# DISTILL WHERE IT MATTERS: SELECTIVE ON-POLICY SELF-DISTILLATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

On-policy self-distillation (OPSD) provides dense supervision without a stronger external teacher by conditioning a frozen teacher initialized from the same model on the reference solution. However, standard OPSD distills every example equally, overlooking that privileged information may benefit the teacher differently across problems. We introduce Information-Gain Weighted OPSD (IG-OPSD), which measures the teacher's advantage over the student through their reference-likelihood gap and assigns greater weight to examples where this advantage is larger. Experiments with Qwen3 models show that IG-OPSD consistently outperforms uniform OPSD across four mathematical reasoning benchmarks at both 8B and 14B scales. On the two most difficult benchmarks under uniform OPSD, the 14B model improves Avg@12 by 4.2 points on AIME25 and 5.6 points on HMMT25. Ablation and diagnostic analyses show that the gains depend on matching weights to the correct examples: IG primarily captures the advantage provided by privileged context and directs learning toward harder problems. These results establish sample selection as a key design axis for on-policy self-distillation: the student should learn more from the teacher precisely when privileged information makes the teacher more informative.

## 1 INTRODUCTION

OPSD constructs supervision without a stronger external teacher (Zhao et al., 2026). The student samples an on-policy rollout from the problem alone, while a frozen teacher initialized from the same model evaluates its prefixes with privileged access to the reference solution. Distilling the student toward this privileged distribution provides dense token-level feedback without a separately trained reward model or a larger teacher.

Standard OPSD, however, treats the teacher's signal as equally useful on every example. Reference access can substantially improve the teacher's view when the student lacks solution-relevant information, but may add little when the privileged and unprivileged views are already similar. Uniform weighting spends the same optimization budget in both cases. This exposes a central question for on-policy self-distillation: *when is the privileged teacher actually worth following?*

We answer this question with Information-Gain Weighted OPSD (IG-OPSD). For each example, IG-OPSD measures the teacher's advantage through the reference log-likelihood gap between the privileged teacher and the unprivileged student. It standardizes this score against a recent reference window and converts positive relative advantage into a sample weight. IG-OPSD changes only how much each example contributes to the loss, leaving rollout generation and the token-level distillation target unchanged. In this view, on-policy distillation has three complementary axes—what distribution to imitate, which tokens to emphasize, and which examples deserve the teacher's influence—and IG-OPSD addresses the third.

Our contributions are threefold. First, we formulate privileged-teacher advantage as a sample-selection signal and instantiate it in a simple weighting rule for OPSD. Second, with 12 sampled responses per problem, IG-OPSD consistently improves Avg@12 over uniform OPSD on AIME24, AIME25, AMC23, and HMMT25 at both 8B and 14B scales; at 14B, it raises Avg@12 by 4.2 percentage points on AIME25 and 5.6 points on HMMT25. At both scales, randomly reassigning the same weights to different examples reduces performance. The gains therefore come from emphasizing the right examples, showing that sample selection complements target design and token weighting. Third, we provide a detailed account of why the method works: decomposition shows that IG primarily tracks the advantage supplied by privileged context rather than parameter drift, weighting ablations identify



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.

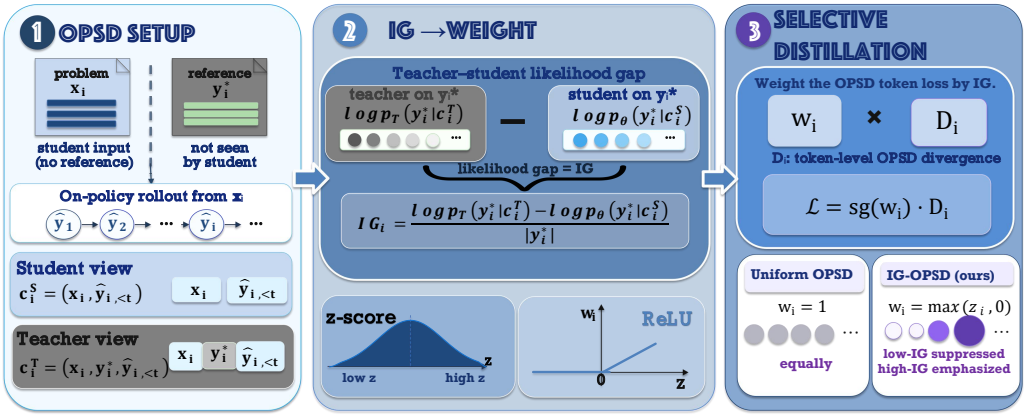


Figure 1: IG-OPSD adds sample selection to OPSD. The student samples from the problem alone, while the fixed teacher receives the reference solution. A teacher-forced reference-likelihood gap measures the teacher's advantage; a stop-gradient z-score and ReLU produce non-negative weights, which are rescaled to global-batch mean one.

sample selection as the main source of improvement, and difficulty-stratified results show that the gains concentrate on problems that uniform OPSD rarely solves.

## 2 Related Work

**On-policy distillation with privileged teachers.** Knowledge distillation traditionally transfers predictive information from a stronger teacher to a weaker student (Hinton et al., 2015). For generative models, MiniLLM trains on student-generated sequences with a reverse-KL objective (Gu et al., 2024), while GKD uses on-policy generations to reduce the mismatch between distillation and deployment (Agarwal et al., 2024). OPSD removes the need for a stronger external teacher by conditioning a frozen teacher initialized from the same model on a reference solution while supervising student rollouts generated from the problem alone (Zhao et al., 2026). IG-OPSD retains this teacher, rollout, and token-level target construction, but challenges the assumption that every example should receive the same amount of teacher influence.

**Selective on-policy self-distillation.** Existing extensions of OPSD modify different parts of the distillation process. PW-OPSD changes token-level weights according to position-dependent teacher reliability (Liu et al., 2026), whereas Purified-OPSD corrects the target distribution using a PMI-based criterion (Shen et al., 2026). Other recent studies examine reasoning branches and the effect of self-distillation on output diversity (Kaur et al., 2026; Nicolicioiu et al., 2026). These methods determine what distribution to imitate or which tokens to emphasize. IG-OPSD is complementary: it assigns a single weight to each example according to the advantage provided by privileged context, determining when the teacher should exert greater influence.

**Information-guided selection and reasoning feedback.** Information-theoretic acquisition methods select informative data using expected information gain (Kirsch & Gal, 2022). Our IG score is different: it is a realized reference-likelihood gain that compares a privileged teacher with an unprivileged student. It is therefore an operational measure of teacher advantage rather than a Bayesian information-gain objective or a causal estimate of context value. In reasoning post-training, GRPO learns from group-relative outcome rewards (Shao et al., 2024), process reward models supervise intermediate reasoning steps (Lightman et al., 2023), and verifiable-reward methods optimize against outcome feedback (DeepSeek-AI, 2025). SDPO instead conditions the current model on environmental feedback and distills the resulting predictions into the policy to obtain dense credit assignment (Hübotter et al., 2026). IG-OPSD requires none of these additional reward or feedback signals: it uses the reference solution already available in OPSD to decide how strongly each example should influence learning.

# 3 METHOD

## 3.1 OPSD BACKGROUND

For a problem-reference pair $(x_i, y_i^\star)$, the student samples a trajectory $\hat{y}_i \sim p_\theta(\cdot \mid x_i)$. OPSD then compares the next-token distribution of a student context $c_i^S = (x_i, \hat{y}_{i,<t})$ with a teacher context $c_i^T = (x_i, y_i^\star, \hat{y}_{i,<t})$. With a generic divergence $D$, the unweighted objective is

$$\mathcal{L}_{\text{OPSD}}(\theta) = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \frac{1}{T_i} \sum_{t=1}^{T_i} D\big(p_T(\cdot \mid c_{i,t}^T) \,\|\, p_\theta(\cdot \mid c_{i,t}^S)\big)\,. \tag{1}$$

In our experiments, $\beta = 0$, so $D$ is the forward-KL endpoint $D_{\text{KL}}(p_T \,\|\, p_\theta)$, evaluated with the support and clipping settings reported in Appendix L.

The OPSD divergence supplies dense token-level supervision, while the reference solution determines how strongly each sample contributes.

