PERSISTENT NEGATIVES FOR ADVERSARIAL BLACK-BOX ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Black-box On-Policy Distillation (OPD) seeks to improve a student from its own generations when the teacher provides sampled responses but not token probabilities. Adversarial distillation offers one route: it learns a discriminator over prompt-matched teacher and student responses and uses its score as the policy reward. However, sampling discriminator negatives from the latest student at each step couples the learned reward to a negative distribution that changes after every policy update. We address this moving-target problem with persistent-negative adversarial distillation, a live-pool method that replaces a fraction of each discriminator batch with historical, prompt-matched teacher–student comparisons. Under matched discriminator compute, historical comparisons train the discriminator, while GRPO remains on-policy with fresh student responses. Our analysis identifies the Bayes-optimal reward as a teacher-to-negative log-density ratio and, under explicit assumptions, shows how persistent negatives anchor the discriminator and reduce reward-estimation MSE relative to fresh-negative training. Across two student families, three judges, and four judged-chat benchmarks, persistent-negative adversarial distillation consistently improves performance over current methods at matched discriminator compute. It also yields smoother fresh-policy discriminator trajectories, with fewer below-chance dips. These findings identify the discriminator's negative distribution as an important design axis in black-box on-policy distillation.

1 INTRODUCTION

Knowledge distillation transfers capabilities from a strong teacher to a smaller student. In the white-box setting, the student can match the teacher's token probabilities or hidden representations (Hinton et al., 2015). Proprietary teachers instead commonly expose only generated text. Sequence-level knowledge distillation (SeqKD; Kim & Rush, 2016) is compatible with this interface, but it trains the student only on teacher trajectories. On-policy distillation addresses this mismatch by learning from the student's own generations, yet established likelihood-based objectives still require teacher probabilities (Gu et al., 2024; Agarwal et al., 2024). The central challenge is therefore how to provide useful feedback on the student's own responses when the teacher supplies only generated text, without token-level probabilities.

Adversarial distillation addresses this gap by learning a discriminator from prompt-matched teacher and student responses and using its scores as sequence-level rewards (Ye et al., 2025). For the same prompt, the discriminator learns to score a teacher response above a student response; its scores then serve as sequence-level rewards for a GRPO update (Shao et al., 2024). This converts text-only teacher outputs into an adaptive policy-learning signal, but it also creates a second estimation problem: the reward model must be learned from a negative distribution over student responses. When all negatives are freshly sampled, that distribution changes after every policy update. Failure modes that are rare in the latest rollout can disappear from discriminator training even when they remain reachable by the student.

Our key observation is that these two notions of on-policy data are distinct. The policy update must use responses sampled from the current student, but the discriminator need not discard every valid comparison from earlier students. We therefore introduce *persistent-negative adversarial distillation*. Our primary live-pool method maintains a bounded collection of complete, prompt-matched

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.

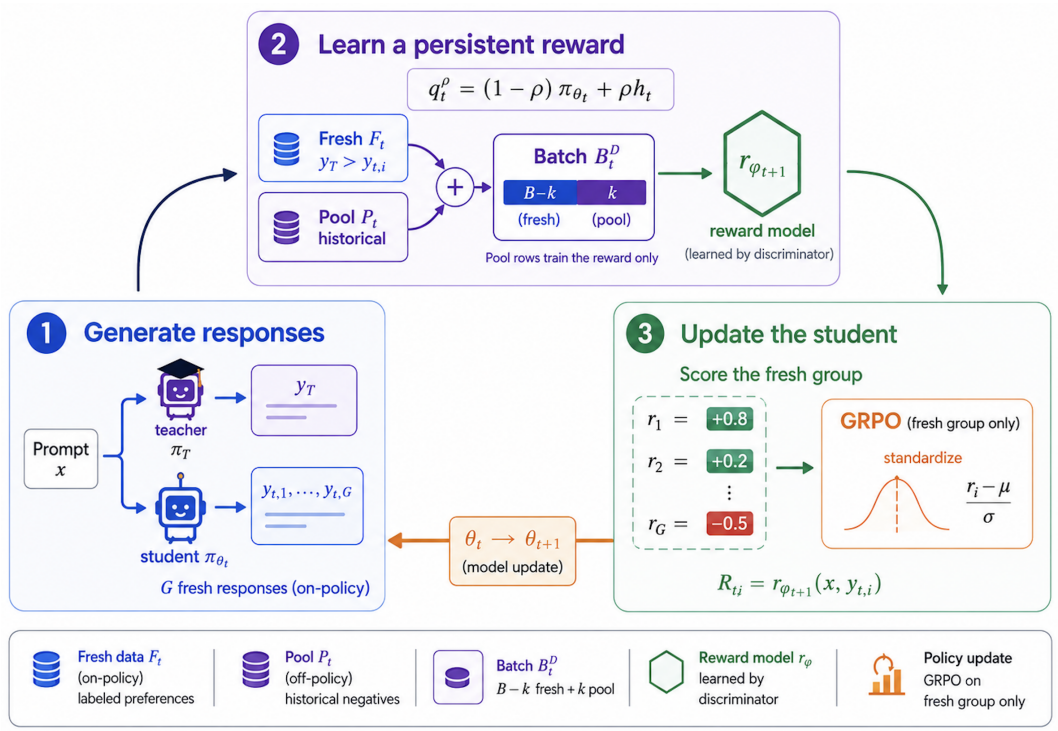


Figure 1: **Workflow of persistent-negative on-policy distillation.** At iteration t, each prompt x is paired with a teacher response y_T and a group of fresh student responses $y_{t,1:G}$. Fresh comparisons F_t are combined with historical comparisons from P_t to form the fixed-size discriminator batch B_t^D, from which the updated reward $r_{\varphi_{t+1}}$ is learned. The discriminator then scores the fresh student group, and the resulting within-group advantages drive the GRPO update $\theta_t \to \theta_{t+1}$ for student.

teacher–student comparisons and replaces a fixed fraction of each discriminator batch with historical comparisons. The updated discriminator then scores only fresh current-student responses for GRPO, so historical data affect the policy exclusively through the learned reward and the student update remains on-policy. Figure 1 illustrates the complete training loop. At each iteration, the current student generates fresh responses; fresh and historical prompt-matched comparisons train the discriminator; and the updated discriminator supplies rewards for a GRPO update computed from fresh student rollouts. The resulting student then generates the responses used in the next iteration.

