CRITSEL-OPSD: CRITERION SELECTION FOR RUBRIC-CONDITIONED ON-POLICY SELF-DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Rubric-conditioned On-Policy Self-Distillation (OPSD) gives the teacher structured criteria as privileged information, but existing formulations typically expose the entire rubric throughout training. We ask whether a rubric, usually treated as an indivisible supervision object, contains redundant teacher-context information. We introduce CritSel-OPSD, a criterion-selection framework that combines leave-one-out teacher-context interventions with Top-K subset construction. Selection uses trajectories from the current student and is refreshed periodically, allowing successive training blocks to use student-state-dependent contexts. Sampled Teacher Influence (STI) provides a trajectory-aligned selector. In our evaluated setting with Qwen3-8B and OpenRubrics, compact teacher contexts preserve the observed benefits of full-rubric conditioning across six diverse benchmarks while using only a subset of rubric criteria. Comparisons with random controls indicate that the central result is the viability of compact teacher-context allocation, not the universal superiority of one ranking signal. These results identify criterion selection as a design dimension distinct from rubric design and reward construction.

1 INTRODUCTION

Rubrics decompose response quality into explicit criteria, providing structured supervision for evaluation and post-training (Christiano et al., 2017; Stiennon et al., 2020; Gunjal et al., 2026; Liu et al., 2026a; Zhang et al., 2026). But a structured rubric need not be an indivisible teacher context: criteria may affect teacher predictions differently on the response trajectories visited by the student, with some supplying redundant information and others inducing larger changes in teacher predictions. This motivates studying criterion-level teacher-context allocation rather than treating rubric availability as an all-or-nothing choice.

OPSD creates a particularly clear allocation problem: a frozen teacher sees information that the rubric-free student does not, yet existing rubric-conditioned formulations typically expose the complete rubric $R = \{c_1, \ldots, c_M\}$ at every training step (Xu et al., 2026a; Rezaei et al., 2026; Gu et al., 2026). We refer to this baseline as **full-rubric OPSD (Full-OPSD)**. Full-rubric exposure leaves the necessity of individual criteria untested; we ask whether the teacher needs the entire rubric.

We study whether a compact subset $R^{\text{sel}} \subseteq R$ can preserve the benefit of full-rubric conditioning. We introduce **CritSel-OPSD**, which constructs $\pi_T(\cdot \mid x, R^{\text{sel}})$ through criterion-level interventions on trajectories generated by the current student. The rubric, teacher, rollout protocol, and distillation objective remain fixed; only the teacher-visible subset changes. Attribution scoring and refresh frequency are design choices within this allocation framework.

Figure 1 highlights the distinction between possessing a complete rubric and choosing which criteria to expose. Adaptive-rubric methods modify rubric content or criterion weights as policies evolve (Tyagi et al., 2026; Rezaei et al., 2025; Ding et al., 2026), whereas CritSel-OPSD keeps the rubric and distillation objective fixed. It periodically scores criterion-level interventions on student trajectories, retains Top-K criteria, and applies standard OPSD with the selected teacher context.

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.

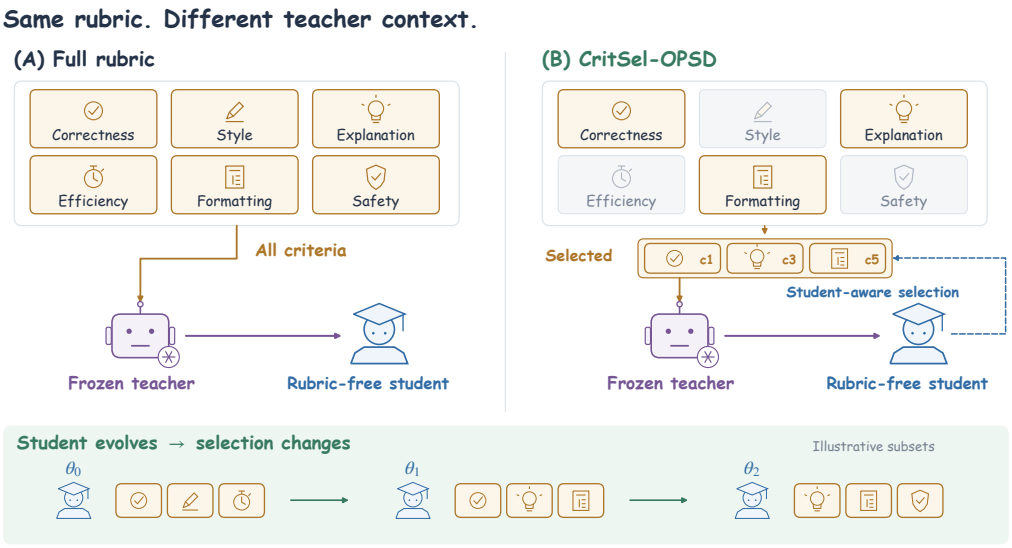

Figure 1: The teacher-context allocation question: must all rubric criteria be exposed throughout training? Left: Full-OPSD exposes the complete rubric to the frozen teacher. Right: CritSel-OPSD retains a compact, student-dependent subset; only the selected criteria enter the teacher context. The bottom strip illustrates that different student states may induce different useful criterion subsets within a fixed rubric.

Experiments with Qwen3-8B on 32,000 processed OpenRubrics examples show that compact criterion selection preserves the observed full-rubric gains. With a four-criterion selected context, STI achieves an Avg-6 score of 60.05 compared with 59.44 for Full-OPSD and 58.10 for Base. Random and original-order subsets are also competitive, indicating that the primary finding is the viability of compact teacher-context allocation rather than the dominance of a single ranking signal. Further analysis shows that STI closely tracks full-distribution teacher sensitivity, whereas teacher-side and student-relative attribution induce substantially different rankings, and the selected subsets exhibit structured student-state dependence rather than simple prefix truncation.

Our contributions are threefold:

- We formulate **criterion selection** as a distinct design problem for OPSD with rubric-conditioned teacher contexts: choosing which criteria from a fixed rubric should construct the privileged teacher context at the current student state.
- We introduce **CritSel-OPSD**, a framework that constructs periodically refreshed, student-state-dependent teacher contexts through leave-one-out teacher-context attribution, while keeping the OPSD objective unchanged.
- We empirically characterize criterion selection: compact contexts can preserve full-rubric performance, selected subsets vary with student state, and different attribution objectives capture different notions of criterion influence.

2 Related Work

2.1 On-Policy and Rubric-Conditioned Distillation

On-policy and privileged-context distillation query a teacher on student-generated trajectories or enriched teacher contexts, rather than only on expert prefixes (Lopez-Paz et al., 2015; Agarwal et al., 2024; Zhao et al., 2026; Ye et al., 2026). Rubric-conditioned variants provide the closest setting: a frozen teacher receives criterion-level rubrics and supplies token-level targets to a rubric-free student (Rezaei et al., 2026; Gu et al., 2026). Related approaches instead use rubric scores for black-box distillation or alter the weighting of supervision across positions (Fang et al., 2026; Liu et al., 2026b).

The closest rubric-conditioned formulations typically study full-rubric teacher contexts. We study the less-explored design choice of which criteria from a fixed rubric should construct that privileged context, while keeping the rollout protocol and distillation objective unchanged.

2.2 Rubric-Based and Adaptive Supervision

Rubric-based post-training has used criteria as rewards, scalable reward-modeling inputs, token- or step-level feedback, rollout scaffolds, and policy-visible reasoning context (Gunjal et al., 2026; Liu et al., 2026a; Xu et al., 2026a;b; Xie et al., 2026; Zhou et al., 2025; Yu et al., 2026). Adaptive methods further elicit or evolve rubric criteria, co-evolve rubrics with the policy, or adjust criterion-level reward weights using current-policy rollouts (Rezaei et al., 2025; Shao et al., 2025; Li et al., 2026b; Ding et al., 2026; Tyagi et al., 2026). These methods therefore adapt rubric content, reward construction, or policy-visible supervision. CritSel-OPSD operates on a different object: the rubric text, teacher parameters, and distillation objective remain fixed, while only the subset of criteria exposed to the frozen teacher changes. The student never receives the rubric, and selected criteria are not rewards, loss weights, or credit-assignment targets.

2.3 Context Attribution and Selection

Context attribution uses perturbation or leave-one-out interventions to measure how context components affect a model's response likelihood or output distribution (Cohen-Wang et al., 2024; Liu et al., 2025; Li et al., 2026a); related work also uses ablation signals for learning and compresses or selects inference context (Chuang et al., 2025; Li et al., 2023; Xu et al., 2023). We use the same technical antecedent but change the intervention object and downstream role: leave-one-out rubric criteria are evaluated through teacher likelihoods on a student-generated trajectory, and the resulting ranking constructs the privileged teacher context for subsequent on-policy distillation. Unlike input attribution or inference-context reduction, our goal is student-state-dependent selection of fixed rubric information for teacher supervision; attribution scores are not optimization rewards.

