-14B	DRB	99 / 100	—	33.04	32.71

D.3 ADDITIONAL EVALUATION DETAILS

Table 10 collects LSB results across training strategies and model sizes.

DRB-II retains the evaluator's task-level scores, including criterion-level blocked judgments. The 8B RSD aggregate reports a blocked rate of approximately 0.00437. Generation uses the common 2,048-token-per-call, live-search protocol specified above.

E PAIRED TASK UNCERTAINTY

For each comparator and each of RQA, DRB, and DRB-II, we align task scores using stable IDs, except that DRB is aligned by exact prompt text. RSD-8B has a full record for every task in these three benchmarks. Each comparison pairs scores on the same input tasks, assigning empty outputs zero. We sample N paired task differences with replacement, repeat this 10,000 times using pseudorandom seed 20260922, and report the 2.5th and 97.5th percentiles of the mean difference.

F TRAINING COMPUTATION

Training uses 48.3 task-level GPU-hours for GRPO, 69.2 for RSD-8B, and 75.3 for CriPO on 16 allocated GPUs, with effective step-time estimates of 47.4, 68.3, and 74.4 GPU-hours. These allocation-based measurements cover GPU training tasks; search and judging use external services.

Fresh rollouts generate approximately 14.09M generated response tokens for GRPO, 20.45M for RSD, and 14.56M for CriPO. These counts cover fresh-branch generation. Each full run contains 7,296 fresh trajectories across 912 groups of eight samples.

Table 10: **LSB results across training strategies and model sizes.** Scores use the composite metric defined in Appendix D; higher is better.

Method	LSB
Qwen3-8B	25.87
GRPO	26.04
CriPO	25.81
APTER	26.10
ReLIFT-32	25.91
Fused-bank	25.84
ReLIFT-64	26.01
Fused-online	25.83
SFT-only	27.43
Random selection	27.83
RSD-8B	28.79
Qwen3-14B	25.89
GRPO-14B	25.68
RSD-14B	**29.02**
Qwen3-32B	26.86
GRPO-32B	26.44
RSD-32B	27.74

Table 11: RSD-8B minus each comparator, in score points. Intervals describe task resampling conditional on the evaluated models, outputs, and judges. Each is a pointwise 95% paired task-bootstrap interval.

Comparator	Benchmark	Difference	95% task interval
GRPO	ResearchQA	+14.23	[+12.83, +15.61]
GRPO	DRB	+7.61	[+5.81, +9.46]
GRPO	DRB-II	+2.84	[+1.13, +4.66]
CriPO	ResearchQA	+14.12	[+12.65, +15.57]
CriPO	DRB	+6.39	[+4.57, +8.12]
CriPO	DRB-II	+2.91	[+1.13, +4.78]
Fused-bank	ResearchQA	+10.94	[+9.50, +12.39]
Fused-bank	DRB	+6.46	[+4.66, +8.24]
Fused-bank	DRB-II	+3.50	[+1.68, +5.39]
SFT-only	ResearchQA	+2.86	[+0.77, +4.92]
SFT-only	DRB	+2.54	[-0.26, +5.45]
SFT-only	DRB-II	+1.25	[-0.64, +3.17]
APTER	ResearchQA	+14.30	[+12.89, +15.72]
APTER	DRB	+7.98	[+6.14, +9.87]
APTER	DRB-II	+2.22	[+0.49, +3.99]
ReLIFT-32	ResearchQA	+14.08	[+12.53, +15.61]
ReLIFT-32	DRB	+7.63	[+6.08, +9.17]
ReLIFT-32	DRB-II	+3.17	[+1.41, +4.90]

G Training Dynamics Across Methods

We compare training dynamics across 23 runs. The main comparisons include GRPO, CriPO, APTER, ReLIFT-32, and RSD; the remaining panels cover fused optimization, random selection, SFT-only, larger models, and exploratory variants. Faint lines show raw observations and solid lines average up to five consecutive steps. Smoothing restarts after gaps, and isolated observations appear as points. Random selection is shown through its evaluated checkpoint at step 19. Loss curves use optimizer steps, not rollout rounds.