This design admits a theoretical account. First, the Bayes-optimal Bradley–Terry score (Bradley & Terry, 1952) is the log-density ratio between the teacher distribution and the negative distribution, so changing the negative pool changes the learned reward rather than merely its offset. Second, persistent comparisons provide explicit lower bounds on the historical loss contribution and discriminator curvature. Third, under variance-reduction and bias-control assumptions, the pooled estimator has lower reward MSE, which tightens the error bounds for the standardized advantages and local reward-gradient term used by GRPO.

Using GPT-5 Chat responses as black-box teacher supervision (OpenAI, 2025; Ye et al., 2025), we evaluate Qwen2.5-7B-Instruct (Qwen Team, 2024) and Llama-3.1-8B-Instruct (Grattafiori et al., 2024) students across three judges and four judged-chat benchmarks. At matched discriminator compute, persistent-negative training outperforms current-only training (Ye et al., 2025) in cross student–judge–dataset evaluation, with equal-set-weighted gains of $+1.0$ to $+2.0$ percentage points. The resulting students also improve over their undistilled initializations by $+1.7$ to $+5.3$ points. In addition, persistent negatives reduce the temporal standard deviation of fresh-policy discriminator accuracy and reduce below-chance dips from six to one.

Our contributions are:

- We identify the discriminator's negative distribution as a first-class design choice in black-box adversarial distillation and introduce persistent-negative training, while keeping GRPO updates fully on-policy and discriminator compute fixed.
- We characterize the Bayes-optimal reward as a teacher-to-negative density ratio and establish explicit historical-loss and curvature bounds, together with a bias–variance condition under which persistent negatives reduce reward MSE.
- Experiments across two student families, three judges, and both in-domain and cross-dataset evaluations show broadly consistent gains over current-only training, together with reduced discriminator volatility and fewer below-chance dips.

2 Related Work

Black-box and on-policy distillation. Classical knowledge distillation matches teacher output distributions (Hinton et al., 2015), whereas response-based methods learn from sampled teacher text (Kim & Rush, 2016; Jiang et al., 2023). On-policy distillation trains on student-generated trajectories, but conventional objectives require teacher probabilities on those trajectories (Gu et al., 2024; Agarwal et al., 2024; Lu & Thinking Machines Lab, 2025). Black-box alternatives instead derive sequence-level supervision from text: OVD uses teacher-provided verbal scores (Xiong et al., 2026), ROPD constructs prompt-specific rubrics (Fang et al., 2026), and GAD learns an adaptive discriminator from teacher–student responses (Ye et al., 2025). We develop the discriminator-based approach by introducing persistent comparisons that stabilize its learned reward while preserving fresh on-policy student updates. In addition, we focus on an orthogonal question within discriminator-based distillation: which student-response distribution should train the learned reward?

Adversarial reward learning. Generative adversarial learning trains a discriminator to distinguish samples from two distributions (Goodfellow et al., 2014), while adversarial imitation learning uses the resulting discriminator as a policy reward (Ho & Ermon, 2016). Under logistic classification, the optimal discriminator logit recovers a log-density ratio between the positive and negative distributions. GAD applies this principle to teacher and student responses. Our analysis specializes it to prompt-conditional language generation and shows that the discriminator's reward depends explicitly on the student distribution used to construct its negatives.

Historical and mixed-policy negatives. History buffers and replay have been used to stabilize adversarial image generation and imitation learning (Shrivastava et al., 2017; Kostrikov et al., 2019). ORPO-Distill similarly mixes responses from different student policies when constructing preference pairs (Singh et al., 2025). In ORPO-Distill, these responses directly enter the student's preference objective. Our historical responses instead remain paired with their prompts and teacher responses and train only the discriminator; GRPO continues to score and optimize fresh responses sampled from the current student.

3 Persistent-Negative Adversarial Distillation

Figure 1 provides an overview of the alternating training procedure. We now formalize its discriminator, negative distribution, and policy updates.

3.1 Adversarial reward formulation

Let $\mathcal{D}$ be a prompt distribution, $\tau(\cdot \mid x)$ the teacher-response distribution, and $\pi_{\theta_t}(\cdot \mid x)$ the student policy at iteration t. We observe teacher-response samples $\mathcal{T} = \{(x, y_T)\}$, where $x \sim \mathcal{D}$ and $y_T \sim \tau(\cdot \mid x)$, but cannot query teacher token probabilities. We therefore learn a sequence-level score $r_\varphi(x, y)$ from prompt-matched teacher and student responses.

This construction connects adversarial generation to policy distillation. In a GAN, a discriminator separates observed from generated samples while the generator changes its distribution to become harder to distinguish (Goodfellow et al., 2014). Generative Adversarial Distillation (GAD; Ye et al., 2025) instantiates this game for conditional text: teacher responses play the role of observed samples and the autoregressive student plays the generator. For any negative response distribution $q(\cdot \mid x)$,

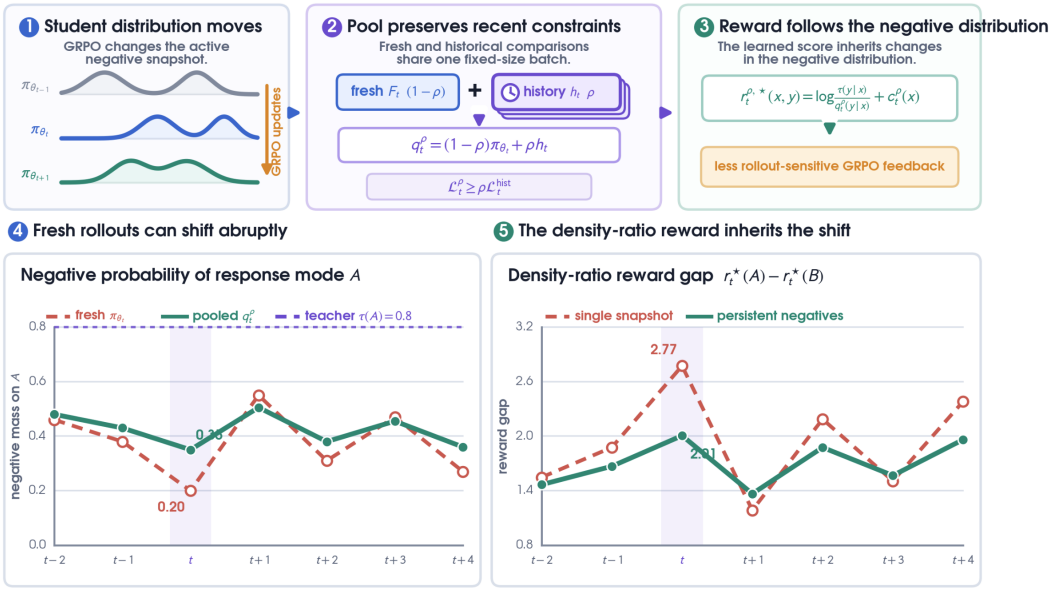


Figure 2: **Why persistent negatives help.** GRPO updates continually change the student distribution, so a discriminator trained from a single policy snapshot can lose constraints supplied by earlier student responses. Persistent-negative training uses $q_t^\rho = (1-\rho)\pi_{\theta_t} + \rho h_t$, allowing fresh and historical comparisons to share a fixed-size discriminator batch. The historical term remains active in the discriminator objective, making the learned reward less dependent on any single rollout and providing steadier feedback for GRPO. Reward trajectories are schematic.

