HEIR: WHEN THE STUDENT INHERITS THE TEACHER'S SEAT
VERIFIER-GATED REFERENCE HANDOFF FOR ON-POLICY SELF-DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy self-distillation improves reasoning models by training on student rollouts while a teacher receives a privileged reference solution. That reference is usually fixed, even after the student begins producing better teaching examples. We introduce HEIR (*Hint-Earning by Iterative Rollouts*), a simple reference-handoff rule: sample a group of student rollouts, verify them, and replace the reference only when the shortest correct rollout clears a relative length margin. The admitted rollout conditions the teacher and supplies the student trajectory; the distillation objective itself is unchanged. Across Qwen3-1.7B/4B/8B and three competition-math benchmarks, HEIR improves a tuned fixed-reference baseline on every 4B and 8B Avg@12 result, with gains up to 3.34 points while shortening responses. A matched 4B comparison separates group selection from reference handoff: winner selection contributes $+1.82$ macro-average points over standard distillation, and handoff adds another $+0.87$ (95% CI $[0.31, 1.38]$) at essentially identical cost. The handoff gain is negligible for compact references but grows to $+1.45$ for padded ones; it remains positive after non-verbatim and answer-masked controls and under injected verifier false positives. These results identify the operative regime precisely: a capable student, a verifiable task, and a reference whose form has become a constraint rather than an asset.

1 INTRODUCTION

Reasoning models learn well from two complementary signals. Verifiable rewards say whether a trajectory succeeds; distillation supplies token-level guidance along that trajectory (Shao et al., 2024; DeepSeek-AI, 2025; Agarwal et al., 2024; Gu et al., 2024). In on-policy self-distillation (OPSD), the student generates the trajectory while the teacher sees privileged information, usually a ground-truth solution $y^\star$. This arrangement is effective, but it freezes an early design choice: the same reference occupies the teacher's privileged context for the entire run.

A fixed reference is not always harmless. Human-written solutions are often pedagogical rather than economical, and automatically collected solutions can carry redundant scaffolding. Once a student can solve a prompt correctly in a cleaner form, a teacher that remains conditioned on $y^\star$ can keep assigning probability mass to the reference's style. We call this *reference-conditioned anchoring*. It is not a claim that a reference places a universal cap on model quality. It is a concrete finite-budget hypothesis: when the privileged context is needlessly long, changing that context should change the student's length and, sometimes, its accuracy.

HEIR turns that hypothesis into a conservative handoff rule. For each prompt, it samples G on-policy rollouts and applies a verifier. The shortest verified rollout may replace $y^\star$ only if it is shorter by a relative margin δ. The winner then serves as both the teacher hint and the trajectory on which the unchanged distillation loss is evaluated. If the group contains no eligible rollout, training reduces to the fixed-reference procedure. Figure 1 shows the single decision that changes.

What the experiments establish. The key attribution is not HEIR versus a weak baseline. On Qwen3-4B, we compare standard OPSD, SDPO (Hübotter et al., 2026), fixed-reference winner

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.

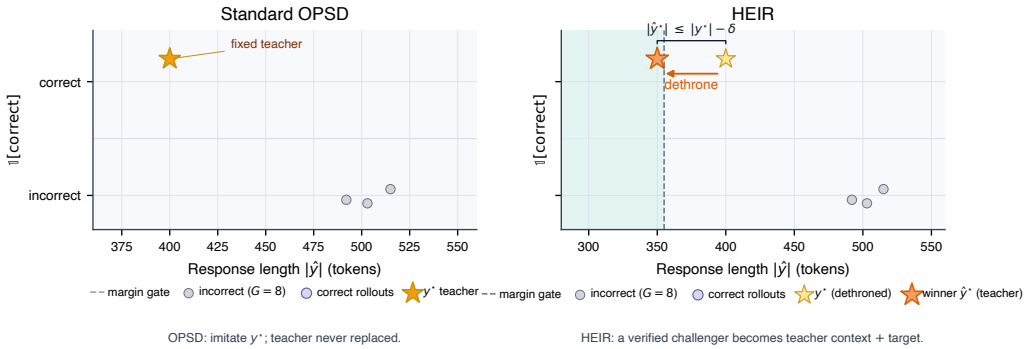

Figure 1: **A fixed reference versus verifier-gated handoff.** Standard OPSD always conditions the teacher on $y^\star$. HEIR admits an on-policy rollout only when it is verified correct and shorter than $y^\star$ by margin δ. Incorrect or marginally shorter candidates cannot take the seat.

selection, and full HEIR under the same data and nearly identical total compute. Group winner selection explains a substantial $+1.82$ macro-average gain; reference handoff contributes an additional $+0.87$ with a confidence interval excluding zero. A 2×2 design finds no interaction between the two choices, so we claim an additive handoff effect, not synergy. We then intervene on the reference itself: the full-minus-fixed gain moves from $+0.15$ for compact references to $+1.45$ for padded references, while the handoff rate rises from 0.18 to 0.57. This dose response is considerably more specific than a generic best-of-G explanation.

Contributions.