## 3.2 MEASURING PRIVILEGED TEACHER ADVANTAGE

For each sampled problem, IG-OPSD measures the reference advantage of the fixed privileged teacher over the current unprivileged student:

$$\text{IG}_i = \frac{1}{|y_i^\star|} \left[\log p_T(y_i^\star \mid x_i, y_i^\star, \hat{y}_i) - \log p_\theta(y_i^\star \mid x_i, \hat{y}_i)\right]. \tag{2}$$

The implementation computes this gap with teacher forcing by appending the reference-solution tokens after the already-constructed teacher and student sequences. Length normalization makes scores from different reference lengths more comparable, and the score is detached from the computation graph. We retain the name information gain because the quantity measures a realized reference log-likelihood gain, but interpret it operationally as *privileged teacher advantage*: it combines the teacher's privileged context with the fixed-teacher/current-student parameter difference.

An absolute positive gap is not by itself a stable threshold because its scale changes across models and training steps. We therefore compare each score with the mean and standard deviation of a reference window formed from the gathered global batch and recent values:

$$z_i = \frac{\text{IG}_i - \mu_{\mathcal{R}}}{\sigma_{\mathcal{R}} + \epsilon}, \quad \tilde{w}_i = \max(z_i, 0), \quad w_i = \begin{cases} \tilde{w}_i \dfrac{|\mathcal{B}_g|}{\sum_{j \in \mathcal{B}_g} \tilde{w}_j}, & \sum_j \tilde{w}_j > 0, \\ 1, & \text{otherwise.} \end{cases} \tag{3}$$

The final loss is

$$\mathcal{L}_{\text{IG-OPSD}}(\theta) = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} w_i \cdot \frac{1}{T_i} \sum_{t=1}^{T_i} D\big(p_T(\cdot \mid c_{i,t}^T) \,\|\, p_\theta(\cdot \mid c_{i,t}^S)\big)\,. \tag{4}$$

The ReLU suppresses examples whose teacher advantage is at or below the recent reference mean. Active weights are rescaled to global-batch mean one, preserving average loss scale; if all weights are zero, the step falls back to uniform OPSD.

Algorithm 1 summarizes the complete training step.

The method adds two teacher-forced reference-likelihood evaluations per sample. A shuffled-weight control tests whether performance depends on attaching each score to the problem that produced it.

# 4 EXPERIMENTAL SETUP

We compare Qwen3-1.7B, 8B, and 14B at checkpoints 25, 50, 75, and 100. The untrained model is *Base*; unweighted fixed-teacher OPSD is the matched baseline. Evaluations use non-thinking inference (temperature 1.0, top-p 0.8, max 8192 tokens). The benchmark suite contains AIME24, AIME25, AMC23, and HMMT25; Avg@12 averages correctness over twelve completions per problem.

**Algorithm 1** Information-Gain Weighted On-Policy Self-Distillation (IG-OPSD)

**Require:** Dataset $\mathcal{D}$, student $p_\theta$, fixed privileged teacher $p_T$, history capacity $M$, stabilizer $\epsilon$

Initialize recent-IG history $\mathcal{H} \leftarrow \varnothing$
**while** training is not complete **do**
  Sample a minibatch $\mathcal{B} = \{(x_i, y_i^\star)\}$
  **for all** $i \in \mathcal{B}$ **do**
    Sample an on-policy rollout $\hat{y}_i \sim p_\theta(\cdot \mid x_i)$
    Form $c_{i,t}^S = (x_i, \hat{y}_{i,<t})$ and $c_{i,t}^T = (x_i, y_i^\star, \hat{y}_{i,<t})$
    Using teacher forcing and no gradient, compute
    $\mathrm{IG}_i \leftarrow \dfrac{1}{|y_i^\star|}\left[\log p_T(y_i^\star \mid x_i, y_i^\star, \hat{y}_i) - \log p_\theta(y_i^\star \mid x_i, \hat{y}_i)\right]$
  **end for**
  Gather current IG values across workers into $\mathrm{IG}_{\mathcal{B}_g}$
  $\mathcal{R} \leftarrow \mathcal{H} \cup \mathrm{IG}_{\mathcal{B}_g}$; compute $\mu_{\mathcal{R}}$ and $\sigma_{\mathcal{R}}$
  **for all** $i \in \mathcal{B}_g$ **do**
    $z_i \leftarrow (\mathrm{IG}_i - \mu_{\mathcal{R}})/(\sigma_{\mathcal{R}} + \epsilon)$
    $w_i \leftarrow \mathrm{stopgrad}(\max(z_i, 0))$
  **end for**
  **if** $w_i = 0$ for every $i \in \mathcal{B}_g$ **then**
    Set $w_i \leftarrow 1$ for every $i \in \mathcal{B}_g$ ▷ uniform OPSD fallback
  **end if**
  $w_i \leftarrow w_i |\mathcal{B}_g| / \sum_{j \in \mathcal{B}_g} w_j$ for every $i \in \mathcal{B}_g$ ▷ mean-one rescaling
  **for all** $i \in \mathcal{B}$ **do**
    $D_i \leftarrow \dfrac{1}{T_i} \sum_{t=1}^{T_i} D\Big(p_T(\cdot \mid c_{i,t}^T) \,\|\, p_\theta(\cdot \mid c_{i,t}^S)\Big)$
  **end for**
  $\mathcal{L}_{\text{IG-OPSD}}(\theta) \leftarrow \dfrac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} w_i D_i$
  Update $\theta$ using $\nabla_\theta \mathcal{L}_{\text{IG-OPSD}}$
  Update $\mathcal{H}$ with $\mathrm{IG}_{\mathcal{B}_g}$ and retain the most recent $M$ values
**end while**

# 5 RESULTS

## 5.1 MAIN ACCURACY RESULTS

Table 1 reports Base and trained-model Avg@12 and Pass@12 on the four competition benchmarks. Both trained methods improve their four-dataset Avg@12 mean over Base at every model size. At 8B and 14B, IG-OPSD improves over OPSD on all four Avg@12 datasets, with the largest 14B gains on AIME25 and HMMT25.

Table 1: Avg@12 and Pass@12 results on AIME24, AIME25, AMC23, and HMMT25. Base is the untrained model; OPSD and IG-OPSD are evaluated at checkpoint 100. Avg. is the unweighted average across datasets. Scores are percentages and means are computed before rounding.

| Model | Method | Metric | AIME24 | AIME25 | AMC23 | HMMT25 | Avg. |
|---|---|---|---|---|---|---|---|
| Qwen3-1.7B | Base | Avg@12 | 13.1 | 10.0 | 45.0 | 5.3 | 18.3 |
| | | Pass@12 | 33.3 | 26.7 | 85.0 | 16.7 | 40.4 |
| | OPSD | Avg@12 | **16.1** | **10.8** | 46.9 | 6.7 | 20.1 |
| | | Pass@12 | 36.7 | 23.3 | **85.0** | 16.7 | 40.4 |
| | **IG-OPSD** | Avg@12 | 14.4 | 10.6 | **49.0** | **7.8** | **20.5** |
| | | Pass@12 | **40.0** | **30.0** | 82.5 | **23.3** | **44.0** |
| Qwen3-8B | Base | Avg@12 | 26.4 | 21.4 | 71.0 | 10.3 | 32.3 |
| | | Pass@12 | 50.0 | 46.7 | 92.5 | 23.3 | 53.1 |
| | OPSD | Avg@12 | 39.2 | 27.5 | 76.9 | 14.7 | 39.6 |
| | | Pass@12 | **66.7** | **50.0** | **92.5** | 33.3 | 60.6 |
| | **IG-OPSD** | Avg@12 | **41.4** | **28.6** | **78.5** | **17.5** | **41.5** |
| | | Pass@12 | 63.3 | **50.0** | **92.5** | **40.0** | **61.5** |
| Qwen3-14B | Base | Avg@12 | 31.1 | 25.6 | 74.2 | 11.7 | 35.6 |
| | | Pass@12 | 63.3 | 40.0 | 95.0 | 26.7 | 56.3 |
| | OPSD | Avg@12 | 52.2 | 39.7 | 86.2 | 20.0 | 49.5 |
| | | Pass@12 | **80.0** | 63.3 | **97.5** | 43.3 | 71.0 |
| | **IG-OPSD** | Avg@12 | **52.5** | **43.9** | **89.4** | **25.6** | **52.8** |
| | | Pass@12 | **80.0** | 63.3 | 95.0 | **50.0** | **72.1** |

At 8B, IG-OPSD improves the four-dataset Avg@12 mean from 39.6% to 41.5%; at 14B, it improves the mean from 49.5% to 52.8%. The equal-weighted 14B improvement is +3.28 points, with a paired problem-level 95% bootstrap interval of [+1.27, +5.43]. The 14B gains on AIME25, AMC23, and HMMT25 also have intervals that exclude zero; Appendix F reports all per-dataset intervals.