define the population Bradley–Terry loss (Bradley & Terry, 1952)

$$\mathcal{L}_D(r_\varphi; q) = \mathbb{E}_{\substack{x\sim\mathcal{D},\, y_T\sim\tau(\cdot|x),\\ y_S\sim q(\cdot|x)}} \left[\mathrm{softplus}(r_\varphi(x, y_S) - r_\varphi(x, y_T))\right]. \tag{1}$$

The discriminator minimizes this loss, increasing the score margin between teacher and student responses. When $q = \pi_\theta$, the corresponding response-level minimax game is

$$\min_\varphi \max_\theta \; \mathcal{V}(\theta, \varphi) = \mathcal{L}_D(r_\varphi; \pi_\theta). \tag{2}$$

The student maximization pushes probability toward responses receiving higher discriminator scores. Because sampled text is discrete, we use the standard non-saturating surrogate: the discriminator score becomes a sequence-level reward, and GRPO performs the student update instead of differentiating through generated tokens (Shao et al., 2024). Concretely, the updated discriminator scores a fresh group $y_{t,i} \sim \pi_{\theta_t}(\cdot \mid x)$, $i = 1, \ldots, G$, producing $R_{t,i} = r_{\varphi_{t+1}}(x, y_{t,i})$. GRPO forms the promptwise advantages

$$A_{t,i} = \frac{R_{t,i} - \bar{R}_t}{s_{R,t} + \varepsilon}, \bar{R}_t = G^{-1} \sum_{j=1}^{G} R_{t,j}, s_{R,t} = \mathrm{std}_j(R_{t,j}), \tag{3}$$

and applies the standard clipped policy objective with a KL penalty to a frozen reference policy (Ye et al., 2025). Only fresh student responses enter this update.

Why the reward transfers teacher behavior. Theorem 4.1 shows that the population-optimal reward is $r_q^\star(x, y) = \log[\tau(y \mid x)/q(y \mid x)] + c(x)$. So it favors responses characteristic of the teacher relative to the student negatives. GRPO increases the probability of fresh responses receiving higher within-group rewards, transferring teacher-compatible behavior without token-level supervision.

3.2 Persistent negatives for a moving reward target

The negative distribution in Eq. (1) determines the reward being learned. Using only $q = \pi_{\theta_t}$ makes this target move after every policy update: response modes absent from the latest rollout stop

Algorithm 1 Persistent-negative adversarial distillation

Require: Teacher-response data $\mathcal{T}$, student π_{θ_0}, and score model r_{φ_0}
Require: Pool P_0, ρ, B, C
for $t = 0, 1, \ldots, N-1$ **do**
  Sample prompt–teacher pairs $(x, y_T) \sim \mathcal{T}$
  Sample fresh groups $y_{t,i} \sim \pi_{\theta_t}(\cdot \mid x)$, $i = 1, \ldots, G$
  Form fresh comparisons $F_t \leftarrow \{(x, y_T, y_{t,i})\}$
  $k_t \leftarrow \min\{\lfloor \rho B \rfloor, |P_t|\}$
  $B_t^D \leftarrow \textsc{Sample}(F_t, B - k_t) \oplus \textsc{Sample}(P_t, k_t)$
  $\varphi_{t+1} \leftarrow \textsc{Disc Update}(\varphi_t, B_t^D)$
  $R_{t,i} \leftarrow r_{\varphi_{t+1}}(x, y_{t,i})$ for fresh responses
  Compute $A_{t,i}$ in (3)
  $\theta_{t+1} \leftarrow \text{GRPO } \textsc{Update}(\theta_t, \{y_{t,i}, A_{t,i}\})$
  $P_{t+1} \leftarrow \textsc{Keep Recent}(P_t \oplus F_t, C)$
end for
return π_{θ_N}

contributing to discriminator training even when the student can produce them again. The learned reward can therefore react strongly to one policy snapshot and forget previously observed failure modes.

We address this moving-target problem with persistent negative distribution. Let P_t be a bounded pool of complete comparisons (x, y_T, y_S), and $h_t(\cdot \mid x)$ be the historical student response distribution represented by that pool. Once the pool is filled, the discriminator negative distribution is

$$q_t^\rho(\cdot \mid x) = (1-\rho)\pi_{\theta_t}(\cdot \mid x) + \rho h_t(\cdot \mid x), \qquad \rho \in [0, 1). \tag{4}$$

Figure 2 summarizes the mechanism. As GRPO updates move the student distribution, the negative distribution constructed from a single policy snapshot also moves. Mixing fresh responses with h_t keeps recent comparisons active in the discriminator objective, which can make the learned reward less sensitive to any individual rollout. Retaining the complete comparison preserves the prompt and teacher response that define the pairwise target for each historical student response. The primary *live pool* retains recent comparisons and refreshes them during training. Pooled comparisons train only r_φ; GRPO continues to use fresh groups sampled from π_{θ_t}. The special case $\rho = 0$ gives the nonpersistent comparator.

3.3 Alternating optimization

Let B be the discriminator batch size, C the pool capacity, and F_t the fresh prompt-matched comparisons constructed at iteration t. We initialize P_0 as empty. During pool startup, $k_t = \min\{\lfloor \rho B \rfloor, |P_t|\}$ pooled rows are available, so the realized mixture weight is $\rho_t = k_t/B$; the remaining slots are filled with fresh comparisons. Algorithm 1 gives the complete alternating update.

Pooled comparisons replace rather than augment fresh comparisons, so every variant uses B discriminator comparisons and the same number of discriminator optimizer steps per iteration. Historical data affect the student only through the updated score model. Moreover, because GRPO removes a promptwise reward offset, persistence changes the policy update only when it changes the ordering or relative spacing of scores within a fresh group. The following section analyzes the resulting reward target, discriminator stability, and GRPO signal.

4 Theoretical Analysis

Our analysis provides three complementary results. The Bayes-optimal Bradley–Terry reward is a teacher-to-negative log-density ratio. Persistent comparisons lower-bound both the historical loss contribution and the discriminator curvature. Finally, a bias–variance condition characterizes when persistence reduces reward MSE. The three results formalize the mechanism illustrated in Figure 2.