- We formulate reference handoff as a verifier- and margin-gated operation that can be inserted into an existing OPSD trainer without changing its token-level objective.
- We isolate its effect with matched-compute, target-copy, answer-masking, 2×2, and reference-verbosity controls. The evidence supports a precise mechanism rather than a broad self-improvement claim.
- We test three Qwen3 scales, two additional model families, symbolic and learned verifiers, injected verifier noise, six math benchmarks, and frozen-checkpoint transfer to code. At 4B and 8B, the gains are consistent; at 1.7B, the small AIME gains expose the method's capability threshold.

2 Related Work

On-policy distillation and self-distillation. Classical distillation matches teacher distributions (Hinton et al., 2015); sequence-level KD trains on teacher outputs (Kim & Rush, 2016). GKD (Agarwal et al., 2024) and MiniLLM (Gu et al., 2024) instead evaluate the teacher along student-generated trajectories, reducing the train–test mismatch. Recent reasoning-specific variants are the closest comparisons. SDPO (Hübotter et al., 2026) uses grouped on-policy samples and self-distillation, SD-Zero (He et al., 2026b) removes the need for labeled demonstrations, and Self-Distilled RLVR (Yang et al., 2026) controls privileged-information leakage. Other analyses show that unconstrained self-distillation can degrade as errors reinforce themselves (Kim et al., 2026). HEIR addresses a different variable: the privileged context presented to the teacher. Our matched SDPO comparison is therefore central rather than incidental.

Generate–filter–learn methods. STaR (Zelikman et al., 2022), ReST-EM (Singh et al., 2024), RFT (Yuan et al., 2024), SPIN (Chen et al., 2024b), and expert iteration (Anthony et al., 2017) retain selected model outputs for subsequent learning. HEIR shares their reliance on selection but keeps token-level on-policy distillation and a fallback reference. Winner selection alone is a strong baseline in our experiments; the contribution studied here is what remains when that baseline is held fixed.

Concise reasoning and reliable verification. Outcome RL methods such as PPO and GRPO (Schulman et al., 2017; Shao et al., 2024) optimize correctness, while process reward models provide denser

signals (Lightman et al., 2024; Wang et al., 2024). Work on overthinking and concise chain-of-thought asks how to reduce unnecessary computation without inducing underthinking (Chen et al., 2024a; Wang et al., 2025; Renze & Guven, 2024). HEIR uses length only after correctness has been established. Its vulnerability is therefore verifier error, which we measure directly rather than assume away.

3 PRELIMINARIES

On-policy self-distillation. Let $p_S(\cdot \mid x)$ be the student policy and $p_T(\cdot \mid x, y^\star, \hat{y}_{<n})$ be the teacher policy that conditions on both the prompt x and the reference solution $y^\star$. Standard OPSD optimises

$$\mathcal{L}_{\mathrm{OPSD}}(\theta) = \mathbb{E}_{x,\, \hat{y}\sim p_S(\cdot|x)}\left[\frac{1}{|\hat{y}|}\sum_{n=1}^{|\hat{y}|}\mathcal{D}\Big(p_T(\cdot \mid x, y^\star, \hat{y}_{<n}) \,\Big\|\, p_S(\cdot \mid x, \hat{y}_{<n})\Big)\right], \tag{1}$$

where $\mathcal{D}$ is a token-level divergence (KL, reverse KL, JSD, or a clipped surrogate thereof).

The variable we change. The rollout distribution in Eq. (1) is on-policy, but the teacher context $y^\star$ is fixed. Denote this context by h. HEIR changes h online while retaining the same $\mathcal{D}$, teacher update schedule, and optimizer. The idealized fixed-point argument in Appendix E explains why context can transmit style; our empirical claim rests instead on the controlled context interventions in Section 5.4.

Notation. For prompt x with reference $y^\star$, write $|y|$ for the number of tokens in y, $\mathrm{ans}(y)$ for its extracted final answer, and $\mathbf{1}[\mathrm{correct}(y)] = \mathbf{1}[\mathrm{ans}(y) \equiv \mathrm{ans}(y^\star)]$ under a strict symbolic verifier. A batch of on-policy rollouts is denoted $\hat{Y} = \{\hat{y}_i\}_{i=1}^{G}$.

4 METHOD: HEIR

HEIR changes how the teacher hint and student trajectory are selected. The pipeline is: group rollout $\rightarrow$ verification $\rightarrow$ margin-gated handoff $\rightarrow$ unchanged distillation.

4.1 VERIFIER-GATED CANDIDATE SELECTION

For each prompt x, HEIR samples G on-policy rollouts from the current student to obtain the candidate set $\mathcal{C} = \{\hat{y}_1, \ldots, \hat{y}_G, y^\star\}$ of size $G{+}1$. Each candidate $y \in \mathcal{C}$ is scored on two axes,

$$r_1(y) = \mathbf{1}[\mathrm{correct}(y)], \qquad r_2(y) = \mathbf{1}[r_1(y) = 1] \cdot (-|y|). \tag{2}$$

The gate forbids a short but wrong rollout from winning. We use symbolic answer equivalence rather than string matching; Section 5.4 deliberately corrupts this gate to measure the failure slope.