3 Method

3.1 CritSel-OPSD Overview

We consider OPSD with rubric-conditioned teacher contexts, using a trainable student π_θ, a frozen teacher π_T, and a frozen rubric-free reference π_{ref}. For input x with rubric $R = \{c_1, \ldots, c_M\}$, the student observes only x, while the teacher receives a selected subset of R. The teacher and reference are copies of the initial student, $\pi_T = \pi_{\text{ref}} = \pi_{\theta_0}$, with different conditioning contexts. Full-OPSD uses all of R. The student and full-rubric teacher distributions on a student-generated prefix are

$$p_t(v) = \pi_\theta(v \mid x, y_{<t}), \qquad q_t^R(v) = \pi_T(v \mid x, R, y_{<t}). \tag{1}$$

CritSel-OPSD changes only the privileged teacher context. The student rollout protocol, frozen model parameters, and distillation objective are unchanged; attribution scores determine subset membership but are not rewards or optimization weights. As summarized in Figure 2, each refresh consists of

$$\text{calibration} \rightarrow \text{intervention} \rightarrow \text{score} \rightarrow \text{Top-}K \rightarrow \text{OPSD}.$$

3.2 Criterion Selection Framework

Criterion selection separates three components: teacher-context intervention, attribution scoring, and subset construction. It is performed at refresh boundaries $h \in \{0, H, 2H, \ldots\}$, where θ_h denotes the student after h successful optimizer updates. For each eligible example x_i, the current student generates one rubric-free greedy calibration trajectory

$$y_i^{(h)} = \text{Generate}\left(\pi_{\theta_h}(\cdot \mid x_i); T = 0\right). \tag{2}$$

This trajectory is fixed for all teacher-context interventions and is used only for attribution. When the example is later trained, the student generates an independent fresh stochastic trajectory.

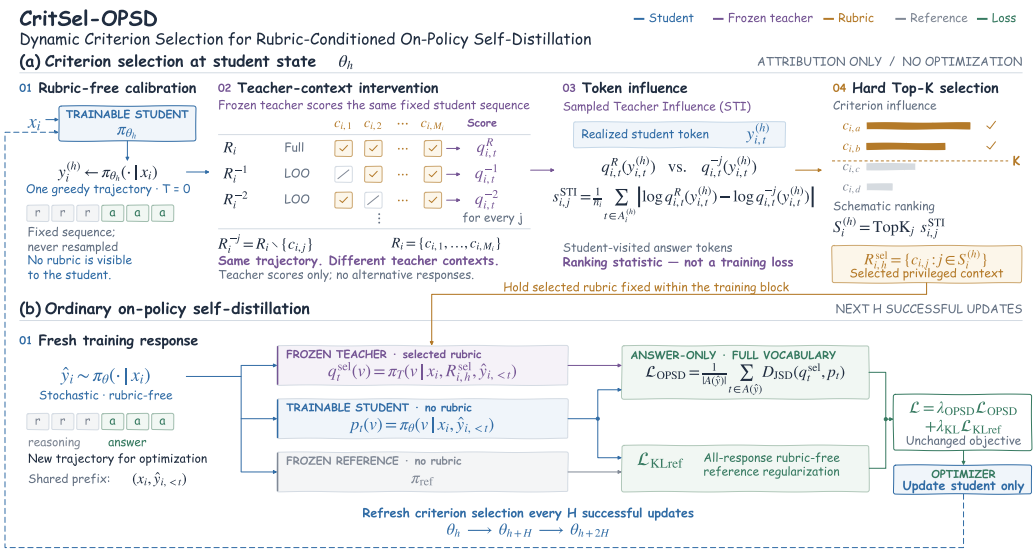

Figure 2: CritSel-OPSD alternates between (A) greedy student calibration followed by full-rubric and leave-one-out (LOO) teacher scoring (Full+LOO) and Top-K criterion selection, and (B) ordinary on-policy self-distillation with the selected rubric. The calibration trajectory is used only for selection; fresh stochastic trajectories are used for training.

For $R_i = \{c_{i,1}, \ldots, c_{i,M_i}\}$, the full-rubric teacher distribution at answer position t is

$$q^R_{i,t}(v) = \pi_T\left(v \mid x_i, R_i, y^{(h)}_{i,<t}\right). \tag{3}$$

For each criterion $c_{i,j}$, we form the leave-one-out (LOO) rubric

$$R_i^{-j} = R_i \setminus \{c_{i,j}\}, \tag{4}$$

and score the same fixed sequence, preserving the remaining criteria's relative order, with

$$q^{-j}_{i,t}(v) = \pi_T\left(v \mid x_i, R_i^{-j}, y^{(h)}_{i,<t}\right). \tag{5}$$

Thus Full+LOO branches differ only in the teacher context; neither generates an alternative response. An attribution objective maps these paired teacher evaluations on the calibration trajectory to a score $s_{i,j}(\theta_h)$ for each criterion.

For a chosen attribution signal $s_{i,j}(\theta_h)$, the selected subset is built by hard Top-K construction:

$$S_i^{(h)} = \mathrm{TopK}_j\ s_{i,j}(\theta_h), \qquad R^{\mathrm{sel}}_{i,h} = \{c_{i,j} : j \in S_i^{(h)}\}. \tag{6}$$

The selected criteria are restored to their original rubric order before teacher prompting. If $M_i \le K$, we retain the complete rubric. The framework is agnostic to the particular attribution signal. We use STI as the default instantiation because it is trajectory-aligned, and evaluate two Jensen–Shannon divergence (JSD) variants—**Teacher Influence JSD** (TI-JSD) and **Student-Relative Influence JSD** (SI-JSD)—as diagnostic alternatives; the empirical relation between STI and distribution-level sensitivity is examined later. For brevity, we use TI and SI below to refer to TI-JSD and SI-JSD, respectively.

Default instantiation: STI. Let $A_i^{(h)}$ be the valid answer-token positions of the calibration trajectory. The default Sampled Teacher Influence (STI) score is

$$\boxed{s^{\mathrm{STI}}_{i,j}(\theta_h) = \frac{1}{|A_i^{(h)}|} \sum_{t \in A_i^{(h)}} \left| \log q^R_{i,t}\left(y^{(h)}_{i,t}\right) - \log q^{-j}_{i,t}\left(y^{(h)}_{i,t}\right) \right|} \tag{7}$$

STI measures the absolute change in the frozen teacher's log-probability of realized student tokens when one criterion is removed. Dependence on the student enters through the calibration trajectory, while the compared distributions remain frozen-teacher distributions. This is local LOO sensitivity, not an additive causal contribution or Shapley value.

3.3 SELECTED-CONTEXT OPSD

For an example assigned to the block prepared at boundary h, write $R^{\mathrm{sel}} = R^{\mathrm{sel}}_{i,h}$. The student then generates a fresh rubric-free stochastic response

$$\hat{y} \sim \pi_\theta(\cdot \mid x). \tag{8}$$

On the same realized prefixes, the selected-rubric teacher and student produce

$$q_t^{\mathrm{sel}}(v) = \pi_T\left(v \mid x, R^{\mathrm{sel}}, \hat{y}_{<t}\right), \qquad p_t(v) = \pi_\theta\left(v \mid x, \hat{y}_{<t}\right). \tag{9}$$

The rubric is supplied only to the frozen teacher; the student and rubric-free reference do not receive it.

We retain the inherited answer-only, full-vocabulary JSD distillation objective (Zhao et al., 2026; Rezaei et al., 2026). If $A(\hat{y})$ denotes answer-token positions,

$$\mathcal{L}_{\mathrm{OPSD}} = \frac{1}{|A(\hat{y})|} \sum_{t \in A(\hat{y})} D_{\mathrm{JSD}}\left(q_t^{\mathrm{sel}}, p_t\right), \tag{10}$$

where D_{JSD} is the standard Jensen–Shannon divergence over the complete vocabulary. A frozen rubric-free reference provides the full-response Kullback–Leibler (KL) regularizer $\mathcal{L}_{\mathrm{KLref}}$ (Appendix E), and the student objective is

$$\mathcal{L}(\theta) = \lambda_{\mathrm{OPSD}}\mathcal{L}_{\mathrm{OPSD}} + \lambda_{\mathrm{KL}}\mathcal{L}_{\mathrm{KLref}}. \tag{11}$$

Only θ is updated.

At each refresh boundary, the current student prepares selected contexts for the next H updates. Each example is attributed once and retains its assigned subset; dynamic selection uses successive student snapshots for successive blocks, rather than repeatedly rescoring the same training example. Block preparation and update accounting are specified in Appendices A.2 and A.