Notation. We use distributions and discriminator loss defined in Section 3. In particular, q_t^ρ is given by Eq. (4), and $\mathcal{L}_D(r; q)$ denotes Eq. (1) evaluated with score function r. For an arbitrary negative distribution q, let $r_q^\star \in \arg\min_r \mathcal{L}_D(r; q)$ denote a population-optimal score over the unrestricted

class of real-valued functions. We abbreviate $r_t^\star := r^\star_{\pi_{\theta_t}}$ and $r_t^{\rho,\star} := r^\star_{q_t^\rho}$. Because the loss depends only on score differences, these optima are defined up to a prompt-dependent additive constant.

4.1 The Bayes-optimal score is a density ratio

Theorem 4.1 (Bayes-optimal discriminator score)**.** *Fix a prompt x. Suppose the response space is finite and $\tau(y \mid x), q(y \mid x) > 0$ for every response y. Over the unrestricted score class,*

$$r_q^\star(x,y) = \log \frac{\tau(y \mid x)}{q(y \mid x)} + c_q(x), \tag{5}$$

where $c_q(x)$ is independent of y. Consequently, we have

$$r_t^\star(x,y) = \log \frac{\tau(y \mid x)}{\pi_{\theta_t}(y \mid x)} + c_t(x), \;\; \textit{and} \;\; r_t^{\rho,\star}(x,y) = \log \frac{\tau(y \mid x)}{q_t^\rho(y \mid x)} + c_t^\rho(x). \tag{6}$$

Example. Consider two response modes, A and B, with $\tau(A,B) = (0.8, 0.2)$ and $\pi_{\theta_t}(A,B) = (0.2, 0.8)$. Under fresh-negative training, $r_t^\star(A) - r_t^\star(B) = \log \frac{0.8}{0.2} - \log \frac{0.2}{0.8} = \log 16 \approx 2.77$. If $h_t(A,B) = (0.5, 0.5)$ and $\rho = 0.5$, then $q_t^\rho(A) = 0.5(0.2) + 0.5(0.5) = 0.35$ and $q_t^\rho(B) = 0.5(0.8) + 0.5(0.5) = 0.65$. The corresponding gap is $\log[(0.8/0.35)/(0.2/0.65)] = \log 7.43 \approx 2.01$. Here the current rollout may make the correct mode look unusually rare; the pool remembers that recent students produced it half of the time and avoids turning this transient shift into an excessively large reward gap. It preserves the useful ordering $A > B$ while making the reward less reactive to one policy snapshot. With realistic multi-mode responses, pooling changes different gaps non-uniformly and therefore can also change the standardized GRPO advantages.

4.2 Persistent negatives anchor the discriminator

For a fixed prompt, let scores lie in the centered subspace $\mathcal{S} = \{r : \mathbf{1}^\top r = 0\}$. Define $\mathcal{L}_t^{\mathrm{fresh}}(r) = \mathcal{L}_D(r; \pi_{\theta_t})$ and $\mathcal{L}_t^{\mathrm{hist}}(r) = \mathcal{L}_D(r; h_t)$. At the population level, the negative-mixture distribution gives $\mathcal{L}_t^\rho = (1-\rho)\mathcal{L}_t^{\mathrm{fresh}} + \rho\mathcal{L}_t^{\mathrm{hist}}$.

Proposition 4.2 (Persistent-negative lower bound)**.** *For every score vector r and $\rho \in [0,1)$, we have $\mathcal{L}_t^\rho(r) \geq \rho\mathcal{L}_t^{\mathrm{hist}}(r)$. Moreover, suppose that, on a bounded convex subset of $\mathcal{S}$, $\nabla^2\mathcal{L}_t^{\mathrm{fresh}}(r) \succeq \mu_{\mathrm{fresh}} I_{\mathcal{S}}$ and $\nabla^2\mathcal{L}_t^{\mathrm{hist}}(r) \succeq \mu_{\mathrm{hist}} I_{\mathcal{S}}$, where $\mu_{\mathrm{fresh}}, \mu_{\mathrm{hist}} \geq 0$. Then $\nabla^2\mathcal{L}_t^\rho(r) \succeq \rho\mu_{\mathrm{hist}} I_{\mathcal{S}}$.*

Example. Consider a prompt with three response modes: a correct answer A, a hallucinated answer B, and an unhelpful refusal C. After a policy update, fresh negatives may shift abruptly from B to C. Fresh-only training then provides little signal about B, even though the student may produce it again. A live pool retains earlier teacher–B comparisons. With $\rho = 0.5$, the pooled objective satisfies $\mathcal{L}_t^\rho(r) \geq 0.5\mathcal{L}_t^{\mathrm{hist}}(r)$, so achieving low pooled loss requires continued control of the historical comparisons involving B. If those comparisons provide curvature in the corresponding score direction, they contribute at least $0.5\mu_{\mathrm{hist}}$ to the pooled curvature.

4.3 Bias–variance tradeoff of persistent negatives

For a reward function f, define the promptwise-centered seminorm as $\|f\|_{t,c}^2 := \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_{\theta_t}(\cdot|x)}\left[\Big(f(x,y) - \mathbb{E}_{y'\sim\pi_{\theta_t}(\cdot|x)} f(x,y')\Big)^2\right]$. This removes prompt-dependent score offsets, which do not affect pairwise comparisons or promptwise GRPO advantages.

Proposition 4.3 (Reward MSE under controlled pool bias)**.** *Let $\widehat{r}_t^{\mathrm{fresh}}$ and $\widehat{r}_t^{\mathrm{pool}}$ be learned reward functions, and let $r_t^\star$ be the ideal fresh-negative reward from Theorem 4.1. Define $V_a = \mathbb{E}\|\widehat{r}_t^a - \mathbb{E}\widehat{r}_t^a\|_{t,c}^2$ for $a \in \{\mathrm{fresh}, \mathrm{pool}\}$ and $B_{\mathrm{pool}}^2 = \|\mathbb{E}\widehat{r}_t^{\mathrm{pool}} - r_t^\star\|_{t,c}^2$. Assume that $\mathbb{E}\widehat{r}_t^{\mathrm{fresh}} = r_t^\star$ and that, for some $\Delta_t > 0$, $V_{\mathrm{pool}} \leq V_{\mathrm{fresh}} - \Delta_t$ and $B_{\mathrm{pool}}^2 < \Delta_t$. Then we have,* $\mathrm{MSE}(\widehat{r}_t^{\mathrm{fresh}}, r_t^\star) = V_{\mathrm{fresh}}$, $\mathrm{MSE}(\widehat{r}_t^{\mathrm{pool}}, r_t^\star) = V_{\mathrm{pool}} + B_{\mathrm{pool}}^2$. *Consequently,* $\mathrm{MSE}(\widehat{r}_t^{\mathrm{fresh}}, r_t^\star) - \mathrm{MSE}(\widehat{r}_t^{\mathrm{pool}}, r_t^\star) \geq \Delta_t - B_{\mathrm{pool}}^2 > 0$.