Figure 2 summarizes the Avg@12 gains by dataset and model size. All four datasets improve at 8B and 14B; the largest gain is +5.6 points on HMMT25 at 14B.

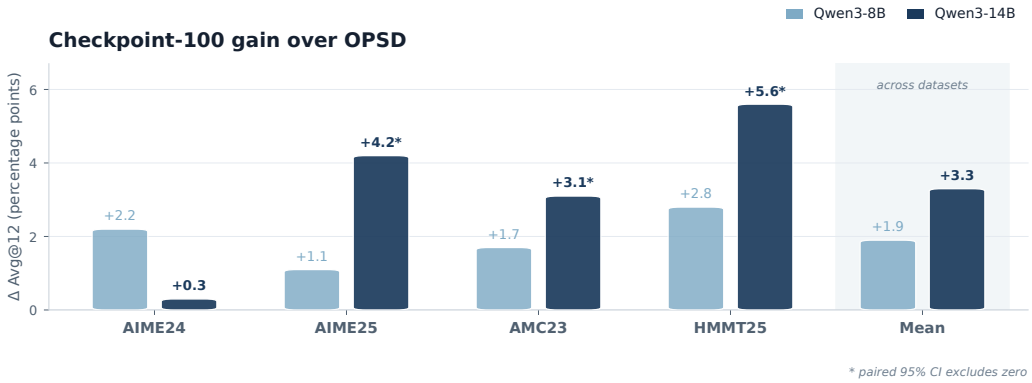


Figure 2: Checkpoint-100 Avg@12 gains on the four competition benchmarks and their unweighted mean, IG-OPSD minus OPSD.

At 1.7B, the Avg@12 mean changes from 20.1% to 20.5%. The contrast across model scales suggests that IG weighting becomes effective only when the student can exploit the privileged-teacher advantage it identifies.

### 5.2 Training Dynamics

At 8B, IG-OPSD separates from OPSD after checkpoint 50 and finishes higher on all four benchmarks. At 14B, the largest late-stage gaps occur on AIME25 and HMMT25 (Figure 3).

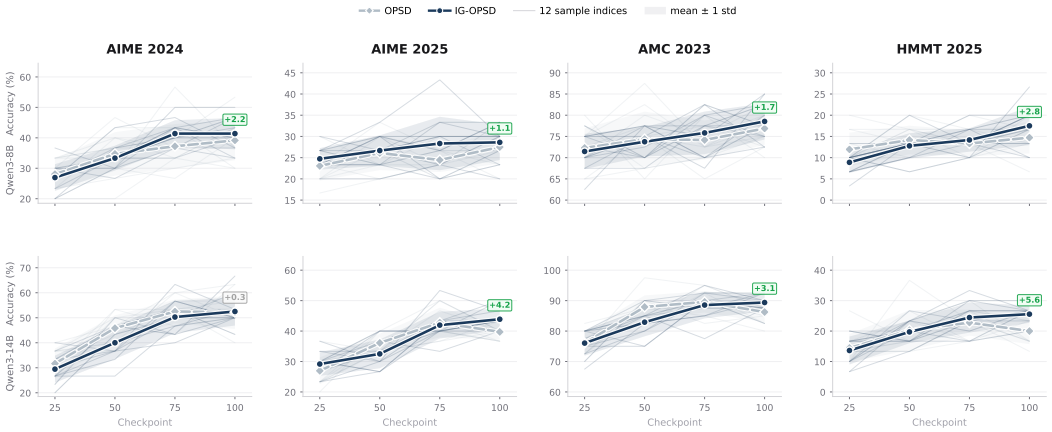


Figure 3: Per-dataset checkpoint curves through step 100 on AIME 2024, AIME 2025, AMC 2023, and HMMT 2025; the top row is Qwen3-8B and the bottom row is Qwen3-14B. Faint lines show twelve sample-index scores per checkpoint, while ribbons show mean ± one standard deviation. Scores are Avg@12 percentages.

### 5.3 Comparison with Alternative Training Methods

Table 2 compares IG-OPSD with OPSD, PW-OPSD (Liu et al., 2026), Purified-OPSD (Shen et al., 2026), SFT, GRPO (Shao et al., 2024), and SDPO (Hübotter et al., 2026). GRPO and SDPO are reinforcement-learning methods trained for 200 steps; all other trained methods use 100 steps.

Table 2: Qwen3-8B and Qwen3-14B Avg@12 comparison (%). Base is untrained; Mean is the unweighted average across the four benchmarks.

| Method | AIME24 | AIME25 | AMC23 | HMMT25 | Avg. |
|---|---|---|---|---|---|
| **Qwen3-8B** | | | | | |
| Base | 26.4 | 21.4 | 71.0 | 10.3 | 32.3 |
| SFT | 25.3 | 18.1 | 62.9 | 12.2 | 29.6 |
| GRPO | 30.6 | 21.9 | 75.2 | 13.1 | 35.2 |
| SDPO | 31.7 | 24.4 | 71.5 | 13.6 | 35.3 |
| OPSD | 39.2 | 27.5 | 76.9 | 14.7 | 39.6 |
| Purified-OPSD | 27.8 | 24.7 | 72.3 | 11.4 | 34.1 |
| PW-OPSD | 33.6 | **29.4** | 74.6 | 13.9 | 37.9 |
| **IG-OPSD** | **41.4** | 28.6 | **78.5** | **17.5** | **41.5** |
| **Qwen3-14B** | | | | | |
| Base | 31.1 | 25.6 | 74.2 | 11.7 | 35.6 |
| SFT | 31.4 | 23.3 | 72.7 | 13.1 | 35.1 |
| GRPO | 49.4 | 36.4 | 82.5 | 17.0 | 46.3 |
| SDPO | 50.1 | 39.1 | 85.2 | 16.4 | 47.7 |
| OPSD | 52.2 | 39.7 | 86.2 | 20.0 | 49.5 |
| Purified-OPSD | 29.2 | 26.1 | 74.6 | 14.4 | 36.1 |
| PW-OPSD | **53.6** | 40.3 | 89.2 | 23.1 | 51.5 |
| **IG-OPSD** | **52.5** | **43.9** | **89.4** | **25.6** | **52.8** |

IG-OPSD has the highest four-dataset mean at both scales: 41.5% at 8B and 52.8% at 14B. It exceeds PW-OPSD by 3.6 and 1.3 mean points, respectively, although PW-OPSD leads one AIME benchmark at each scale. Purified-OPSD remains near Base and below uniform OPSD, indicating that target correction alone does not recover the gains from selecting examples by teacher advantage. OPSD, PW-OPSD, and IG-OPSD all exceed the SFT and RL baselines in mean Avg@12. Figure 4 shows the per-dataset scores.

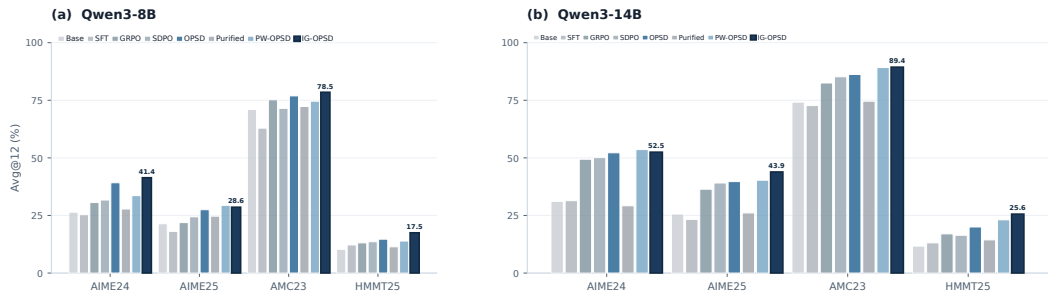


Figure 4: Per-dataset Avg@12 at Qwen3-8B and 14B for the methods in Table 2. IG-OPSD values are labeled above each bar.