Training reward is averaged over fresh trajectories on each run's sampled queries. Table 3 reports the corresponding 14B/32B GRPO checkpoint evaluations; those evaluations use checkpoint 53, while the training curves extend through round 56.

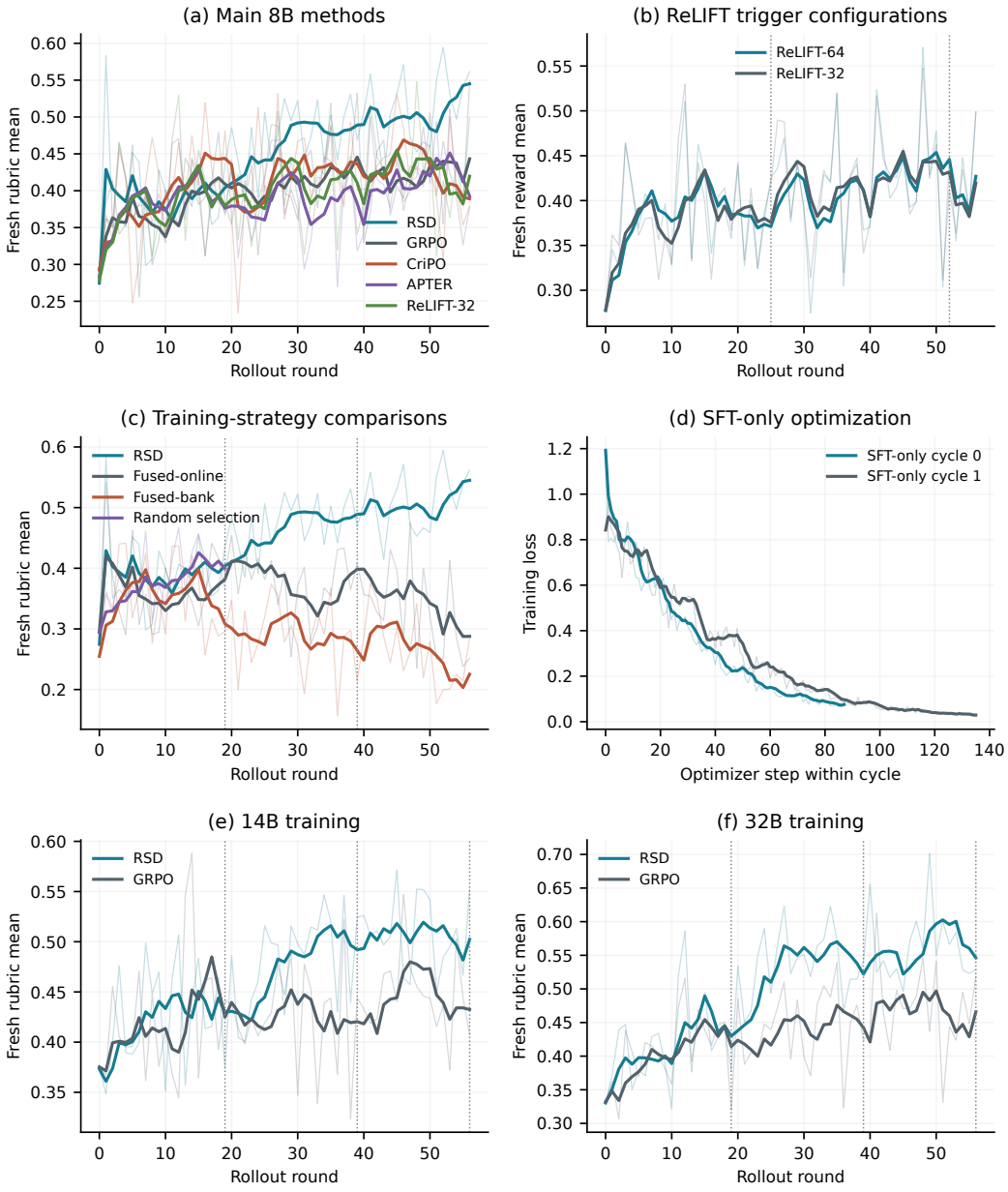

Figure 3: **Main methods, strategy comparisons, and scale.** The metric is stated on each axis. Panel (b) compares ReLIFT-64 with ReLIFT-32; its vertical lines mark the latter's SFT updates at rounds 25 and 52. Panel (a) uses the mean rubric score. Panel (c) includes Fused-online and Fused-bank. Vertical lines in (c), (e), and (f) indicate the RSD SFT boundaries. SFT-only loss is plotted separately by cycle. Each curve corresponds to a training run.

G.1 What the exploratory variants change

These 8B exploratory variants vary how assisted trajectories enter RL, providing context for RSD's dedicated SFT design.

Failure replay. Weak queries enter a replay queue and are retried with rubric-derived hints. The assisted trajectories are used in the RL replay path, alongside fresh unhinted groups, rather than accumulated for RSD's separate SFT phase. "Unhinted" measures the fresh branch; "hinted replay" measures the assisted retries on selected queries.

Mixed-group hints. Unhinted and rubric-hinted trajectories are placed in the same query group for relative-reward learning. This configuration uses eight unhinted and four hinted trajectories per group. The two curves separate these generation conditions even though they share a group.

Separate-group hints. The same query is sampled under two conditions, with and without rubric hints, and the two sets form separate advantage-normalization groups. This configuration uses four trajectories per condition. The curves measure assisted generation and unhinted behavior separately.

Panel suffixes distinguish two failure-replay runs (A/B), two mixed-group runs (A/B), and four separate-group runs (A/B/C/D).

How to read the curves. The horizontal axis is the rollout round and the vertical axis is the mean training reward of the indicated branch. In failure replay A, we have 37 fresh and 51 replay observations through round 56. The replay branch reaches higher scores, while fresh observations remain lower. Each branch follows its sampled queries; gaps mark missing scalar records. Dots retain isolated observations and smoothing never crosses a missing round.