Appendix B provides complete proofs of above theorems and propositions.

5 EXPERIMENTAL SETUP

5.1 MODELS, DATA, AND TRAINING

We evaluate two comparably sized student models from different families: Qwen2.5-7B-Instruct (Qwen Team, 2024) and Llama-3.1-8B-Instruct (Grattafiori et al., 2024), so that we can test whether the effect of persistent negatives transfers across model families rather than depending on a particular architecture or initialization. For each family, we evaluate the initial instruction-tuned checkpoint as the undistilled student.

The teacher-response corpus contains $192{,}014$ responses generated by GPT-5 Chat (OpenAI, 2025) and released with GAD over LMSYS-Chat-1M (Ye et al., 2025; Zheng et al., 2024). We use only sampled teacher responses: we neither query the teacher during training nor access its token probabilities. Training begins with one epoch of supervised fine-tuning on the teacher responses, during which the discriminator is also initialized. This warmup is followed by two adversarial epochs.

Each adversarial step samples eight student responses per prompt. Our persistent-negative configuration uses mixture weight $\rho = 0.5$, uniform pool sampling, and a bounded live pool containing $4,096$ complete comparisons (x, y_T, y_S). Once sufficient history is available, half of each discriminator batch is sampled from the pool. Pooled comparisons replace current comparisons within the fixed discriminator batch size $B = 256$. Consequently, GAD and persistent-negative training use the same number of discriminator examples and optimizer steps. Complete optimization, sequence-length, and hardware configurations appear in Appendix A.

For each student family, we train GAD and our method with three paired random seeds. Within each pair, both methods share the warmup checkpoint, data order, initialization, and optimization settings. Across seeds, we vary data shuffling, response sampling, and optimizer randomness. Both student families use the same, one-epoch supervised warmup, two adversarial epochs, group size $G = 8$, discriminator batch size $B = 256$, mixture weight $\rho = 0.5$, and pool capacity of 4096 comparisons. We change only model-specific tokenization and chat templates. Memory-related microbatching may differ, but the effective batch size is held fixed.

5.2 COMPARED METHODS

For each student model, we compare the undistilled checkpoint, SeqKD (Kim & Rush, 2016), standard GAD (Ye et al., 2025), and our persistent-negative method. SeqKD performs supervised fine-tuning on teacher responses without a discriminator or GRPO. GAD and persistent-negative training resume from the same jointly initialized warmup checkpoint. Standard GAD uses only current-policy negatives, corresponding to $\rho = 0$. Our method instead replaces half of each discriminator batch with historical comparisons sampled from the live pool. Appendix A provides the complete configurations.

5.3 EVALUATION AND INFERENCE

We generate greedy responses for LMSYS ($n = 479$ prompts), Dolly ($n = 500$; Conover et al., 2023), Vicuna/MT-Bench ($n = 80$; Zheng et al., 2023), and Self-Instruct ($n = 252$; Wang et al., 2023). Evaluation prompts are excluded from the distillation cohort.

All distillation prompts are drawn from LMSYS-Chat-1M. The held-out LMSYS evaluation set therefore measures in-domain generalization. Dolly, Vicuna/MT-Bench, and Self-Instruct are not used during training and provide cross-dataset out-of-domain evaluations.

An offline judge assigns scores from 1 to 10 to the candidate and reference answers. Following Ye et al. (2025), we report candidate/(candidate+reference), for which 0.5 denotes parity. We use three judges to reduce dependence on the preferences and calibration of any single evaluator. Qwen2.5-72B-Instruct (Qwen Team, 2024) provides a large, established Qwen-family judge, while Qwen3.5-27B (Qwen Team, 2026) tests consistency under a newer Qwen generation. Gemma-3-27B-IT (Gemma Team et al., 2025) provides a cross-family judge, allowing us to test whether the conclusions persist outside the Qwen model family.

Table 1: **Judged-chat performance with Qwen2.5-7B-Instruct.** Scores are percentages, with 50 denoting parity with the reference. Base is the undistilled student, and GAD uses current-only negatives. Improvements are reported in percentage points (pp). The average weights the four evaluation sets equally.

Dataset	Base	SeqKD	GAD	Persistent-negative (ours)	Δ vs. Base ↑ (pp)	Δ vs. GAD ↑ (pp)
Qwen2.5-72B-Instruct judge						
LMSYS	48.5	48.9	49.1	**50.7**	+2.2	+1.6
Dolly	47.5	46.8	47.9	**49.1**	+1.6	+1.2
Vicuna	48.4	48.6	49.2	**50.2**	+1.8	+1.0
Self-Instruct	48.1	48.5	48.7	**49.4**	+1.3	+0.7
Average	48.1	48.2	48.7	**49.9**	+1.7	+1.1
Qwen3.5-27B judge						
LMSYS	39.1	39.5	39.6	**41.6**	+2.5	+2.0
Dolly	42.4	42.7	42.5	**43.4**	+1.0	+0.9
Vicuna	42.0	44.7	44.3	**45.0**	+3.0	+0.7
Self-Instruct	42.4	43.2	43.9	**45.6**	+3.2	+1.7
Average	41.5	42.5	42.6	**43.9**	+2.4	+1.3
Gemma-3-27B-IT judge						
LMSYS	41.2	42.8	44.6	**47.2**	+6.0	+2.6
Dolly	43.2	43.8	45.1	**46.8**	+3.6	+1.7
Vicuna	42.7	43.6	45.9	**47.5**	+4.8	+1.6
Self-Instruct	42.4	42.7	44.6	**46.3**	+3.9	+1.7
Average	42.4	43.2	45.1	**47.0**	+4.6	+1.9

All three judges score the same fixed student responses using the same evaluation rubric and deterministic decoding. Because their absolute scoring scales need not be calibrated, we report results separately for each judge and compare methods only within judge. The offline judges are independent of the discriminator used during training.

We also measure pre-update discriminator accuracy on fresh current-policy comparisons. Every arm uses the same prompt–teacher slice, but student responses are generated by the corresponding current policy. This diagnostic therefore describes the coupled student–discriminator process rather than isolating discriminator variation alone.