4 EXPERIMENTS

Our experiments address three questions:

(1) Can criterion selection preserve the benefit of full-rubric teacher conditioning?

(2) What criteria does selection expose, and is it equivalent to fixed prefix truncation?

(3) How does selection vary with student state, and how do different attribution objectives characterize criterion influence?

Additional refresh-frequency and budget-sensitivity results are described as ablations. Rubrics are visible only to the teacher; student training inputs and downstream inference remain rubric-free. Unless otherwise stated, we evaluate CritSel-OPSD with STI as the default attribution signal.

4.1 EXPERIMENTAL SETUP

Models and training. We initialize the student, frozen teacher, and frozen reference from Qwen3-8B. Training uses 32,000 processed examples from OpenRubrics (Liu et al., 2026a). Examples originating from the Instruction Following Evaluation (IFEval) dataset are excluded, invalid rubrics are removed, and normalized prompts are deduplicated. Unless otherwise stated, runs use 500 optimizer updates, batch size 64, selection budget $K = 4$, and refresh interval $H = 100$. In Eq. 11, we set $\lambda_{\mathrm{OPSD}} = 0.1$ and $\lambda_{\mathrm{KL}} = 0.00025$, with JSD mixing coefficient $\alpha = 0.5$. Distillation covers answer tokens; reference regularization covers the full response.

Attribution and optimization use separate student trajectories; selected rubrics remain fixed within each block. Full execution semantics are given in Appendix A.2.

Baseline taxonomy. Baselines include the rubric-free initialization (**Base**), **Full-OPSD**, CritSel-OPSD with random and original-order selection controls, static CritSel-OPSD (STI), dynamic CritSel-OPSD with random selection, and alternative attribution objectives (**TI-JSD** and **SI-JSD**). All subset selectors expose at most K criteria; detailed definitions are provided in Appendix B.

Benchmark coverage and evaluation. We evaluate six benchmarks covering reasoning, coding, knowledge, and instruction following: Grade School Math 8K (GSM8K), BIG-Bench Hard (BBH), Instruction Following Evaluation (IFEval), Graduate-Level Google-Proof Q&A (GPQA), Mostly Basic Python Problems Plus (MBPP+), and PopQA (Cobbe et al., 2021; Suzgun et al., 2023; Zhou et al., 2023; Rein et al., 2023; Liu et al., 2023; Mallen et al., 2023). We report their unweighted six-benchmark average (Avg-6); dataset splits, decoding settings, and metric definitions are provided in Appendix D.3.

For each reported in-distribution checkpoint, we average benchmark scores over three answer-generation seeds (42, 1348, and 4527), which vary the stochastic decoding while keeping the checkpoint fixed.

4.2 Does Criterion Selection Preserve the Benefits of Rubric Conditioning?

Table 1: Rubric-free downstream performance after 500 updates. All subset selectors use $K = 4$; dynamic selectors use $H = 100$, while static selectors prepare contexts at initialization. Avg-6 is the unweighted six-benchmark average. Δ reports the percentage-point change in Avg-6 relative to Base. The best and second-best results are shown in **bold** and underlined, respectively.

Method	GSM8K	BBH	IFEval	GPQA	MBPP+	PopQA	Avg-6 ↑	Δ vs. Base
Initialization								
Base	95.63	64.72	34.07	54.02	69.16	31.02	58.10	–
Full-OPSD								
Full-OPSD	95.53	70.06	37.03	53.94	68.93	31.13	59.44	+1.34
Static compact subset controls								
CritSel-OPSD (Random-4)	95.58	70.09	37.95	53.20	**69.40**	31.21	59.57	+1.47
CritSel-OPSD (Original-Order Top-4)	95.68	70.61	36.85	54.54	69.37	31.13	59.70	+1.60
Static CritSel-OPSD (STI)	95.60	70.02	37.52	53.79	69.06	31.16	59.53	+1.43
Dynamic compact selection								
Dynamic CritSel-OPSD (Random-4)	95.43	70.49	**38.14**	54.39	69.01	31.15	59.77	+1.67
CritSel-OPSD (TI-JSD)	95.55	70.18	36.85	54.39	69.16	31.13	59.54	+1.44
CritSel-OPSD (SI-JSD)	95.48	71.00	37.15	54.46	68.92	31.15	59.69	+1.59
CritSel-OPSD (STI)	**95.78**	**71.67**	37.46	**54.91**	69.16	**31.29**	**60.05**	**+1.95**

Table 1 first establishes the reference effect of rubric conditioning: Full-OPSD reaches an Avg-6 of 59.44, compared with 58.10 for Base. Compact contexts preserve this observed benefit without exposing the entire rubric: the comparison with random and original-order subsets tests whether this effect comes from merely reducing context size or from criterion-aware selection. Static CritSel-OPSD (STI), CritSel-OPSD (Random-4), CritSel-OPSD (Original-Order Top-4), and Dynamic CritSel-OPSD (Random-4) reach 59.53, 59.57, 59.70, and 59.77, respectively.

CritSel-OPSD (STI) obtains the highest point estimate at 60.05, but its 0.28-point margin over Dynamic CritSel-OPSD (Random-4) is modest. We therefore interpret these results primarily as evidence that compact criterion selection is feasible, rather than as evidence that one attribution signal universally dominates others. Task-level results are heterogeneous: different controls lead on different benchmarks.

Criterion exposure and performance. Table 2 examines whether exposing more teacher-visible criteria consistently improves downstream performance. The results show that a small selected context can retain the observed full-rubric performance, suggesting that complete rubric exposure is not always necessary. Criterion counts are reported for context characterization only; they do not imply token or runtime savings, and the available budgets do not establish that shorter contexts are always better.

4.3 What Does Criterion Selection Expose?

We examine the structure of selected teacher contexts: how much of the rubric is retained, whether selection is equivalent to original-order truncation, and whether selected subsets vary with student state.

Training artifacts show that Top-4 selection retains only a subset of the teacher context (Figure 3a). Although early criteria are selected more frequently, later criteria remain selected with non-zero probability. Selection probabilities deviate from the random-K expectation (Figure 3b), while the

Table 2: Criterion exposure versus downstream performance, using existing runs. Bounds for $K = 2, 6$ follow from $\min(K, M_i)$; measured mean exposure is reported only for the full and default contexts.

Teacher context	Criteria per example	Avg-6
Full-OPSD	8.53 (mean)	59.44
CritSel-OPSD (STI, $K = 2$)	≤ 2	59.71
CritSel-OPSD (STI, $K = 4$)	4.00 (mean, rounded)	60.05
CritSel-OPSD (STI, $K = 6$)	≤ 6	59.59

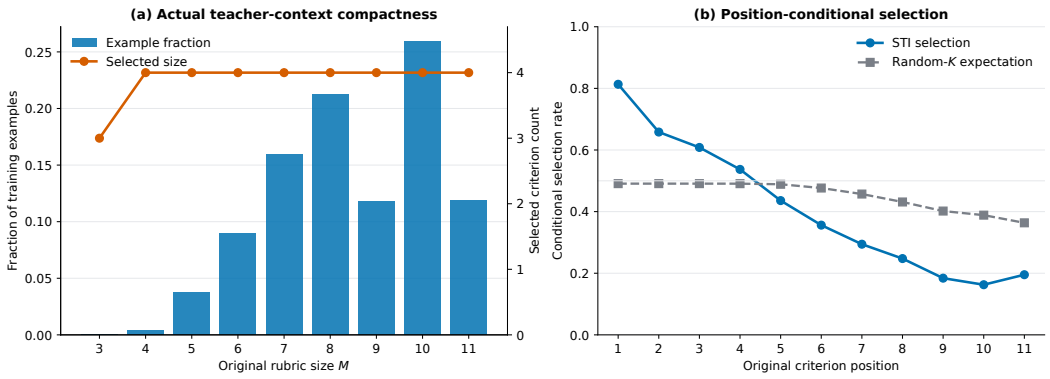

Figure 3: Training-context reduction and position bias. (a) Distribution of original rubric size M with the selected size $\min(4, M)$ overlaid. (b) Conditional selection probability by original criterion position, together with the analytical random-K expectation computed under the observed rubric-size distribution.

selection of later criteria shows that STI is not equivalent to deterministic prefix truncation. This does not rule out position effects.

Position-wise selection rates, prefix overlap, temporal consistency, selection frequencies, and cutoff-margin statistics are reported in Appendix C. The refresh-frequency comparison remains an ablation because it changes the training protocol in addition to the selected context.

4.4 How Does Criterion Selection Vary with Student State?

We first examine whether selected teacher contexts vary with the student state on a fixed probe panel. We then analyze refresh frequency as a secondary operating choice. Static CritSel-OPSD (STI) selects at initialization, whereas dynamic CritSel-OPSD (STI) uses the student snapshot at each refresh boundary.