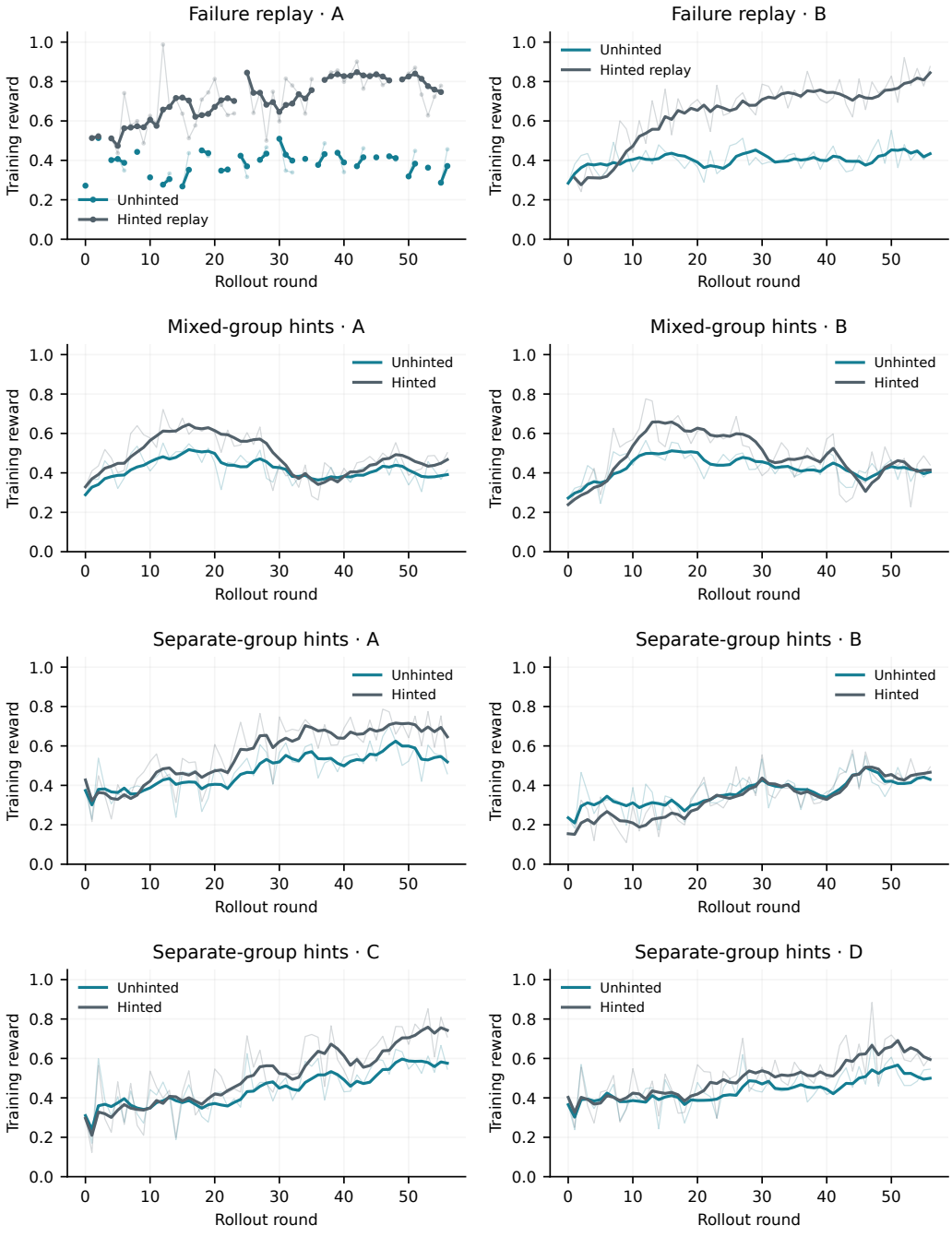

Figure 4: **Exploratory training variants.** The three families are defined in Appendix G.1; letters identify separate runs. Light traces show raw rewards; darker lines average up to five consecutive rounds, restarting after gaps. Dots show sparse observations. Missing records are not interpolated. Assisted and unhinted scores describe their respective query streams.

H A UNIFIED VIEW OF ASSISTED-TO-UNASSISTED TRANSFER

The transfer analysis connects rubric-guided target construction to behavior under the original prompt. It decomposes unhinted improvement into demonstration advantage, imitation error, and query-distribution shift.

Distributions and assumptions. Fix an evaluation environment, a task distribution μ, and a bounded reward $R(q, y) \in [0, 1]$. Write $J_\mu(p) = \mathbb{E}_{q\sim\mu, y\sim p(\cdot|q)} R(q, y)$, where p is the deployed trajectory distribution without hints. Let p_0 be the reference policy before a learning cycle, S the query distribution represented in accepted demonstrations, and $\nu(y \mid q)$ their normalized trajectory distribution. Guidance, resampling, and acceptance are absorbed into ν; after collection, ν is treated as fixed. All three trajectory laws are evaluated with the same reward and environment. Assume finite

$$K_S(\nu\|p) = \mathbb{E}_{q\sim S} \mathrm{KL}(\nu(\cdot \mid q)\|p(\cdot \mid q)) . \tag{8}$$

The mathematical ν denotes a population trajectory law. Query support outside S can be handled by any extension of ν, since the bound below only uses ν on S.