6 Results

6.1 Persistent negatives improve judged-chat performance

Persistent-negative training improves judged-chat performance consistently across student and judge families. With Qwen2.5-7B-Instruct as the student, our method outperforms GAD on every evaluation set under all three judges (Table 1). The equal-set-weighted gains over GAD are $+1.1\%$ under Qwen2.5-72B-Instruct, $+1.3\%$ under Qwen3.5-27B, and $+1.9\%$ under Gemma-3-27B-IT. Relative to the undistilled student, the corresponding gains are $+1.7\%$, $+2.4\%$, and $+4.6\%$.

The improvement is not limited to the LMSYS evaluation set, which shares the source domain of the training corpus. Persistent-negative training also outperforms GAD on Dolly, Vicuna, and Self-Instruct, providing evidence of cross-dataset generalization. Although the judges use different absolute score scales, they agree on the direction of every comparison.

The same pattern transfers to Llama-3.1-8B-Instruct (Table 2). Persistent-negative training improves the equal-set-weighted average over GAD by $+1.0\%$ under Qwen2.5-72B-Instruct, $+1.1\%$ under Qwen3.5-27B, and $+2.0\%$ under Gemma-3-27B-IT. Relative to the undistilled Llama student, the corresponding gains are $+3.1$, $+5.2\%$, and $+5.3\%$. The Qwen student retains higher absolute scores, but the persistent-negative improvement over GAD is similar across the two model families.

Across both students, all three judges favor persistent-negative training over GAD on all four evaluation sets, yielding positive point estimates in all 24 student–judge–dataset comparisons. These com-

Table 2: **Judged-chat performance with Llama-3.1-8B-Instruct.**

Dataset	Base	SeqKD	GAD	Persistent-negative (ours)	Δ vs. Base ↑ (pp)	Δ vs. GAD ↑ (pp)
Qwen2.5-72B-Instruct judge						
LMSYS	46.7	47.2	48.5	**49.4**	+2.7	+0.9
Dolly	45.8	45.9	46.7	**48.2**	+2.4	+1.5
Vicuna	45.2	47.7	48.8	**49.8**	+4.6	+1.0
Self-Instruct	46.3	47.2	48.2	**48.9**	+2.6	+0.7
Average	46.0	47.0	48.1	**49.1**	+3.1	+1.0
Qwen3.5-27B judge						
LMSYS	36.5	37.9	39.6	**40.7**	+4.2	+1.1
Dolly	38.1	39.2	41.2	**42.2**	+4.1	+1.0
Vicuna	37.6	40.0	42.8	**44.1**	+6.5	+1.3
Self-Instruct	37.8	39.6	42.7	**43.8**	+6.0	+1.1
Average	37.5	39.2	41.6	**42.7**	+5.2	+1.1
Gemma-3-27B-IT judge						
LMSYS	39.5	39.8	42.3	**45.6**	+6.1	+3.3
Dolly	40.1	42.0	43.9	**44.9**	+4.8	+1.0
Vicuna	39.8	41.9	43.7	**45.8**	+6.0	+2.1
Self-Instruct	39.9	42.0	42.5	**44.2**	+4.3	+1.7
Average	39.8	41.4	43.1	**45.1**	+5.3	+2.0

parisons are correlated because they share prompts, references, and evaluation rubrics; we therefore treat their directional consistency as robustness evidence rather than as 24 independent tests.

6.2 Persistent negatives produce smoother discriminator trajectories

We measure pre-update discriminator accuracy on fresh current-policy comparisons. In the seed run, persistent-negative training reduces the temporal standard deviation from 0.087 to 0.068, raises the minimum accuracy from 0.414 to 0.461, and reduces below-chance dips from six to one. Because each student generates its own negatives, this diagnostic describes the coupled student–discriminator trajectory rather than the discriminator in isolation. Appendix D reports the complete statistics and common-probe analysis.

7 Conclusion

We introduce persistent-negative adversarial distillation for transferring behavior from a black-box teacher without access to token probabilities. A sequence-level discriminator learns rewards from prompt-matched teacher and student responses, while GRPO updates the student using fresh on-policy generations. Standard adversarial distillation trains the discriminator only against the current student distribution, so previously observed failure modes can disappear from discriminator training as the policy changes. We address this moving-target problem by retaining historical student responses as persistent negatives. This design preserves broader discriminator supervision while maintaining a fully on-policy student update.

AI use statement

We used generative AI tools to assist with manuscript editing and formatting, and with portions of research-code development and data-analysis workflows. All AI-assisted text was reviewed and revised by the authors. AI-assisted code was inspected and tested, and reported numerical results were checked against the underlying analysis outputs. We did not use generative AI tools as authoritative sources for scientific claims or citations. The language models used as the teacher, evaluators, and objects of study are described separately as part of the methodology. The authors take full responsibility for the final content of this work, including all text, claims, code, and artifacts produced with the aid of generative AI.

Ethics statement

This work studies persistent negatives for black-box model distillation. It involves no human subjects, annotation efforts, or newly collected data. Training uses teacher responses released with public research over LMSYS-Chat-1M, while evaluation uses publicly available benchmarks. Because conversational data and pretrained models may reproduce biases, sensitive content, or harmful behavior, our results do not establish deployment safety.

Reproducibility statement

Section 3.3 specifies the complete training algorithm, including the fixed discriminator-compute intervention. We report the student and judge models, data construction, training schedule, paired-seed protocol, and evaluation procedure in Section 5. Appendix A provides the full optimization, sequence-length, hardware, and pool configurations. Theoretical proofs, seed-level results, and additional discriminator diagnostics appear in Appendices B, C, and D, respectively. We identify all released datasets and pretrained checkpoints through citations so that the experiments can be reconstructed from the documented configurations.

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 *The Twelfth International Conference on Learning Representations*, 2024.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world's first truly open instruction-tuned LLM. Databricks blog, 2023. URL https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm.

Pavlos S. Efraimidis and Paul G. Spirakis. Weighted random sampling with a reservoir. *Information Processing Letters*, 97(5):181–185, 2006.

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. URL https://arxiv.org/abs/2605.07396.

Gemma Team et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. URL https://arxiv.org/abs/2503.19786.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Yuxin Jiang, Chunkit Chan, Mingyang Chen, and Wei Wang. Lion: Adversarial distillation of proprietary large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3134–3154, 2023.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327, 2016.

Ilya Kostrikov, Kumar Krishna Agrawal, Debidatta Dwibedi, Sergey Levine, and Jonathan Tompson. Discriminator-actor-critic: Addressing sample inefficiency and reward bias in adversarial imitation learning. In *International Conference on Learning Representations*, 2019.