### 5.4 Weighting ablations

Table 3 compares the Qwen3-14B weighting rules on AIME24, AIME25, and HMMT25 at checkpoint 100.

Table 3: Qwen3-14B checkpoint-100 Avg@12 ablation results (%). Mean is the unweighted average of AIME24, AIME25, and HMMT25.

| Method | AIME24 | AIME25 | HMMT25 | Avg. |
|---|---|---|---|---|
| OPSD (uniform) | 52.2 | 39.7 | 20.0 | 37.3 |
| Binary IG gate | **53.6** | 43.3 | 22.8 | 39.9 |
| Shuffled IG weights | 49.4 | 43.3 | 20.3 | 37.7 |
| Current-batch z-score | 48.9 | 41.1 | 23.6 | 37.9 |
| Positive raw IG | 50.0 | 41.4 | 20.8 | 37.4 |
| **IG-OPSD** | 52.5 | **43.9** | **25.6** | **40.6** |

IG-OPSD reaches the highest three-dataset mean (40.6%), followed by the binary gate (39.9%). Shuffling the same weights across problems reduces the mean to 37.7%, close to uniform OPSD at 37.3%, showing that sample identity matters. Current-batch normalization (37.9%) and positive raw IG (37.4%) also trail the reference-window rule.

Table 4 aggregates the two central comparisons. IG-OPSD improves over uniform OPSD at 14B, and over shuffled-weight controls at both 8B and 14B; all paired intervals exclude zero.

Table 4: Aggregate checkpoint-100 Avg@12 contrasts. $\Delta$ is IG-OPSD minus the comparator in percentage points. The four-set comparison averages dataset means equally; shuffled controls pool three equally sized 30-problem sets. Intervals are paired problem-level bootstrap 95% CIs and do not include training-seed variation.

| Model | Comparator | Sets | $\Delta$ Avg@12 | Paired 95% CI |
|---|---|---|---|---|
| Qwen3-14B | OPSD | 4 | +3.28 | [+1.27, +5.43] |
| Qwen3-8B | Shuffled IG weights | 3 | +3.98 | [+1.02, +7.22] |
| Qwen3-14B | Shuffled IG weights | 3 | +2.96 | [+0.56, +5.56] |

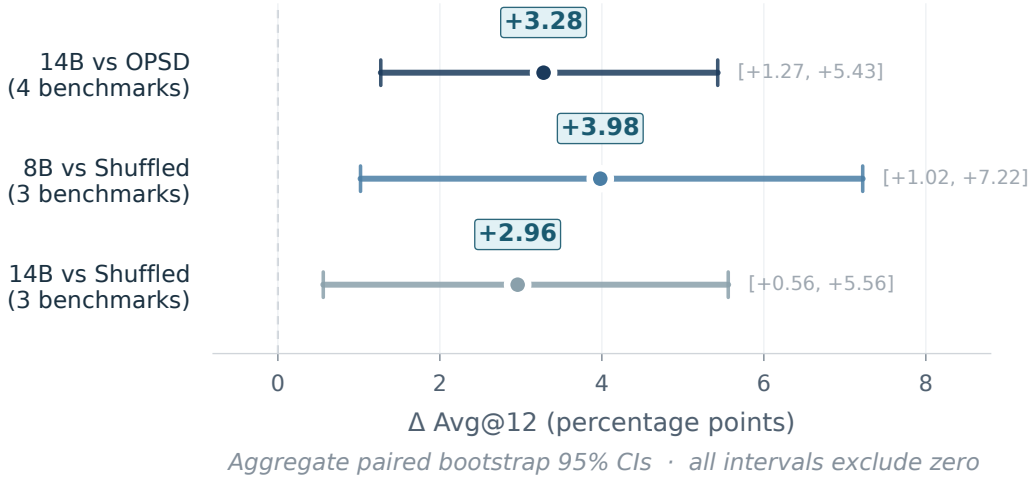


Figure 5: Aggregate contrasts from Table 4. Diamonds show point estimates and bars show paired bootstrap 95% CIs.

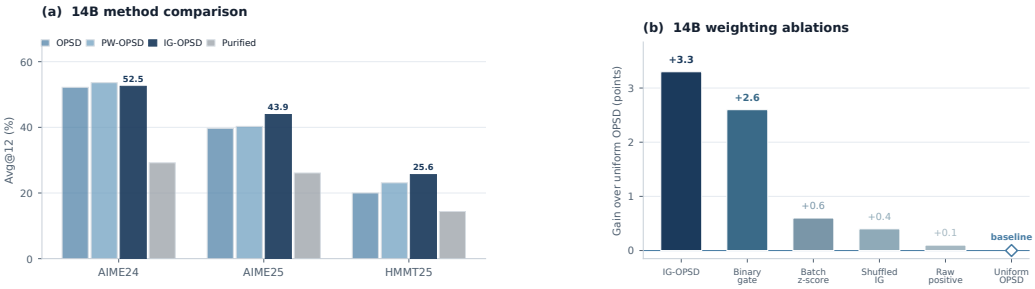


Figure 6: Qwen3-14B checkpoint-100 controls. (a) Distillation-method comparison on AIME24, AIME25, and HMMT25. (b) Three-dataset mean gains of the weighting ablations over uniform OPSD.

### 5.5 Information-gain decomposition

We decompose teacher-forced IG into *context advantage*, measured between privileged and unprivileged views of the fixed teacher, and *parameter drift*, measured between the unprivileged fixed teacher and the current student.

Context advantage accounts for 70% of mean total IG at 8B and 85% at 14B at checkpoint 100. It also closely matches the total-IG sample ranking (Spearman $\rho > 0.85$ at 8B and $\rho > 0.94$ at 14B), with active-set Jaccard similarity of at least 0.74. Parameter drift increases over training, while the context component remains nearly constant. Appendix D reports the checkpoint-wise ranking and active-set statistics.

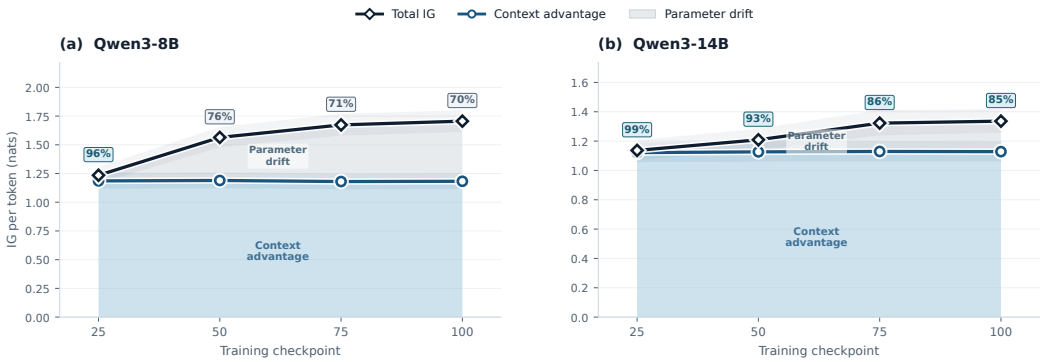


Figure 7: Offline IG decomposition into context advantage and parameter drift on the pooled competition evaluation sets. Shaded areas stack the two components; percentage labels show the context advantage share of the mean total IG. Context advantage is nearly constant across checkpoints; the total IG increase is driven almost entirely by parameter drift, yet context advantage dominates the sample ranking at every checkpoint.

### 5.6 Where do the gains occur?

We stratify the two most difficult 14B benchmarks using uniform OPSD's per-problem Avg@12: easy problems have Avg@12 at least $2/3$, medium problems at least $1/3$, and hard problems below $1/3$. On AIME25, IG-OPSD improves hard-problem Avg@12 by 5.7 points, compared with 0.9 points on easy problems; the hard subset contributes 73% of the dataset-level gain. On HMMT25, the corresponding changes are +5.4 points on hard problems and 0.0 on easy problems, with 75% of the gain contributed by the hard subset. Thus, the aggregate improvement is concentrated on problems that uniform OPSD solves infrequently rather than on already-saturated items. This stratification is descriptive: its bins are defined post hoc using the comparator's evaluation accuracy, not an independent difficulty annotation.