Transfer bound. Define the demonstration advantage $A_S(\nu, p_0) = J_S(\nu) - J_S(p_0)$ and query-shift penalty $V = \mathrm{TV}(\mu, S)$. Then

$$\boxed{J_\mu(p) - J_\mu(p_0) \;\geq\; A_S(\nu, p_0) - \sqrt{\tfrac{1}{2} K_S(\nu\|p)} - 2V.} \tag{9}$$

Proof. On each query, bounded reward and Pinsker's inequality give $|\mathbb{E}_\nu R - \mathbb{E}_p R| \leq \mathrm{TV}(\nu, p) \leq \sqrt{\mathrm{KL}(\nu\|p)/2}$. Averaging over S and applying Jensen's inequality yields $J_S(p) \geq J_S(\nu) - \sqrt{K_S/2}$. The function $f(q) = \mathbb{E}_p R - \mathbb{E}_{p_0} R$ lies in $[-1, 1]$, so $\mathbb{E}_\mu f \geq \mathbb{E}_S f - 2\,\mathrm{TV}(\mu, S)$. Combining the two statements proves Eq. 9.

The decomposition links three parts of the learning mechanism: constructing advantageous targets, imitating them under the original prompt, and covering queries on which the learned behavior will be used.

Connection to the training objective. For exact, unnormalized trajectory cross-entropy, $\mathbb{E}_\nu[-\log p] = H(\nu) + \mathrm{KL}(\nu\|p)$; fitting a fixed ν therefore targets the imitation term. Our supervised objective masks environment tokens and normalizes by the number of supervised tokens. This token-normalized surrogate targets assistant-action prediction. Full trajectory KL decomposes across the complete interaction history:

$$\mathrm{KL}(\nu(y \mid q)\|p(y \mid q)) = \sum_t \mathbb{E}_{\eta_t\sim\nu} \mathrm{KL}\big(\nu(y_t \mid \eta_t, q)\|p(y_t \mid \eta_t, q)\big). \tag{10}$$

Here η_t denotes the interaction history before step t. For variable-length trajectories, terminate and pad with an absorbing state. The decomposition includes assistant actions and tool observations under the accepted trajectory distribution. Removing the hint before SFT gives the student the same initial prompt as at deployment; K_S measures how closely it imitates the target trajectories.

Reward and environment transfer. The training rubric and retrieval environment differ from external evaluation. Let $\widetilde{J}_S$ denote the corresponding training-domain score and assume that $|J_S(v) - \widetilde{J}_S(v)| \leq \epsilon_v$ for $v \in \{\nu, p_0\}$. The corresponding bound incorporates the reward and environment discrepancy:

$$J_\mu(p) - J_\mu(p_0) \geq \widetilde{A}_S(\nu, p_0) - \epsilon_\nu - \epsilon_{p_0} - \sqrt{K_S/2} - 2V. \tag{11}$$

The ϵ terms express the assumed training-to-evaluation discrepancy. The acceptance threshold 0.8 applies to training rubric scores; Eq. 11 specifies the sufficient condition for transfer to the evaluation reward.

Connection to the experiments. RSD accepts trajectories with training reward at least 0.8 and learns them under original prompts. Fused-bank and dedicated consolidation use a matched demonstration-exposure budget: the former interleaves supervised targets with RL in joint updates, while the latter learns the targets in dedicated SFT phases. SFT-only retains demonstration construction and learning; random selection changes the query-selection rule. These design choices connect the strategy comparisons to target construction, imitation, and query coverage in Eq. 9.

Rubric guidance and acceptance shape demonstration advantage; original-prompt learning addresses imitation; and query selection determines the coverage of supervision. The benchmark comparisons evaluate the unhinted behavior produced by these training strategies.

A sufficient condition for positive transfer. For any candidate policy satisfying $A_S \geq a$, $K_S \leq k$, and $V \leq v$, Eq. 9 has a positive lower bound whenever

$$a > \sqrt{k/2} + 2v. \tag{12}$$

Under training/evaluation mismatch, the right side additionally includes the two error bounds from Eq. 11.