Kevin Lu and Thinking Machines Lab. On-policy distillation. Thinking Machines Lab: Connectionism, 2025. URL https://thinkingmachines.ai/blog/on-policy-distillation.

OpenAI. GPT-5 Chat. https://developers.openai.com/api/docs/models/gpt-5-chat-latest, 2025. Model documentation; accessed September 12, 2026.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

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.

Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Josh Susskind, Wenda Wang, and Russ Webb. Learning from simulated and unsupervised images through adversarial training. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2107–2116, 2017.

Aasheesh Singh, Vishal Vaddina, and Dagnachew Birru. ORPO-Distill: Mixed-policy preference optimization for cross-architecture LLM distillation. *arXiv preprint arXiv:2509.25100*, 2025. URL https://arxiv.org/abs/2509.25100. Accepted at the NeurIPS 2025 Workshop on Efficient Reasoning.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 13484–13508, 2023.

Jing Xiong, Hui Shen, Shansan Gong, Yuxin Cheng, Jianghan Shen, Chaofan Tao, Haochen Tan, Haoli Bai, Lifeng Shang, and Ngai Wong. OVD: On-policy verbal distillation. *arXiv preprint arXiv:2601.21968*, 2026. URL https://arxiv.org/abs/2601.21968.

Tianzhu Ye, Li Dong, Zewen Chi, Xun Wu, Shaohan Huang, and Furu Wei. Black-box on-policy distillation of large language models. *arXiv preprint arXiv:2511.10643*, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P. Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. LMSYS-Chat-1M: A large-scale real-world LLM conversation dataset. In *International Conference on Learning Representations*, 2024.

A Implementation and training details

A pool entry is a Bradley–Terry row (x, y_T, y_S). The reference live-pool GAD arm uses a FIFO queue tagged with the step that produced each row. With 256 discriminator rows per rank per step, capacity 4096 covers 16 steps. Each pooled row replaces a fresh row; it does not enlarge the discriminator batch or enter the policy update.

Table 3: Training configuration. Pool rows describe the only intervention.

Component	Configuration
Student / teacher	Qwen2.5-7B-Instruct / sampled GPT-5-Chat responses
Warmup	one epoch of SFT on teacher responses (374 steps at 50%); discriminator jointly initialized; GAD arms resume, SeqKD does not
Adversarial phase	two epochs; 748 steps at 50%, 492 at 33%
Optimization	learning rate 10^{-6}; batch 256; rollout group 8; temperature 0.8
Lengths	maximum prompt 2048 tokens; maximum response 1536
Regularization	KL coefficient 10^{-3}; gradient clip 0.2; discriminator fp32
Reference pool	4096 BT rows; $\rho = 0.5$; uniform sampling
SeqKD	learning rate 5×10^{-6}; batch 256; no discriminator or GRPO
Hardware	one node; 8×H200; tensor parallelism 1

For exploratory live-pool sweeps, stale rows are sampled uniformly, by recency $w_i = \exp[-\lambda(t - s_i)]$ with $\lambda = 10^{-3}$, or by push-time loss $w_i = p_i^\alpha$ with $p_i = \mathrm{softplus}(-m_i)$ and $\alpha \in \{0.5, 1\}$. Weighted draws use A-Res (Efraimidis & Spirakis, 2006): sample $u_i \sim \mathrm{Uniform}(0, 1)$, form u_i^{1/w_i}, and take the largest keys.

B Proofs of the Theoretical Results

We provide complete proofs of the three results in Section 4. All expectations over learned score functions are taken over the randomness of the training procedure.

B.1 Proof of Theorem 4.1

Proof. Fix a prompt x and suppress it from the notation. Let the finite response space be $\mathcal{Y} = \{1, \dots, m\}$, and write $\tau_i = \tau(i \mid x)$, $q_i = q(i \mid x)$, and $r_i = r(x, i)$. The population loss is

$$\mathcal{L}_D(r; q) = \sum_{i,j} \tau_i q_j \,\mathrm{softplus}(r_j - r_i).$$

Terms with $i = j$ are constant. For an unordered pair $i \neq j$, let $d = r_i - r_j$. Its contribution is

$$g_{ij}(d) = \tau_i q_j \,\mathrm{softplus}(-d) + \tau_j q_i \,\mathrm{softplus}(d).$$

Since all probabilities are positive, g_{ij} is strictly convex. Its derivative vanishes exactly when

$$d = \log \frac{\tau_i q_j}{\tau_j q_i} = \log \frac{\tau_i}{q_i} - \log \frac{\tau_j}{q_j}.$$

These pairwise conditions are simultaneously satisfied by

$$r_i = \log \frac{\tau_i}{q_i} + c,$$

where c is independent of the response. Because the complete loss is convex and this score minimizes every pairwise contribution, it is a global minimizer. Positivity of τ_i and q_i makes the comparison graph connected, so all minimizing score differences are fixed. The only remaining freedom is the additive constant c. Restoring the prompt dependence gives

$$r_q^\star(x, y) = \log \frac{\tau(y \mid x)}{q(y \mid x)} + c_q(x).$$

□

B.2 Proof of Proposition 4.2

Proof. By linearity of expectation in the negative distribution,

$$\mathcal{L}_t^\rho(r) = (1 - \rho)\mathcal{L}_t^{\mathrm{fresh}}(r) + \rho\mathcal{L}_t^{\mathrm{hist}}(r).$$

The softplus loss is nonnegative. Therefore,

$$\mathcal{L}_t^{\rho}(r) \geq \rho \mathcal{L}_t^{\mathrm{hist}}(r).$$

The softplus function is convex, so both component losses are convex in the score vector. Differentiating the mixture gives

$$\nabla^2 \mathcal{L}_t^{\rho}(r) = (1-\rho)\nabla^2 \mathcal{L}_t^{\mathrm{fresh}}(r) + \rho \nabla^2 \mathcal{L}_t^{\mathrm{hist}}(r).$$

Under the curvature assumptions in the theorem, it follows that

$$\nabla^2 \mathcal{L}_t^{\rho}(r) \succeq \big[(1-\rho)\mu_{\mathrm{fresh}} + \rho \mu_{\mathrm{hist}}\big] I_{\mathcal{S}} \succeq \rho \mu_{\mathrm{hist}} I_{\mathcal{S}}.$$

For completeness, consider the stated connectedness condition for the historical comparison graph. For any direction $v \in \mathcal{S}$,

$$v^\top \nabla^2 \mathcal{L}_t^{\mathrm{hist}}(r) v = \sum_{i,j} \tau_i h_{t,j}\, \ell''(r_j - r_i)(v_j - v_i)^2,$$

where $\ell(u) = \mathrm{softplus}(u)$ and $\ell''(u) = \sigma(u)\sigma(-u) > 0$.