Table 5: Qwen3-14B checkpoint-100 Avg@12 by OPSD difficulty stratum (%). Share is the stratum's contribution to the dataset-level IG-OPSD gain; strata contain different numbers of problems.

| Dataset | Stratum | $n$ | OPSD | IG-OPSD | $\Delta$ | Share |
|---|---|---|---|---|---|---|
| AIME25 | Easy | 9 | 97.2 | 98.1 | +0.9 | 7% |
| | Medium | 5 | 51.7 | 56.7 | +5.0 | 20% |
| | Hard | 16 | 3.6 | 9.4 | +5.7 | 73% |
| HMMT25 | Easy | 5 | 88.3 | 88.3 | +0.0 | 0% |
| | Medium | 2 | 41.7 | 62.5 | +20.8 | 25% |
| | Hard | 23 | 3.3 | 8.7 | +5.4 | 75% |

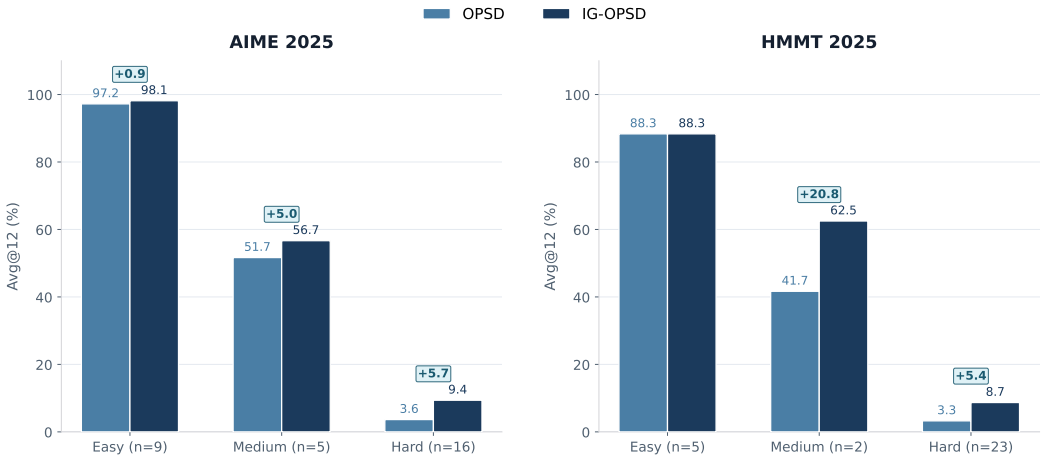


Figure 8: Qwen3-14B checkpoint-100 Avg@12 by the OPSD difficulty strata in Table 5.

Together with the decomposition in Section 5.5, these findings provide a consistent explanation for the larger gains on AIME25 and HMMT25. IG primarily reflects the advantage introduced by privileged context, while the improvements on these benchmarks concentrate on problems that uniform OPSD rarely solves. This suggests that IG-OPSD is most useful when the reference solution supplies information missing from the student's unprivileged view.

## 6 Discussion

Taken together, the analyses support a simple account of why IG-OPSD works. Privileged context largely determines which examples receive high IG scores, and the ablations show that selecting the right examples matters more than the precise weight magnitude. The resulting gains concentrate on problems that uniform OPSD rarely solves. Thus, IG-OPSD improves OPSD by allocating teacher influence where it is most useful, while leaving the rollout and token-level target unchanged.

## 7 Conclusion

IG-OPSD turns the teacher's reference advantage into a sample-allocation rule, allowing the student to learn more when privileged information makes the teacher more informative. With 12 sampled responses per problem, it consistently improves Avg@12 over uniform OPSD on AIME24, AIME25, AMC23, and HMMT25 at both 8B and 14B scales; at 14B, it raises Avg@12 by 4.2 percentage points on AIME25 and 5.6 points on HMMT25. Correctly paired weights outperform shuffled controls at both scales, providing direct evidence that example-specific selection complements target design and token weighting. Our analyses further show that IG primarily captures privileged-context advantage and focuses learning on problems that uniform OPSD struggles to solve. Effective self-distillation is therefore not about following the teacher everywhere, but about recognizing when the teacher has something worth teaching.

### Ethics statement

This work does not involve human subjects, personal data, or other ethical concerns requiring special review. All models, codebases, and datasets were used in accordance with the licenses of their respective open-source projects.

### Reproducibility statement

Section 4 summarizes the evaluation setting, Appendix J specifies the evaluation and scoring protocol, and Section 5.2 analyzes checkpoint-wise training dynamics. Appendix A gives the weighting procedure, and Appendix L records the training hyperparameters. The per-problem evaluation and analysis artifacts are enumerated in Appendix I. The supplementary material includes the core IG-OPSD implementation, unit tests, and compact diagnostic results used in our analysis.

### AI use statement

Generative AI tools were used to identify potentially relevant literature, to assist with editorial restructuring and language revision, and to support limited scripting and experiment orchestration. The authors reviewed and tested AI-assisted code and configuration changes, and checked the resulting artifacts. All decisions about the methodology, reported results, interpretation, and inclusion of references were made by the authors. The authors take full responsibility for the final content of this work.

## References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2306.13649.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=5h0qf7IBZZ.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026. URL https://arxiv.org/abs/2601.20802.

Simran Kaur, Narutatsu Ri, Yinghui He, Liam Fowl, and Sanjeev Arora. Rethinking on-policy self-distillation for thinking models, 2026. URL https://arxiv.org/abs/2607.05184.

Andreas Kirsch and Yarin Gal. Unifying approaches in active learning and active sampling via fisher information and information-theoretic quantities, 2022. URL https://arxiv.org/abs/2208.00549.

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

Xiaogeng Liu, Xinyan Wang, Yingzi Ma, Yechao Zhang, and Chaowei Xiao. When are teacher tokens reliable? position-weighted on-policy self-distillation for reasoning, 2026. URL https://arxiv.org/abs/2605.21606.

Andrei Liviu Nicolicioiu, Mohammad Pezeshki, and Aaron Courville. On-policy self-distillation with sampled demonstrations reduces output diversity, 2026. URL https://arxiv.org/abs/2606.26091.

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

Zhanming Shen, Jintao Tong, Shaotian Yan, Chen Shen, Hao Chen, Wentao Ye, Xiaomeng Hu, Rui Miao, Haobo Wang, Junbo Zhao, Gang Chen, and Jieping Ye. Purified OPSD: On-policy self-distillation without losing how to think, 2026. URL https://arxiv.org/abs/2607.02234.

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, 2026. URL https://arxiv.org/abs/2601.18734.

## A Implementation Notes for IG Weighting

The information-gain score is computed with teacher forcing over the reference solution tokens, after the teacher and student sequences for the sampled rollout have already been constructed. The score is stop-gradient: it selects and rescales examples, but it is not itself optimized as a differentiable auxiliary objective. A z-score and ReLU produce the sample weights. When the ReLU gate removes all examples in a step, the implementation falls back to the uniform OPSD coefficient so that training does not drop the batch.

Figure 9 provides an expanded view of the training flow, including the two conditioning contexts, the teacher-forced IG calculation, and the unchanged per-token OPSD divergence.