On a fixed probe panel, selected subsets remain similar across nearby checkpoints but drift over longer intervals, while remaining more structured than random selection (Figure 4). The frequency profile suggests a persistent core together with a state-dependent boundary.

Table 3: CritSel-OPSD (STI) refresh controls at $K = 4$ and 500 updates. Static CritSel-OPSD (STI) selects once at initialization; it is a selection-mode comparator, not an H-frequency condition.

Method	H	Updates	GSM8K	BBH	IFEval	GPQA	MBPP+	PopQA	Avg-6
Matched refresh controls									
CritSel-OPSD (STI)	1	500	95.53	65.19	37.46	53.79	68.92	31.15	58.67
CritSel-OPSD (STI)	25	500	95.53	70.82	37.40	54.09	69.12	31.23	59.70
CritSel-OPSD (STI)	50	500	95.38	70.51	37.15	53.27	69.19	31.23	59.46
CritSel-OPSD (STI)	100	500	95.78	71.67	37.46	54.91	69.16	31.29	60.05
CritSel-OPSD (STI)	250	500	95.70	71.17	37.22	54.91	69.22	31.26	59.91
Static comparator									
Static CritSel-OPSD (STI)	–	500	95.60	70.02	37.52	53.79	69.06	31.16	59.53

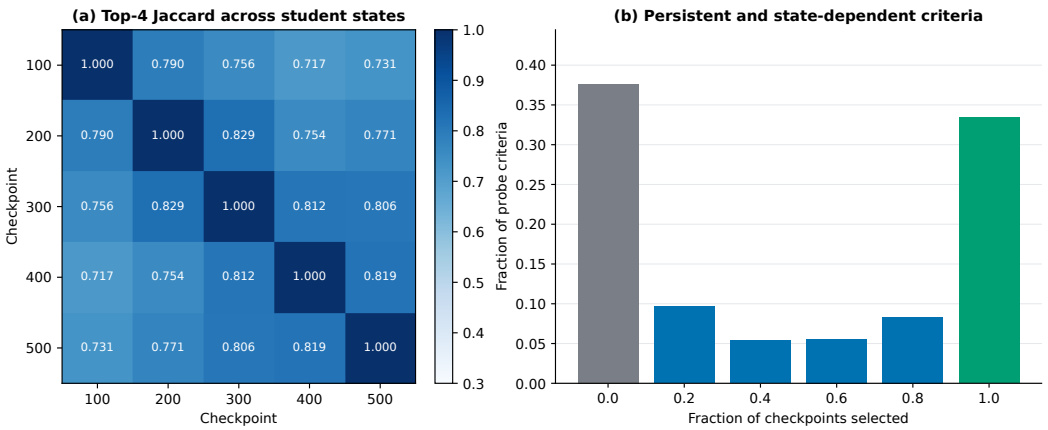

Figure 4: Student-state-dependent criterion dynamics on a fixed probe panel. (a) Pairwise Top-4 Jaccard similarity across checkpoints, compared with a random-set reference. (b) Frequency with which each criterion position is selected across checkpoints.

Table 3 gives two observations. First, matched dynamic CritSel-OPSD (STI) runs span 58.67–60.05 for $H = 1$–250; $H = 100$ achieves the best observed point estimate, whereas the most frequent $H = 1$ condition gives 58.67. Second, Static CritSel-OPSD (STI) reaches 59.53, which is 0.52 points below the default dynamic run but above the $H = 1$ row. The non-monotonic sweep describes the effect of refresh frequency in this implementation; it does not establish that dynamic refreshing or any particular interval universally dominates. These point estimates leave the preferred schedule and variability across independent training runs unresolved.

4.5 Comparing Criterion Attribution Objectives

We now compare how different attribution objectives rank criteria within the same intervention framework. We therefore compare criterion rankings induced by STI with two alternatives that summarize different sensitivity notions. **TI-JSD** asks which criteria change the teacher's full next-token distribution. **SI-JSD** asks which criteria change the teacher–student discrepancy in the sampled direction. These are different attribution objectives: high TI identifies large teacher shifts, whereas high positive SI identifies criteria whose presence increases the teacher–student discrepancy, equivalently indicating that removing the criterion reduces the discrepancy.

Probe protocol. On a fixed probe panel, we evaluate the student at multiple training checkpoints. At each checkpoint, the student generates a fresh greedy, rubric-free response, and full-rubric and leave-one-out teacher contexts are scored on the same answer tokens. This offline diagnostic is separate from one-pass training attribution; protocol details are in Appendix D.4.

STI tracks full-distribution sensitivity rankings. We compute Spearman correlation across criteria within each prompt and then average across prompts. STI–TI agreement remains high, ranging from 0.805 to 0.836 across checkpoints (Figure 5a). The same conclusion appears at the level of selected teacher contexts: Top-4 overlap averages 0.838 and Jaccard similarity averages 0.749, both above random-set references (Figure 5b). Thus, a realized-token score can recover most of the ranking structure found by full-vocabulary teacher sensitivity in this probe, while remaining a distinct statistic rather than a numerical equivalent of TI-JSD. Appendix Tables 7 and 8 give the exact checkpoint statistics; Appendix D.4 derives the chance baselines.

Student-relative attribution is a different objective. These teacher-sensitivity and student-relative attribution objectives also behave differently. TI–SI rank correlations are only 0.125–0.187, with mean Top-4 overlap of 0.575 and mean Jaccard similarity of 0.433 (Figure 5c). This low agreement does not show that either signal is intrinsically better; it shows that "changing the teacher" and "moving the teacher toward the current student" are not the same selection goal. Accordingly, the

downstream performance of TI- and SI-based selectors should be interpreted as an empirical comparison of two attribution objectives rather than as a validation of one underlying notion of criterion importance. Within CritSel-OPSD, STI is attractive because it is both trajectory-aligned and highly consistent with broader teacher-side criterion sensitivity.

4.6 Ablation and Sensitivity Analysis

Attribution objective. CritSel-OPSD (STI) obtains the highest Avg-6 point estimate, although all attribution variants and controls remain competitive (Table 4). Attribution choice affects selected contexts, but is not the sole determinant of downstream performance.

Table 4: Attribution-objective controls at $K = 4, H = 100$ and 500 updates. CritSel-OPSD (Bottom-STI) retains the lowest-STI criteria. CritSel-OPSD (STI) denotes Top-K STI selection and is repeated as the reference row. CritSel-OPSD (Bottom-SI-JSD) selects the lowest signed SI-JSD scores, not the lowest absolute magnitudes.

Method	GSM8K	BBH	IFEval	GPQA	MBPP+	PopQA	Avg-6
CritSel-OPSD (TI-JSD)	95.55	70.18	36.85	54.39	69.16	31.13	59.54
CritSel-OPSD (SI-JSD)	95.48	71.00	37.15	54.46	68.92	31.15	59.69
Dynamic CritSel-OPSD (Random-4)	95.43	70.49	38.14	54.39	69.01	31.15	59.77
CritSel-OPSD (Bottom-STI)	95.32	70.59	37.34	54.99	69.01	31.31	59.76
CritSel-OPSD (Bottom-SI-JSD)	95.43	71.41	37.34	54.24	69.51	31.18	59.85
Reference							
CritSel-OPSD (STI)	95.78	71.67	37.46	54.91	69.16	31.29	60.05

Selection budget. Table 5 varies K while keeping the default refresh schedule. CritSel-OPSD (STI) reaches 59.71, 60.05, and 59.59 for $K = 2$, 4, and 6, respectively. These three sparse points show that a small selected context can support strong performance, but they do not identify a preferred budget or support a monotonic capacity trend.

Table 5: CritSel-OPSD (STI) selection-budget controls at $H = 100$ and 500 updates. The sparse $K = 2, 4, 6$ sweep is descriptive and does not identify a preferred budget.

Method	K	GSM8K	BBH	IFEval	GPQA	MBPP+	PopQA	Avg-6
CritSel-OPSD (STI)	2	95.40	71.76	37.09	53.87	69.02	31.14	59.71
CritSel-OPSD (STI)	4	95.78	71.67	37.46	54.91	69.16	31.29	60.05
CritSel-OPSD (STI)	6	95.68	69.76	37.34	54.54	68.96	31.28	59.59

5 Conclusion

We study criterion allocation as a distinct design dimension of rubric-conditioned on-policy self-distillation. CritSel-OPSD constructs selected teacher contexts from a fixed rubric through leave-one-out teacher-context attribution on student-generated trajectories, while leaving the underlying distillation objective unchanged. In the evaluated setting, compact teacher contexts achieve comparable downstream performance to full-rubric conditioning, showing that rubrics need not be treated as indivisible teacher contexts. Selected subsets exhibit structured student-state-dependent variation, while competitive random and original-order controls indicate that the evidence supports compact teacher-context feasibility rather than the universal superiority of STI. Criterion sensitivity should therefore not be interpreted as a direct measure of downstream learning utility.