On a bounded score set, $\ell''(r_j - r_i)$ is uniformly bounded away from zero. If the historical comparison graph is connected, the quadratic form can vanish only when $v_i = v_j$ for every connected pair. Thus v must be constant. The centered constraint $v \in \mathcal{S}$ then implies $v = 0$. Hence the historical Hessian is uniformly positive definite on $\mathcal{S}$, establishing $\mu_{\mathrm{hist}} > 0$ and completing the proof. □

B.3 Proof of Proposition 4.3

Proof. Define the promptwise-centering operator

$$(C_t f)(x,y) = f(x,y) - \mathbb{E}_{y' \sim \pi_{\theta_t}(\cdot|x)} f(x,y').$$

Then $\|f\|_{t,c}^2 = \|C_t f\|_t^2$, where $\|\cdot\|_t$ is the ordinary L_2 norm under $x \sim \mathcal{D}$ and $y \sim \pi_{\theta_t}(\cdot \mid x)$.

For either estimator $\widehat{r}_t^a$, write

$$\widehat{r}_t^a - r_t^\star = \big(\widehat{r}_t^a - \mathbb{E}\widehat{r}_t^a\big) + \big(\mathbb{E}\widehat{r}_t^a - r_t^\star\big).$$

The first term has zero expectation. Consequently, the expected cross term vanishes after applying the linear centering operator C_t, giving the bias–variance decomposition

$$\mathbb{E}\|\widehat{r}_t^a - r_t^\star\|_{t,c}^2 = V_a + \|\mathbb{E}\widehat{r}_t^a - r_t^\star\|_{t,c}^2.$$

The assumed unbiasedness of the fresh estimator therefore gives

$$\mathrm{MSE}(\widehat{r}_t^{\mathrm{fresh}}, r_t^\star) = V_{\mathrm{fresh}},$$

whereas the pooled estimator satisfies

$$\mathrm{MSE}(\widehat{r}_t^{\mathrm{pool}}, r_t^\star) = V_{\mathrm{pool}} + B_{\mathrm{pool}}^2.$$

Using $V_{\mathrm{pool}} \leq V_{\mathrm{fresh}} - \Delta_t$, we obtain

$$\mathrm{MSE}(\widehat{r}_t^{\mathrm{fresh}}, r_t^\star) - \mathrm{MSE}(\widehat{r}_t^{\mathrm{pool}}, r_t^\star) \geq \Delta_t - B_{\mathrm{pool}}^2.$$

Finally, $B_{\mathrm{pool}}^2 < \Delta_t$ makes the right-hand side strictly positive. Hence the pooled estimator has lower reward MSE under the stated bias–variance condition. □

C Seed-Level Robustness

We evaluate GAD and persistent-negative training using three paired random seeds for each student family. Within each pair, both methods start from the same seed-specific warmup checkpoint and use matched training configurations, differing only in the discriminator negative distribution.

Figure 3 reports the persistent-negative-minus-GAD difference for each seed. Each effect is averaged equally across the four judged-chat datasets. The individual points expose training-run variability; the mean and standard deviation summarize the three paired effects. Because only three seeds are available, the standard deviation is descriptive and should not be interpreted as a precise uncertainty interval.

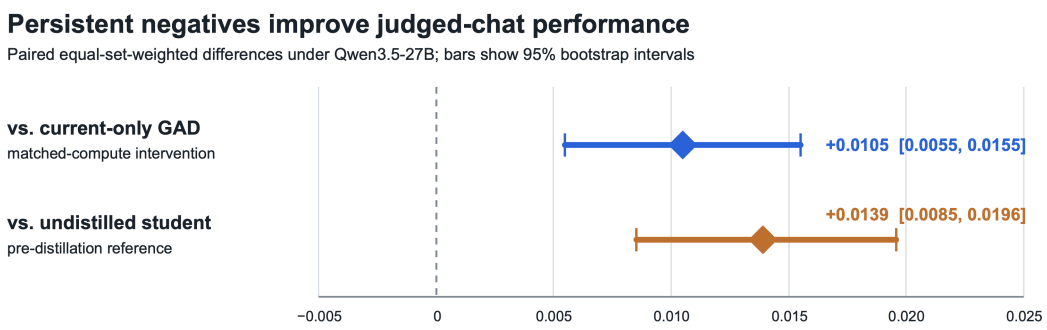


Figure 3: **Performance gains across paired training seeds.** Each point shows the equal-set-weighted improvement of persistent-negative training over GAD for one paired seed, measured in percentage points. Large markers denote the mean across three seeds, and error bars show one standard deviation. Positive values favor persistent-negative training.

Table 4: **Fresh-policy discriminator accuracy in the initial primary run.** Standard deviation is computed across adversarial updates, and "dips" counts pre-update evaluations below chance.

Method	Mean	Std.	Minimum	Dips < 0.5
GAD	0.894	0.087	0.414	6
Persistent-negative (ours)	**0.903**	**0.068**	**0.461**	**1**

D Additional Discriminator Diagnostics

D.1 Fresh-policy discriminator accuracy

We evaluate the discriminator immediately before each update using fresh current-policy comparisons. The prompt–teacher slice is fixed across methods, but each student generates its own negative responses. Consequently, this diagnostic measures the behavior of the coupled student–discriminator system rather than holding the student distribution fixed.

Table 4 reports results from the initial primary run across 748 adversarial updates. Persistent-negative training has slightly higher mean accuracy, lower temporal variability, a higher minimum, and fewer below-chance evaluations than GAD. Because consecutive measurements come from the same training trajectory, these statistics are descriptive rather than independent observations.

These measurements are consistent with a smoother discriminator trajectory under persistent-negative training. They do not, however, isolate changes in the discriminator from changes in the evolving student policy.

D.2 Common-probe reward geometry

Fresh-policy accuracy does not reveal whether two discriminators assign different rewards to identical responses. We therefore construct a common probe by sampling eight responses for each of 128 held-out prompts from the shared warmup policy. The resulting response groups are fixed and scored by the final GAD and persistent-negative discriminators.

Across groups, the discriminators have a mean within-group Spearman rank correlation of 0.803 and a mean correlation of 0.846 between their standardized reward profiles. Thus, the learned reward functions are substantially aligned, but they are not related solely by transformations removed by promptwise GRPO normalization. Differences in response ordering and relative score spacing therefore remain visible to the policy update.

This diagnostic uses one final discriminator pair and should be interpreted as evidence that the pool changes the policy-facing reward geometry, not as evidence that this change mediates the final performance improvement.