## B Prompt and Scoring Templates

The following templates specify the student, teacher, and scoring contexts; exact tokenization follows the Qwen3 tokenizer and repository data collator.

```
Student context:
Problem: {x}
Answer:
{student_rollout_prefix}
```

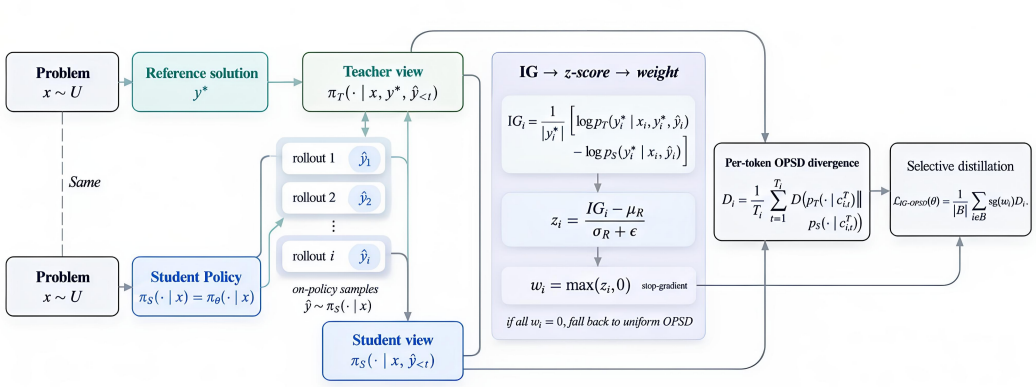


Figure 9: Detailed training flow for IG-OPSD. The student samples on-policy rollouts from the problem-only context, while the privileged teacher also sees the reference solution. A stop-gradient ReLU z-score weight multiplies the per-sample OPSD divergence; if every weight is zero, the step falls back to uniform OPSD.

```
Teacher context:
Problem: {x}
Here is a reference solution: {y_star}
After understanding the reference solution, please try to solve
this problem using your own approach below:
Answer:
{student_rollout_prefix}
```

The IG score appends the reference solution tokens after each context and computes

$$\mathrm{IG}_i = \frac{1}{|y_i^\star|} \sum_{u=1}^{|y_i^\star|} \left[\log p_T(y_{i,u}^\star \mid c_i^T, y_{i,<u}^\star) - \log p_\theta(y_{i,u}^\star \mid c_i^S, y_{i,<u}^\star)\right].$$

This score is detached before weighting the OPSD loss.

## C IG Weight Diagnostics

The implementation logs raw IG mean and standard deviation, reference-window standard deviation and size, active weight fraction, and uniform-fallback frequency. These quantities are recorded under `ig_mean`, `ig_std`, `ig_ref_std`, `ig_ref_size`, `ig_weight_active_frac`, and `ig_fallback_uniform`.

Figure 10 shows the weight transform. The z-score places each example relative to the current global-batch/history reference window, and the ReLU gate removes examples at or below the reference mean.

The logged quantities are implementation diagnostics rather than evaluation metrics. Because a single logged step can be dominated by its local batch composition, we do not compare one-step active or fallback fractions across model sizes; the reproducibility artifacts retain the complete histories.

## D Checkpoint-Wise IG Decomposition

Table 6 expands the competition-set decomposition in Section 5.5. Context share is the ratio of mean absolute context advantage to mean absolute total IG. Rank correlation compares total IG with context advantage across the pooled evaluation problems. Active-set Jaccard compares the examples selected by thresholding each score against its corresponding reference mean. High values throughout training show that privileged context, rather than parameter drift alone, determines most of the total-IG ordering and selected set.

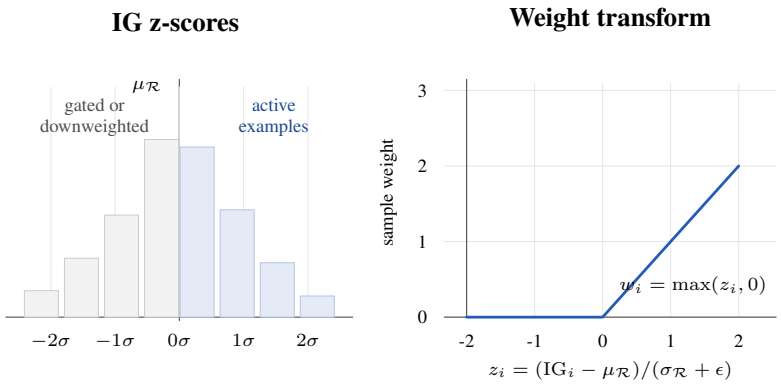


Figure 10: Diagnostic view of the IG-OPSD weighting rule. The left panel shows examples relative to the reference IG window; the right panel shows the ReLU z-score transform.

Table 6: Checkpoint-wise IG decomposition on the pooled competition evaluation sets. $\rho$ is Spearman rank correlation between total IG and context advantage; Jaccard compares their active sets.

| Model | Checkpoint | Context share | $\rho$(total, context) | Active-set Jaccard |
|---|---|---|---|---|
| Qwen3-8B | 25 | 0.958 | 0.995 | 0.965 |
| | 50 | 0.761 | 0.910 | 0.742 |
| | 75 | 0.709 | 0.874 | 0.692 |
| | 100 | 0.696 | 0.859 | 0.746 |
| Qwen3-14B | 25 | 0.986 | 0.999 | 0.927 |
| | 50 | 0.931 | 0.981 | 0.857 |
| | 75 | 0.858 | 0.944 | 0.807 |
| | 100 | 0.846 | 0.944 | 0.807 |

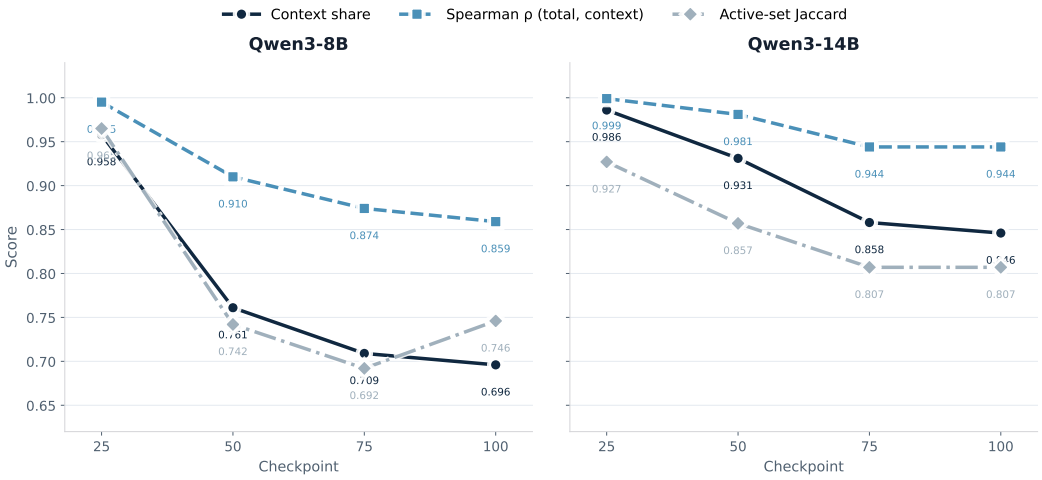


Figure 11: Checkpoint-wise IG decomposition metrics from Table 6. Context share, Spearman rank correlation, and active-set Jaccard all decrease from checkpoint 25 to 100 as parameter drift grows, but remain above 0.69 at 8B and 0.81 at 14B, confirming that context advantage dominates the total-IG ordering throughout training.

## E Computational Overhead

IG-OPSD adds two non-generative, teacher-forced likelihood passes per sample, each scaling with the reference-solution length.

Table 7 reports the profiler's `OPSDTrainer.training_step` timing from the available local run records. To align the Qwen3-8B row with the evaluation endpoint, we use the mean of all available

profiler calls in the original runs through checkpoint 100 (799 calls for OPSD and 800 for IG-OPSD); the single missing OPSD call reflects an incomplete final logger flush, while the checkpoint-100 artifact itself is complete. Job runtime is omitted because it includes scheduling, checkpointing, and resume effects.

Table 7: Available local training-step profiles for OPSD and IG-OPSD. Times are seconds per profiled training step; overhead is IG-OPSD relative to OPSD. The 8B row averages the original runs through checkpoint 100; the other rows use their checkpoint-100 run summaries.

| Model | OPSD step | IG-OPSD step | Overhead | Endpoint |
|---|---|---|---|---|
| Qwen3-1.7B | 20.84 | 21.36 | +2.5% | 100 |
| Qwen3-8B | 57.60 | 60.15 | +4.4% | 100 |
| Qwen3-14B | 18.19 | 18.96 | +4.2% | 100 |

Table 8 reports `train_runtime` through step 100 for the Qwen3-14B distillation runs in Table 2. The totals include rollout generation and parameter updates. All four runs used four GPUs; overhead is calculated from unrounded runtimes relative to OPSD.