Scope and limitations. Our evaluation uses one model family and one rubric source. The results establish feasibility in this setting, while generalization to other models and rubric distributions remains open. Decoding-seed averages do not quantify training-seed variation, and sensitivity rankings do not directly measure transferable criterion utility. Compact teacher contexts also require additional attribution computation (Appendix A.2); we make no claim of end-to-end efficiency gains.

AI USE STATEMENT

We used Codex with GPT-5.6-sol for AI assistance during the preparation of this paper for four purposes: writing assistance, including language polishing; retrieval and discovery of related work; research ideation and execution, including analysis planning and code-level support; and drafting portions of the manuscript. The authors verified the relevant sources, checked the analyses and experimental results, and take full responsibility for the final text, claims, and conclusions.

REPRODUCIBILITY STATEMENT

The method and experimental setting are specified in Sections 3 and 4. Detailed implementation, execution, evaluation, and diagnostic protocols are provided in the appendix, including the algorithm summary (Appendix A), cost details (Appendix A.2), and evaluation protocols (Appendix D). The default setting uses Qwen3-8B, 32,000 processed OpenRubrics examples, $K = 4$, $H = 100$, and 500 successful optimizer updates.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Yung-Sung Chuang, Benjamin Cohen-Wang, Shannon Zejiang Shen, Zhaofeng Wu, Hu Xu, Xi Victoria Lin, James Glass, Shang-Wen Li, and Wen-tau Yih. Selfcite: self-supervised alignment for context attribution in large language models. *arXiv preprint arXiv:2502.09604*, 2025.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Mądry. Contextcite: Attributing model generation to context. *Advances in Neural Information Processing Systems*, 37:95764–95807, 2024.

Hongxin Ding, Baixiang Huang, Yue Fang, Weibin Liao, Zheng Li, Jinyang Zhang, Zhijing Wu, Junfeng Zhao, and Yasha Wang. Evorubrics: Dynamic rubrics as rewards via adversarial co-evolution for llm reinforcement learning. *arXiv preprint arXiv:2606.23038*, 2026.

Junfeng Fang, Zhepei Hong, Mao Zheng, Mingyang Song, Gengsheng Li, Houcheng Jiang, Dan Zhang, Haiyun Guo, Xiang Wang, and Tat-Seng Chua. Rubric-based on-policy distillation. *arXiv preprint arXiv:2605.07396*, 2026.

Siyi Gu, Jialin Chen, Sophia Zhou, Arman Cohan, and Rex Ying. Rethinking reward supervision: Rubric-conditioned self-distillation. *arXiv preprint arXiv:2606.19327*, 2026.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. In *International Conference on Learning Representations*, volume 2026, pp. 127924–127945, 2026.

Ruizhe Li, Chen Chen, Yuchen Hu, Yanjun Gao, Xi Wang, and Emine Yilmaz. Attributing response to context: A jensen–shannon divergence driven mechanistic study of context attribution in retrieval-augmented generation. In *International Conference on Learning Representations*, volume 2026, pp. 146805–146837, 2026a.

Shuyue Stella Li, Rui Xin, Teng Xiao, Yike Wang, Rulin Shao, Zoey Hao, Melanie Sclar, Sewoong Oh, Faeze Brahman, Pang Wei Koh, et al. Evolm: Self-evolving language models through co-evolved discriminative rubrics. *arXiv preprint arXiv:2605.03871*, 2026b.

Yucheng Li, Bo Dong, Frank Guerin, and Chenghua Lin. Compressing context to enhance inference efficiency of large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 6342–6353, 2023.

Fengyuan Liu, Nikhil Kandpal, and Colin Raffel. Attribot: A bag of tricks for efficiently approximating leave-one-out context attribution. In *International Conference on Learning Representations*, volume 2025, pp. 16687–16710, 2025.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. *Advances in neural information processing systems*, 36:21558–21572, 2023.

Tianci Liu, Ran Xu, Tony Yu, Ilgee Hong, Carl Yang, Tuo Zhao, and Haoyu Wang. Openrubrics: Towards scalable synthetic rubric generation for reward modeling and llm alignment. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17417–17437, 2026a.

Xiaogeng Liu, Xinyan Wang, Yingzi Ma, Yechao Zhang, and Chaowei Xiao. When are teacher tokens reliable? position-weighted on-policy self-distillation for reasoning. *arXiv preprint arXiv:2605.21606*, 2026b.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. *arXiv preprint arXiv:1511.03643*, 2015.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 9802–9822, 2023.

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.

MohammadHossein Rezaei, Robert Vacareanu, Zihao Wang, Clinton Wang, Bing Liu, Yunzhong He, and Afra Feyza Akyürek. Online rubrics elicitation from pairwise comparisons. *arXiv preprint arXiv:2510.07284*, 2025.

MohammadHossein Rezaei, Anas Mahmoud, Zihao Wang, Utkarsh Tyagi, Advait Gosai, Razvan-Gabriel Dumitru, Aakash Sabharwal, Bing Liu, and Yunzhong He. Rubric-guided self-distillation: Post-training without rubric verifiers. *arXiv preprint arXiv:2606.12507*, 2026.

Rulin Shao, Akari Asai, Shannon Zejiang Shen, Hamish Ivison, Varsha Kishore, Jingming Zhuo, Xinran Zhao, Molly Park, Samuel G Finlayson, David Sontag, et al. Dr tulu: Reinforcement learning with evolving rubrics for deep research. *arXiv preprint arXiv:2511.19399*, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021, 2020.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed H Chi, Denny Zhou, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13003–13051, 2023.

Utkarsh Tyagi, Xingang Guo, MohammadHossein Rezaei, Daniel George, Anas Mahmoud, Jackson Lee, Bing Liu, and Yunzhong He. Not every rubric teaches equally: Policy-aware rubric rewards for rlvr. *arXiv preprint arXiv:2605.20164*, 2026.

Weichu Xie, Haozhe Zhao, Wenpu Liu, Yongfu Zhu, Liang Chen, Minghao Ye, Zirong Chen, Yuqi Xu, Shuai Dong, Ziyue Wang, et al. Step-wise rubric rewards for llm reasoning. *arXiv preprint arXiv:2605.17291*, 2026.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. Recomp: Improving retrieval-augmented lms with compression and selective augmentation. *arXiv preprint arXiv:2310.04408*, 2023.

Ran Xu, Tianci Liu, Zihan Dong, Tony Yu, Ilgee Hong, Carl Yang, Linjun Zhang, Tao Zhao, and Haoyu Wang. Alternating reinforcement learning for rubric-based reward modeling in non-verifiable llm post-training. *arXiv preprint arXiv:2602.01511*, 2026a.

Tianze Xu, Yanzhao Zheng, Pengrui Lu, Lyumanshan Ye, Yong Wu, Zhentao Zhang, Yuanqiang Yu, Chao Ma, Jihuai Zhu, Pengfei Liu, et al. Rubrics to tokens: Bridging response-level rubrics and token-level rewards in instruction following tasks. *arXiv preprint arXiv:2604.02795*, 2026b.

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-policy context distillation for language models. *arXiv preprint arXiv:2602.12275*, 2026.

Jiachen Yu, Zhihao Xu, Junjie Wang, and Yujiu Yang. Think-with-rubrics: From external evaluator to internal reasoning guidance. *arXiv preprint arXiv:2605.07461*, 2026.

Junkai Zhang, Zihao Wang, Lin Gui, Swarnashree Mysore Sathyendra, Jaehwan Jeong, Victor Veitch, Wei Wang, Yunzhong He, Bing Liu, and Lifeng Jin. Chasing the tail: Effective rubric-based reward modeling for large language model post-training. In *International Conference on Learning Representations*, volume 2026, pp. 133430–133457, 2026.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

Yang Zhou, Sunzhu Li, Shunyu Liu, Wenkai Fang, Kongcheng Zhang, Jiale Zhao, Jingwen Yang, Yihe Zhou, Jianwei Lv, Tongya Zheng, et al. Breaking the exploration bottleneck: Rubric-scaffolded reinforcement learning for general llm reasoning. *arXiv preprint arXiv:2508.16949*, 2025.

A Algorithm and Training Procedure

A.1 CritSel-OPSD Algorithm

Algorithm 1 summarizes the two-stage loop in Figure 2. The teacher and reference parameters remain frozen throughout; only the student is updated. For a final partial block, $m = \min(H, U-h)$ replaces H in the block-size calculation.