Table 8: Trainer-reported Qwen3-14B runtime through checkpoint 100. Overhead is relative to OPSD.

| Method | Runtime (s) | Runtime (h) | Overhead |
|---|---|---|---|
| OPSD | 14,034 | 3.90 | 0.0% |
| PW-OPSD | 14,353 | 3.99 | +2.3% |
| IG-OPSD | 14,637 | 4.07 | +4.3% |
| Purified-OPSD | 16,399 | 4.56 | +16.9% |

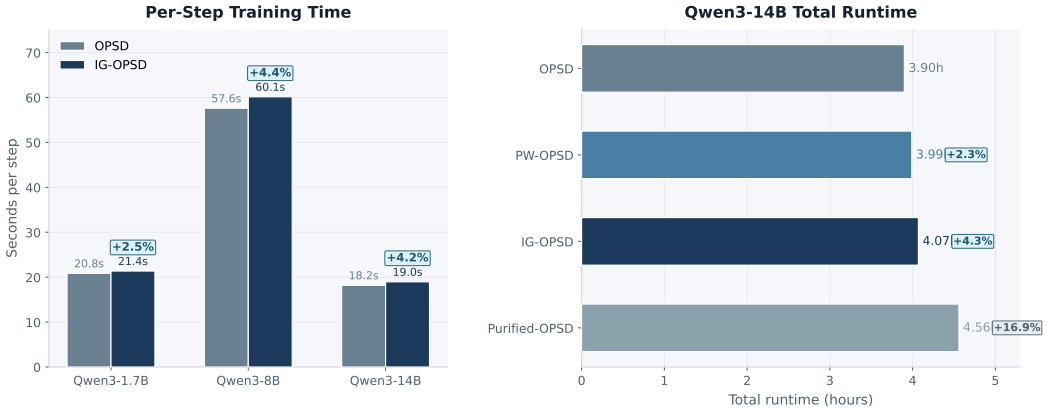


Figure 12: Training-time overhead visualization. Left: per-step time for OPSD versus IG-OPSD across model sizes (Tables 7). Right: Qwen3-14B total runtime through step 100 for all distillation methods (Table 8). IG-OPSD adds 2.5–4.4% per-step overhead; Purified-OPSD is the most expensive at +16.9%.

## F Paired Bootstrap Uncertainty

For each benchmark, single-checkpoint differences should be interpreted with problem-level uncertainty. Table 9 reports paired bootstrap 95% confidence intervals over problem IDs for checkpoint-100 deltas, using the same problem resamples for OPSD and IG-OPSD. Avg@12 uses the per-problem fraction of correct samples; Pass@12 uses the per-problem indicator that at least one of twelve samples is correct. Table 4 additionally reports the precomputed aggregate contrasts: the four-set comparison resamples problems within each dataset and then averages the four dataset means equally, whereas the shuffled controls pool three equally sized 30-problem sets. None of these intervals includes training-seed variation.

Table 9: Checkpoint-100 paired bootstrap confidence intervals for IG-OPSD minus OPSD. Values are percentage points.

| Metric | Model | AIME24 | AIME25 | AMC23 | HMMT25 |
|---|---|---|---|---|---|
| Avg@12 | Qwen3-1.7B | -1.7 [-4.7, +0.8] | -0.3 [-3.6, +2.2] | +2.1 [-2.3, +6.5] | +1.1 [-0.6, +3.3] |
| | Qwen3-8B | +2.2 [-3.6, +8.1] | +1.1 [-2.5, +6.4] | +1.7 [-3.3, +8.3] | +2.8 [-1.1, +7.2] |
| | Qwen3-14B | +0.3 [-4.2, +4.7] | +4.2 [+0.3, +8.6] | +3.1 [+0.2, +6.3] | +5.6 [+1.4, +10.8] |
| Pass@12 | Qwen3-1.7B | +3.3 [-6.7, +16.7] | +6.7 [-6.7, +20.0] | -2.5 [-7.5, +0.0] | +6.7 [-6.7, +20.0] |
| | Qwen3-8B | -3.3 [-16.7, +10.0] | +0.0 [-13.3, +13.3] | +0.0 [-7.5, +7.5] | +6.7 [-6.7, +20.0] |
| | Qwen3-14B | +0.0 [+0.0, +0.0] | +0.0 [-13.3, +13.3] | -2.5 [-7.5, +0.0] | +6.7 [-10.0, +23.3] |

Figure 13 visualizes Avg@12 confidence intervals across all four datasets; Table 9 lists the numeric intervals for all four benchmarks. At 14B, the intervals for AIME25, AMC23, and HMMT25 exclude zero.

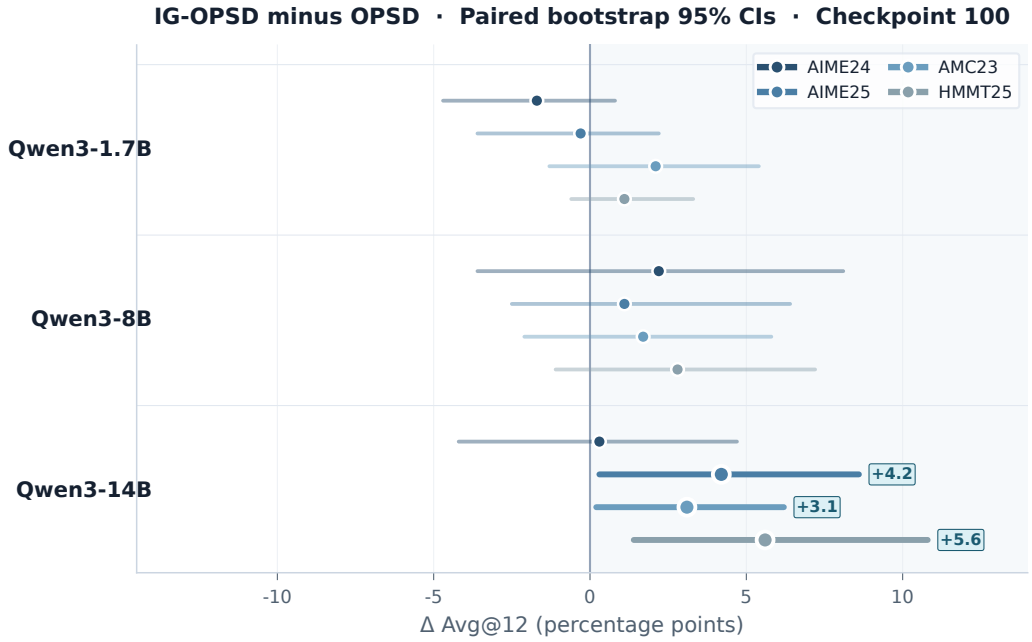


Figure 13: Paired bootstrap 95% intervals for checkpoint-100 Avg@12 deltas on all four competition sets. Points show IG-OPSD minus OPSD; intervals resample problem IDs. Green annotations mark intervals that exclude zero.

## G Per-Problem Pass@12 Win/Loss Decomposition

Table 10 decomposes checkpoint-100 Pass@12 into matched problem categories. "Both" means both methods solve the problem at least once among twelve samples; "IG only" and "OPSD only" identify changed coverage; "Neither" means no correct sample from either method.

Table 10: Checkpoint-100 Pass@12 coverage decomposition. Entries are problem counts out of 30: Both / IG only / OPSD only / Neither.

| Model | AIME24 | AIME25 | HMMT25 |
|---|---|---|---|
| Qwen3-1.7B | 10 / 2 / 1 / 17 | 6 / 3 / 1 / 20 | 4 / 3 / 1 / 22 |
| Qwen3-8B | 17 / 2 / 3 / 8 | 13 / 2 / 2 / 13 | 9 / 3 / 1 / 17 |
| Qwen3-14B | 24 / 0 / 0 / 6 | 17 / 2 / 2 / 9 | 11 / 4 / 2 / 13 |

HMMT25 is the most difficult benchmark in our suite under uniform OPSD. On this benchmark, IG-OPSD improves Pass@12 by 6.7 points at all three model scales (1.7B, 8B, and 14B), suggesting that its gains on difficult problems also translate into broader problem coverage. At 14B, this improvement comprises four problems solved only by IG-OPSD, compared with two solved only by OPSD.

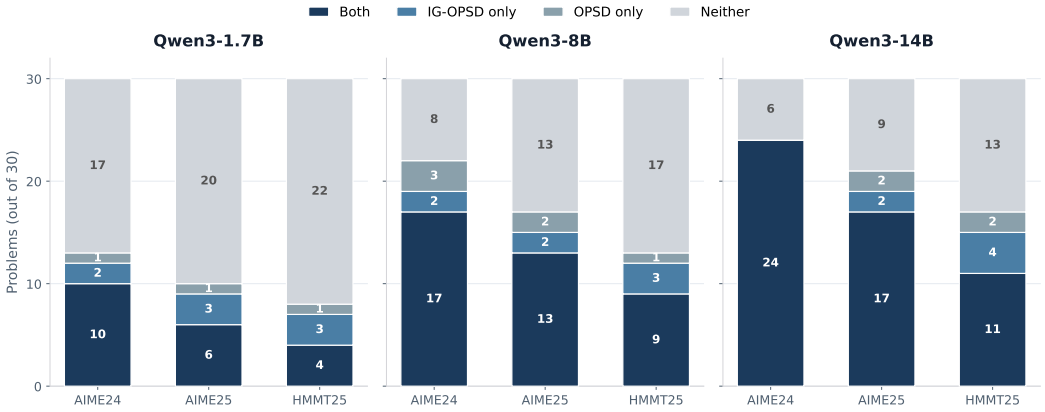


Figure 14: Pass@12 coverage decomposition from Table 10. Each bar shows the 30-problem split into Both solved, IG-OPSD only, OPSD only, and Neither. At 14B, AIME24 shows complete overlap while HMMT25 has the most IG-OPSD-exclusive problems.

## H Output-Quality Sanity Check

At Qwen3-14B checkpoint 100, the IG-OPSD–OPSD changes in repeated 4-gram rate are +0.0006 on AIME24, −0.0028 on AIME25, +0.0042 on AMC23, and +0.0060 on HMMT25. All paired problem-level 95% bootstrap intervals include zero, providing no evidence that the accuracy gain is accompanied by increased local repetition.

## I Artifact Manifest

The implementation and result artifacts used by this draft are:

- `ig_weighting.py`: pure weighting helpers, including z-score, ReLU, hard-gate, raw-IG, rank-gate, fallback, and uniform-mixing utilities.
- `opsd_trainer.py`: OPSD integration, teacher-forced IG scoring, global/history reference-window construction, sample weighting, and training diagnostics.
- `data_collator.py`: solution tokenization and batch fields used for teacher-forced IG scoring.
- `tests/test_ig_weighting.py`: unit tests for z-score behavior, single-sample history-window handling, gating modes, and all-zero fallback behavior.
- `eval/eval_results/`: raw per-problem evaluation JSON files used to compute Avg@12, Pass@12, paired bootstrap intervals, and win/loss decompositions.
- `analysis/purified_opsd/`: saved-generation quality reports used for the output-repetition sanity check.
- `analysis/ig_component_decomposition/`: checkpoint-wise total-IG, context-advantage, parameter-drift, rank-correlation, and active-set reports.
- `analysis/paired_ci/` and `analysis/ig_diagnostics/`: aggregate paired-bootstrap reports and competition-set difficulty stratifications.
- `analysis/runtime_profile/` and local run records: extracted checkpoint-100 timing aggregates used in Appendix E; complete IG histories underlying the implementation diagnostics remain with the run artifacts.

## J Evaluation Protocol and Dataset Handling

Table 11: Competition-math evaluation sets.

| Benchmark | Dataset ID | Split | Problems |
|---|---|---|---|
| AIME24 | `HuggingFaceH4/aime_2024` | train | 30 |
| AIME25 | `yentinglin/aime_2025` | train | 30 |
| AMC23 | `math-ai/amc23` | test | 40 |
| HMMT25 | `MathArena/hmmt_feb_2025` | train | 30 |

Each problem is sampled 12 times in non-thinking mode with temperature 1.0, top-p 0.8, top-k disabled, min-p 0, presence penalty 0, and at most 8192 new tokens. Avg@12 is the fraction correct over all generated solutions; Pass@12 is the fraction of problems with at least one correct solution.

We extract the balanced-brace content of the final `\boxed` occurrence. Answers are parsed with `math_verify` using `fallback_mode=no_fallback` and checked for mathematical equivalence with a five-second timeout. If verification raises an exception, we compare lowercased strings after removing spaces and dollar signs. Missing or malformed boxed answers are counted as incorrect.

## K Reproducibility and Data Sources

All reported numbers are calculated from the local per-problem artifacts under `eval/eval_results/`.

## L Training and Evaluation Hyperparameters

Tables 12–14 record the executed configurations for the primary OPSD versus IG-OPSD comparison and the Qwen3-14B weighting ablations. We separate shared optimization settings from the vocabulary support used by the distillation loss because the latter is not identical in every archived run.

Table 12: Shared training hyperparameters for the primary non-thinking OPSD and IG-OPSD runs. A global batch contains 32 prompts.

| Component | Setting |
|---|---|
| Training data | `siyanzhao/Openthoughts_math_30k_opsd`, train split (29,434 examples); seeded random training sampler, without an additional subset or curriculum. |
| Models and budget | Qwen3-1.7B, 8B, and 14B; 100 optimizer steps; checkpoints every 25 steps. At batch size 32, checkpoint 100 corresponds to 3,200 sampled prompt instances. |
| Batch construction | Per-device batch 1; 2 GPUs $\times$ 16 accumulation steps (1.7B), or 4 GPUs $\times$ 8 accumulation steps (8B/14B); global batch size 32. |
| Optimizer | Fused AdamW, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, weight decay 0, gradient clipping 0.1; learning rate $5 \times 10^{-6}$ with a linear scheduler and 0 warmup steps. |
| Numerics | bfloat16, gradient checkpointing, SDPA attention, maximum sequence length 8192. |
| LoRA | Rank 64, $\alpha = 128$, dropout 0.05, no bias; target modules `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj`. |
| Rollout sampling | On-policy student; maximum completion length 1024, temperature 1.1, top-p 0.95, top-k 20; student and fixed teacher both use non-thinking mode. |
| Token objective | $\beta = 0$ gives the forward divergence $D_{\mathrm{KL}}(p_T \,\|\, p_\theta)$. OPSD averages it uniformly over samples; IG-OPSD applies one detached sample weight before the batch reduction. |
| IG weighting | Length-normalized reference log-likelihood gap; z-score over the gathered global batch plus a 256-value history; ReLU; $\epsilon = 10^{-6}$; mean-one renormalization; uniform fallback if all weights are zero. |

Table 13: Executed divergence support and clipping for the primary OPSD and IG-OPSD runs. "Full" means no `top_k_loss` restriction. The clip is applied to each vocabulary-component contribution before summation.

| Model | OPSD support | IG-OPSD support | Token clip | $\beta$ |
|---|---|---|---|---|
| Qwen3-1.7B | Teacher top-128 | Teacher top-128 | 0.05 | 0 |
| Qwen3-8B | Full vocabulary | Full vocabulary | $10^{-7}$ | 0 |
| Qwen3-14B | Teacher top-256 | Teacher top-256 | $10^{-7}$ | 0 |

Table 14: Hyperparameter differences among the Qwen3-14B checkpoint-100 weighting ablations in Table 3. Parameters not shown are shared with Table 12.

| Variant | Weight rule | History | Shuffle | Renorm. | Loss support |
|---|---|---|---|---|---|
| OPSD (uniform) | $w_i = 1$ | — | No | — | Top-256 |
| Binary IG gate | $\mathbf{1}[\mathrm{IG}_i > \mu_{\mathcal{R}}]$ | 256 | No | Yes | Top-256 |
| Shuffled IG weights | ReLU z-score | 256 | Yes | Yes | Top-256 |
| Current-batch z-score | ReLU z-score | 0 | No | Yes | Top-256 |
| Positive raw IG | $\max(\mathrm{IG}_i, 0)$ | Not used | No | Yes | Top-256 |
| IG-OPSD | ReLU z-score | 256 | No | Yes | Top-256 |

All rows use global batch size 32, $\beta = 0$, token clipping at $10^{-7}$, and teacher top-256 loss support. Thus the ablations retain the same token-level objective and differ only in the sample-weighting rule shown in the table.