Algorithm 1: CritSel-OPSD: Dynamic Criterion Selection and On-Policy Self-Distillation

Input: Training stream $\mathcal{D} = \{(x_i, R_i)\}$; initial parameters θ_0; criterion budget K; refresh interval H; batch size B; successful-update budget U; loss weights $\lambda_{\text{OPSD}}, \lambda_{\text{KL}}$.
Output: Trained rubric-free student π_θ.

$\theta \leftarrow \theta_0$; $\pi_T, \pi_{\text{ref}} \leftarrow \pi_{\theta_0}$ (frozen); $h \leftarrow 0$;
while $h < U$ **do**
  $m \leftarrow \min(H, U-h)$; synchronize rollout workers to θ_h;
  Stage A: Select privileged teacher context at θ_h.;
  $\mathcal{B}_h \leftarrow [\,]$;
  while $|\mathcal{B}_h| < mB$ **do**
    $(x_i, R_i) \leftarrow \text{Next}(\mathcal{D})$; $M_i \leftarrow |R_i|$;
    Generate greedy rubric-free $y_i^{(h)}$ using **Eq. (2)**;
    if *no valid answer span in* $y_i^{(h)}$ **then continue**;
    if $M_i \le K$ **then**
      $R_{i,h}^{\text{sel}} \leftarrow R_i$; `// skip attribution`
    else
      Score fixed $y_i^{(h)}$ with the full-rubric teacher $q_{i,t}^{R}$;
      for each $j \in \{1, \dots, M_i\}$ **do**
        Score the same $y_i^{(h)}$ under R_i^{-j} to obtain $q_{i,t}^{-j}$;
        Compute $s_{i,j}^{\text{STI}}(\theta_h)$ using **Eq. (7)**;
      Select $R_{i,h}^{\text{sel}}$ using **Eq. (6)**; restore original criterion order;
    Append $(x_i, R_{i,h}^{\text{sel}})$ to $\mathcal{B}_h$;
  Validate and atomically activate the complete selected-rubric snapshot;
  Stage B: Run m successful ordinary OPSD updates.;
  $u \leftarrow h$;
  while $u < h + m$ **do**
    $\mathcal{M} \leftarrow$ batch $(u-h+1)$ of $\mathcal{B}_h$, containing B examples;
    for each $(x_i, R_{i,h}^{\text{sel}}) \in \mathcal{M}$ **do**
      Sample fresh $\hat{y}_i \sim \pi_\theta(\cdot \mid x_i)$; `// no calibration replay`
      Evaluate $p_{i,t}$, $q_{i,t}^{\text{sel}}$, and π_{ref} on the same prefixes;
      Compute $\mathcal{L}_i(\theta)$ using **Eq. (11)**;
    Attempt a student-only optimizer step on $\frac{1}{B}\sum_{i \in \mathcal{M}} \mathcal{L}_i(\theta)$;
    if *the optimizer step succeeds* **then** $u \leftarrow u + 1$;
  $h \leftarrow u$; `// refresh using the updated student`
return π_θ;

A.2 Blockwise Refresh Semantics

At a refresh boundary h, rollout workers are synchronized to the student parameters θ_h. The next block contains HB accepted examples, where B is the training batch size and H is the number of successful optimizer updates per block. Each example is assigned a rubric-free greedy calibration response with an identifiable answer span; invalid examples are replaced from the same training stream. Examples with $M_i > K$ undergo Full+LOO attribution, whereas examples with $M_i \le K$ retain their complete rubric without teacher attribution evaluations.

The selected-rubric snapshot is activated only after all block artifacts have been persisted and validated, so optimization never consumes a partially prepared block. The snapshot, data-stream offset, and successful-update counter are checkpointed together. Within the block, training uses fresh stochastic student responses rather than replaying calibration responses. Refresh boundaries are

Table 6: Information required at each calibration answer position. Teacher quantities are evaluated under both full-rubric and leave-one-out contexts.

Score	teacher information	student information
STI	Realized-token log-probabilities	Realized token
TI-JSD	Full-vocabulary distributions	Realized prefix
SI-JSD	Full-vocabulary distributions	Full-vocabulary distribution

indexed by successful optimizer updates, not trainer-loop iterations; skipped or failed updates therefore do not advance the refresh schedule.

A.3 Computational Cost

Let $\mathcal{B}_h$ denote the accepted examples prepared at boundary h, with $|\mathcal{B}_h| = HB$. The number of teacher-context evaluations used exclusively for attribution is

$$N_{T,h}^{\text{attr}} = \sum_{i \in \mathcal{B}_h} \mathbf{1}\{M_i > K\}(M_i + 1). \tag{12}$$

Each eligible example requires one full-rubric evaluation and M_i leave-one-out evaluations on its fixed calibration sequence. This count excludes calibration generation, student evaluations needed by SI-JSD, and the ordinary OPSD training computation.

Thus the number of attribution contexts is $O(M_i)$ per eligible example; this is not a complete statement about FLOPs or wall-clock complexity. Evaluation cost also depends on rubric and response lengths, model size, batching, and implementation details. The HB accepted examples require HB retained calibration responses; invalid candidates can incur additional generation work.

For a fixed eligible example, reducing K does not reduce the M_i+1 evaluations: every criterion must be scored before ranking. Increasing K can eliminate attribution only when the example satisfies $M_i \leq K$. Likewise, increasing H groups more examples into each block rather than reducing the per-example intervention count by a factor of H. The refresh interval therefore controls selection staleness and block-management frequency, not an asymptotic reduction in attribution evaluations per example.

Attribution statistics and runtime scope. Table 6 distinguishes the distributional information needed to compute each score. All three scores use the same full-rubric and leave-one-out teacher contexts and answer positions on the fixed calibration trajectory. STI requires only $\log q_{i,t}^{R}(y_{i,t}^{(h)})$ and $\log q_{i,t}^{-j}(y_{i,t}^{(h)})$ to form its score, whereas SI-JSD additionally evaluates the current student distribution on the calibration prefixes. These information requirements do not by themselves establish a proportional reduction in model forward-pass cost.

Our diagnostic implementation computes STI, TI-JSD, and SI-JSD together in the same attribution pass for controlled comparison. Consequently, we do not claim an end-to-end runtime advantage for STI in this implementation. Its distinguishing property here is the locality of the ranking statistic, not a measured speedup.

B Attribution Objectives and Diagnostics

Let $n_i = |A_i^{(h)}|$ denote the number of calibration answer tokens. Throughout, JSD uses equal mixture weights:

$$D_{\text{JSD}}(P, Q) = \tfrac{1}{2} D_{\text{KL}}(P \| P_{\text{mix}}) + \tfrac{1}{2} D_{\text{KL}}(Q \| P_{\text{mix}}), \qquad P_{\text{mix}} = \tfrac{1}{2}(P + Q). \tag{13}$$

We define two alternatives to STI using the same full-rubric and leave-one-out contexts and calibration trajectory. First, **Teacher Influence JSD** (TI-JSD) measures full teacher-distributional sensitivity:

$$s_{i,j}^{\text{TI}}(\theta_h) = \frac{1}{n_i} \sum_{t \in A_i^{(h)}} D_{\text{JSD}}\left(q_{i,t}^{R}, q_{i,t}^{-j}\right). \tag{14}$$

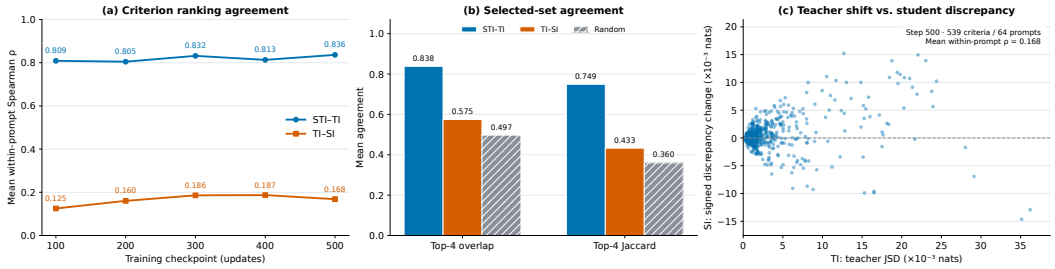

Figure 5: Criterion attribution on 64 fixed probe prompts. (a) Mean within-prompt Spearman rank correlation across checkpoints. (b) STI–TI Top-4 overlap and Jaccard agreement, with analytical random-selection baselines using each rubric's size. (c) TI versus signed SI for 539 criteria at update 500, in 10^{-3} nats. The annotated correlation is the mean within-prompt Spearman correlation, not a pooled correlation of the plotted points. Checkpoints come from one training run, not independent training replicates.

TI-JSD measures sensitivity over the full vocabulary rather than only the realized-token slice.

Second, define the current student distribution on the same calibration prefix as

$$p_{i,t}^{(h)}(v) = \pi_{\theta_h}\left(v \mid x_i, y_{i,<t}^{(h)}\right). \tag{15}$$

We define **Student-Relative Influence JSD** (SI-JSD) by

$$s_{i,j}^{\mathrm{SI}}(\theta_h) = \frac{1}{n_i} \sum_{t \in A_i^{(h)}} \left[D_{\mathrm{JSD}}\left(q_{i,t}^{R}, p_{i,t}^{(h)}\right) - D_{\mathrm{JSD}}\left(q_{i,t}^{-j}, p_{i,t}^{(h)}\right) \right]. \tag{16}$$

SI-JSD is a student-relative sensitivity score; it is not a direct measure of learning progress or downstream quality. STI, TI-JSD, and SI-JSD share the same intervention structure but summarize different changes.

For either ablation, we replace $s_{i,j}^{\mathrm{STI}}(\theta_h)$ in the Top-K rule with the corresponding score above and retain the criteria with the largest scores. For SI-JSD, ranking uses the signed score, not its absolute value. The criterion budget, refresh schedule, fresh training rollouts, and OPSD objective remain unchanged.

Random and static controls. All controls retain at most $K = 4$ criteria per example. Dynamic Random-4 uses the same $H = 100$ block-preparation and training pipeline as dynamic STI, but selects criteria uniformly without replacement using a deterministic seed derived from the training seed, refresh step, and question key; attribution diagnostics are still computed but do not determine the selected context. Static Random-4 instead uses a fixed, precomputed selection cache, with per-example randomness derived from selection seed 0, the question key, and the full-rubric hash. Static Original-Order Top-4 uses a fixed original-order selection cache prepared before training. Both static controls keep their cached subsets unchanged throughout training, without calibration rollouts or leave-one-out scoring. All controls preserve the original order of retained criteria and retain the full rubric when $M_i \leq K$.

B.1 Attribution Agreement

STI–TI agreement remains high across the five fixed-probe checkpoints, whereas TI–SI agreement is substantially lower. These results support treating STI, TI-JSD, and SI-JSD as different attribution objectives rather than interchangeable numerical variants. The random-selection derivation and probe identity checks are given in Appendix D.4.

C Additional Selected-Context Analysis

We separate pooled training-context composition from longitudinal evidence on a fixed probe panel.

Table 7: Mean within-prompt Spearman correlations on the fixed probe panel. Checkpoints are from a single STI training trajectory.

Update	$\rho(\mathrm{STI}, \mathrm{TI})$	$\rho(\mathrm{TI}, \mathrm{SI})$
100	0.809	0.125
200	0.805	0.160
300	0.832	0.186
400	0.813	0.187
500	0.836	0.168

Table 8: Top-4 subset agreement. O denotes overlap and J denotes Jaccard similarity. The final row weights the five checkpoints equally.

Update	STI–TI O	STI–TI J	TI–SI O	TI–SI J
100	0.836	0.746	0.574	0.428
200	0.828	0.733	0.566	0.423
300	0.844	0.758	0.582	0.440
400	0.832	0.742	0.570	0.432
500	0.852	0.767	0.582	0.440
Mean	0.838	0.749	0.575	0.433

C.1 Training Context Composition

Training selection rows are joined to rollout rows using $(\texttt{refresh_step}, \texttt{sample_id})$, because sample IDs restart within each refresh block. Probe criterion identity is represented by $(\texttt{prompt_index}, \texttt{criterion_index})$; criterion text and criterion ID are retained as metadata because duplicate text can occur at different rubric positions. Probe-panel construction, identity checks, and bootstrap details are given in Appendix D.4. Top-4 ranking uses STI descending order and criterion index ascending as the tie-break.

Table 9: Training-context summary statistics. Overall retention is the mean per-example fraction $N^{-1}\sum_{i=1}^{N}\min(4, M_i)/M_i$, not the ratio of mean selected size to mean rubric size. Eligible-only retention averages this fraction over examples with $M_i > 4$.

Statistic	Value
Training examples	32,000
Eligible examples ($M > 4$)	31,875 (99.61%)
Mean original rubric size	8.5271
Mean selected size	3.9998
Overall retention	0.4908
Eligible-only retention	0.4888
Original-Order Top-4 mean overlap	0.6529
Original-Order Top-4 mean Jaccard	0.5218
Original-Order Top-4 exact match	0.1304
Normalized cutoff-margin median	0.0413

C.2 Criterion Position Bias

C.3 Student-State-Dependent Selection Dynamics

Table 11: Temporal consistency on the fixed 64-prompt probe panel. Turnover is the number of replaced criteria out of four.

Transition	Retention	Jaccard	Turnover (count)
$100 \to 200$	.8633	.7896	.5469
$200 \to 300$	.8867	.8292	.4531
$300 \to 400$	.8789	.8125	.4844
$400 \to 500$	.8828	.8188	.4688
$100 \to 500$	.8203	.7313	.7188
Random-set expectation	.4972	.3599	—

Table 10: Conditional selection rate by original criterion position. The denominator is restricted to examples for which the position exists ($M \geq j$).

Position	Available	Selected	STI rate	Random-K	Excess
1	32,000	26,024	.8133	.4908	+.3224
2	32,000	21,064	.6583	.4908	+.1674
3	32,000	19,466	.6083	.4908	+.1175
4	31,993	17,182	.5371	.4907	+.0464
5	31,875	13,891	.4358	.4888	−.0530
6	30,668	10,922	.3561	.4766	−.1204
7	27,811	8,183	.2942	.4570	−.1628
8	22,701	5,621	.2476	.4313	−.1837
9	15,890	2,924	.1840	.4018	−.2178
10	12,112	1,971	.1627	.3885	−.2258
11	3,816	745	.1952	.3636	−.1684

Table 12: Criterion selection frequency across the five fixed-probe checkpoints.

Selected checkpoints	Criteria	Fraction
0/5	203	37.6623%
1/5	52	9.6475%
2/5	29	5.3803%
3/5	30	5.5659%
4/5	45	8.3488%
5/5	180	33.3952%

For examples with $M_i > 4$, let $s_{i,(1)} \geq s_{i,(2)} \geq \cdots \geq s_{i,(M_i)}$ denote the STI scores in descending order. The raw Top-4 cutoff margin and its normalized form are

$$\Delta_i^4 = s_{i,(4)} - s_{i,(5)}, \qquad \widetilde{\Delta}_i^4 = \frac{s_{i,(4)} - s_{i,(5)}}{s_{i,(1)} + \epsilon}, \quad \epsilon = 10^{-12}. \tag{17}$$

The normalized STI cutoff margin has median 0.0413, and only 0.0847% of eligible examples have an exact zero margin. The selected-minus-non-selected STI gap remains between 0.039 and 0.045 across the five probe checkpoints; this is descriptive concentration induced by Top-K selection, not an independent effectiveness test.

D Experimental Details

D.1 Data Preparation and Training Configuration

The reported training configuration uses 32,000 examples obtained after processing the original OpenRubrics dataset: invalid rubrics are removed, normalized prompts are deduplicated, and examples originating from IFEval are excluded. Default runs use a global batch size of 64 prompts and 500 successful optimizer steps. Calibration uses greedy decoding, whereas optimization uses temperature 1.0 and top-p 1.0. The teacher and reference parameters remain frozen. Block preparation and successful-update accounting follow Appendix A.2.

Optimizer and training seed. We use PyTorch AdamW with a constant learning rate of 10^{-6}, no warmup, weight decay 0.01, Adam betas $(0.9, 0.999)$, and a gradient-clipping threshold of 1.0. Dynamic runs use training seed 0, distinct from the answer-generation seeds used for downstream evaluation.

Sequence budgets and hardware. The prompt length limit is 8,192 tokens, and the student response length limit is 16,384 tokens. Scoring and teacher sequences use a budget of 32,768 tokens. Training uses two nodes with 16 devices per node, for 32 devices in total.

Loss configuration. The objective in Eq. 11 is $\mathcal{L} = 0.1\mathcal{L}_{\mathrm{OPSD}} + 2.5 \times 10^{-4}\mathcal{L}_{\mathrm{KLref}}$, where $\mathcal{L}_{\mathrm{OPSD}}$ is the answer-token JSD loss with mixing coefficient 0.5, and $\mathcal{L}_{\mathrm{KLref}}$ is the full-response reference KL regularizer.

D.2 Training and Attribution Prompt Templates

The question is taken directly from the dataset's prompt field. The templates below show the system and user messages before applying the model's chat template; angle-bracketed fields are placeholders.

Student prompt. Both training rollouts and refresh-stage calibration rollouts use the following rubric-free messages:

> **System:** You are a helpful assistant. Answer the user's question as well as possible.
>
> **User:**

```
Question: <question>

Provide your answer.
```

Qwen3 thinking remains enabled. Calibration uses greedy decoding (temperature 0.0), whereas training rollouts use temperature 1.0.

Teacher prompt. During training, the frozen teacher uses the same question with the selected rubric inserted into its user message:

> **System:** You are a helpful assistant. Answer the user's question as well as possible. A rubric describing what a good answer looks like is provided for your reference.
>
> **User:**

```
Question: <question>

Reference rubric for a good answer:
<selected_rubric_json>

Provide your answer.
```

The rubric is serialized as JSON rather than a free-form prose block. For the default Top-4 setting, its displayed structure is:

```
{
  "criteria": [
    {"criterion": "<first retained criterion text>"},
    {"criterion": "<second retained criterion text>"},
    {"criterion": "<third retained criterion text>"},
    {"criterion": "<fourth retained criterion text>"}
  ]
}
```

The retained criteria remain in their original order; rubrics with fewer than four criteria retain all available criteria.

Full-rubric and leave-one-out scoring prompts. At each attribution refresh, the student generates one rubric-free calibration response per question. The full-rubric teacher and each leave-one-out teacher use the teacher template above, replacing the rubric placeholder with the complete rubric or the rubric with criterion j removed, respectively. All teacher contexts score the same student-generated response on its answer-token positions; no separate response is generated for each intervention. Thus, a rubric with 11 criteria requires one full-rubric and 11 leave-one-out teacher contexts, evaluated on one shared student response.

D.3 Benchmark Coverage and Evaluation Protocols

All downstream evaluations are rubric-free. The configuration details below were checked against the saved evaluation configurations for the STI checkpoint at update 500 and its three decoding seeds, 42, 1348, and 4527.

Datasets and evaluation scope. Table 13 records the actual dataset identifiers, released split names, evaluated sample counts, and few-shot settings. All six configurations use `limit=null`,

with no evaluation sample cap. The `train` split names for IFEval and GPQA refer to the datasets' published evaluation splits; they do not indicate that these examples were used for OPSD training. BBH's 23 task families are instantiated as 27 evaluated configurations.

Table 13: Audited six-benchmark evaluation configuration for STI at update 500. Dataset split names are reproduced literally. Metric definitions and parsing details are given below.

Benchmark	Dataset	Subset	Split	Examples	Shots
GSM8K	`openai/gsm8k`	`main`	test	1,319	0
BBH	`lukaemon/bbh`	All 27 configurations	test	6,511	0
IFEval	`HuggingFaceH4/ifeval`	—	train	541	0
GPQA	`Idavidrein/gpqa`	`gpqa_main`	train	448	0
MBPP+	`evalplus/mbppplus`	—	test	378	0
PopQA	`akariasai/PopQA`	—	test	14,267	15

Benchmark metrics. We use the benchmark evaluators with the following reported metrics: GSM8K `exact_match_flex`; the official micro-aggregated `exact_match` metric for BBH; IFEval `prompt_level_loose_acc`; GPQA `exact_match`; MBPP+ pass@1 from 10 generated candidates per task; and PopQA alias-based exact matching. The benchmark table records dataset identifiers, splits, sample counts, and few-shot settings.

Shared generation settings. All six benchmarks use chat formatting (`use_chat_format=true`), stochastic sampling (`do_sample=true`), temperature 0.6, and top-p 0.95. We retain Qwen3's default thinking behavior and do not set `enable_thinking=false`. Answer-generation seeds are 42, 1348, and 4527, while the few-shot seed remains 42 in all three evaluations. MBPP+ uses 10 candidates per task per seed, and the other five benchmarks use one. Input truncation is disabled (`truncate_context=false`) and there are no additional stop sequences.

Length budget. Generation uses the evaluator-defined 32,768-token context and maximum length, with input and output sharing the available context.

Aggregation and statistical scope. Each of the six benchmark scores averages the three decoding-seed results for one trained checkpoint. The primary aggregate is

$$\text{Avg-6} = \frac{\text{GSM8K} + \text{BBH} + \text{IFEval} + \text{GPQA} + \text{MBPP+} + \text{PopQA}}{6}. \tag{18}$$

It is computed from the displayed benchmark means before rounding the aggregate to two decimals. All scores are percentages and higher is better. Because the benchmarks assess distinct capabilities with different scoring rules, Avg-6 is a descriptive cross-task summary rather than a unified reasoning metric. The three decoding seeds are used to reduce sampling noise for each checkpoint.

D.4 Probe Panel and Random-Selection Baselines

Panel construction and scope. The fixed 64-prompt panel was constructed post hoc from the R0 OpenRubrics training pool for offline attribution diagnostics; it is not an independently held-out evaluation set or a probe used to drive online selection. We scan the R0 cache in row order, require nonempty questions and rubrics, deduplicate by question, and retain the first 64 records with parseable rubrics containing 2–11 criteria. Construction is deterministic, uses no additional random selection seed, and does not prefilter for $M_i > 4$. The resulting panel contains 539 criteria, with rubric sizes ranging from 5 to 11; the size frequencies are 3, 7, 10, 14, 5, 20, and 5 for sizes 5 through 11, respectively.

Checkpoint alignment and aggregation. Across checkpoints 100, 200, 300, 400, and 500, we verified identical prompt identities, rubric hashes, criterion indices, criterion IDs, criterion texts, and criterion counts for all 64 probes. Each prompt has exactly four selected criteria at every checkpoint, and all STI scores are finite. At each checkpoint, the student generates a fresh greedy, rubric-free

response; all teacher interventions for that prompt score the same response's answer tokens. Correlations are calculated within each prompt before averaging; subset statistics are likewise averaged across prompts, with equal weighting across checkpoints for the final summary. Spearman correlations use average ranks for ties; undefined correlations are tracked through coverage rather than replaced with zero.

Bootstrap analysis. In the probe analysis, we compute 95% percentile confidence intervals using prompt-level nonparametric bootstrap with 10,000 repetitions and seed 20260923. Each replicate samples 64 prompts with replacement, retaining each sampled prompt's complete criterion structure and its paired observations across checkpoints. Intervals are computed for cross-checkpoint Top-4 retention, Jaccard similarity, turnover, within-prompt Spearman correlations across checkpoints, and raw and normalized selected/non-selected STI gaps. The tables and figures in this paper report point estimates rather than these intervals. This analysis conditions on one training run and does not quantify variation across independent training runs.

Random-selection baselines. For two size-K criterion-index sets S, T, define

$$\mathrm{Overlap}(S,T) = \frac{|S \cap T|}{K}, \qquad \mathrm{Jaccard}(S,T) = \frac{|S \cap T|}{|S \cup T|}. \tag{19}$$

For a uniformly sampled size-K subset against a fixed size-K subset of a rubric with M_i criteria, their intersection size X_i follows

$$\Pr(X_i = x) = \frac{\binom{K}{x}\binom{M_i - K}{K - x}}{\binom{M_i}{K}}. \tag{20}$$

Consequently,

$$\mathbb{E}[\mathrm{Overlap}_i] = \frac{K}{M_i}, \qquad \mathbb{E}[\mathrm{Jaccard}_i] = \sum_{x=\max(0,2K-M_i)}^{K} \Pr(X_i = x)\frac{x}{2K - x}. \tag{21}$$

At $K = 4$, averaging over the actual rubric sizes gives expected overlap 0.497209 and expected Jaccard 0.359925. The expected Jaccard is computed directly, not as a ratio obtained by substituting the expected intersection size. These baselines are constant across checkpoints.

E Reference Regularization

Let $\mathcal{R}(\hat{y})$ denote all generated response positions, including reasoning tokens. For the realized token $\hat{y}_t$, define

$$d_t = \mathrm{clip}\left(\log \pi_{\mathrm{ref}}(\hat{y}_t \mid x, \hat{y}_{<t}) - \log \pi_\theta(\hat{y}_t \mid x, \hat{y}_{<t}), -20, 20\right), \tag{22}$$

$$\ell_{\mathrm{KLref},t} = \mathrm{clip}\left(e^{d_t} - d_t - 1, -10, 10\right), \tag{23}$$

$$\mathcal{L}_{\mathrm{KLref}} = \frac{1}{|\mathcal{R}(\hat{y})|} \sum_{t \in \mathcal{R}(\hat{y})} \ell_{\mathrm{KLref},t}. \tag{24}$$

We use the sampled low-variance KL form employed in Group Relative Policy Optimization (GRPO)-style reference regularization (Shao et al., 2024). The clipping bounds are numerical safeguards in our implementation. Unlike the answer-only distillation term, this regularizer covers the full generated response, including reasoning positions that are excluded from the privileged-teacher distillation loss.