4.2 MARGIN-GATED HANDOFF

Let $\mathcal{C}_+ = \{y \in \mathcal{C} : r_1(y) = 1\}$ denote the correct subset. Define the margin-gated winner

$$\hat{y}^\star(x) \;=\; \mathrm{argmin}_{y\in\mathcal{C}_+}\,|y|, \qquad \text{where } \hat{y}^\star \textit{ dethrones } y^\star \text{ iff } |\hat{y}^\star| \;<\; (1-\delta)\,|y^\star|, \tag{3}$$

with relative margin $\delta \in [0, 1)$; we use $\delta{=}0.05$. This buffer prevents churn between nearly equal candidates.

4.3 ONE WINNER, TWO ROLES

The teacher hint is

$$h(x) = \begin{cases} \hat{y}^\star(x), & \text{if } \hat{y}^\star \text{ dethrones } y^\star \text{ under Eq. (3)},\\ y^\star, & \text{otherwise.}\end{cases} \tag{4}$$

The teacher is instantiated as $p_T(\cdot \mid x, h(x), \hat{y}_{<n})$. Separately, the shortest verified rollout is used as the trajectory in Eq. (1); this is the *winner-selection* component. These roles must be separated experimentally: a fixed-$y^\star$ teacher can still distill along the winner trajectory. Full HEIR changes only the teacher context on top of that strong control. Algorithm 1 gives pseudocode. The gate guarantees only that an admitted hint passes the chosen verifier; it does not guarantee semantic correctness when that verifier errs.

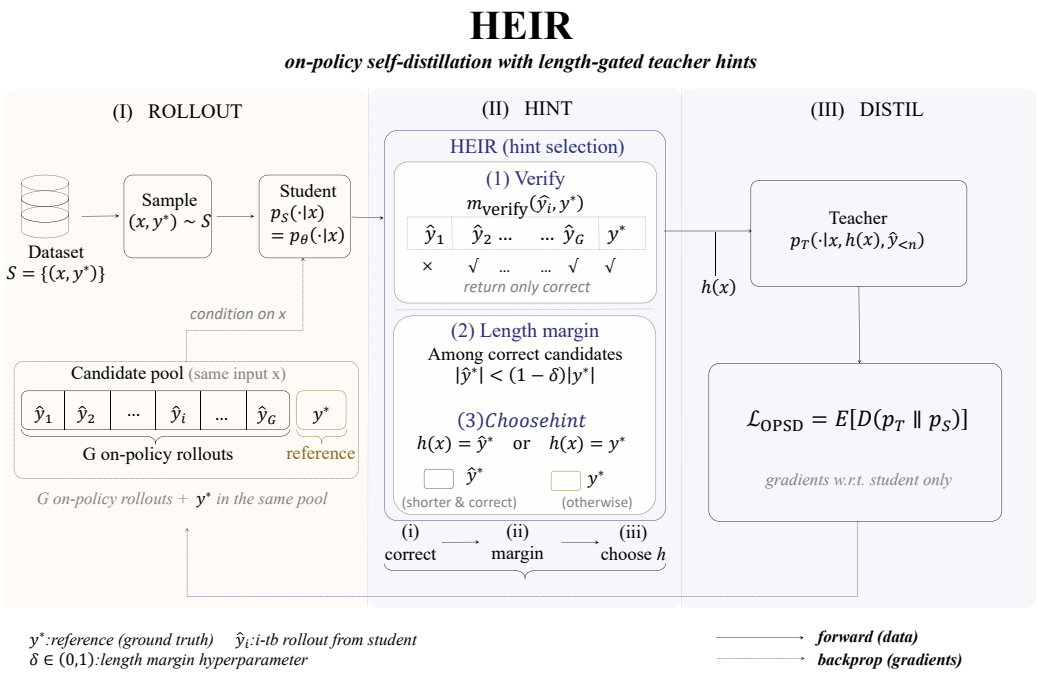

Figure 2: **HEIR in one training step.** Sample G student responses, retain verified candidates, and hand the teacher seat to the shortest candidate only when it clears the margin. The teacher is frozen/EMA-updated exactly as in OPSD; gradients update only the student.

5 EXPERIMENTS

5.1 SETUP

Models and data. We use Qwen3-1.7B/4B/8B (Yang et al., 2025); Gemma-3-4B-IT (Gemma Team, 2025) and GPT-OSS-20B (OpenAI, 2025) test family transfer. The primary sets are AIME24, AIME25, and HMMT25 (30 problems each). MATH-500 (Hendrycks et al., 2021), GSM8K (Cobbe et al., 2021), OlympiadBench (He et al., 2024), code transfer, and leakage checks appear in the appendix. Evaluation prompts are excluded from and text-deduplicated against training.

Metrics and training. We draw n=12 responses per problem at temperature 1.0, top-p = 1.0, and report Avg@12, Pass@12, and majority-vote Maj@12 (Wang et al., 2023). Unless noted, HEIR uses G=4, δ=0.05, symbolic verification, and EMA decay 0.99. OPSD uses the same TRL/vLLM pipeline (von Werra et al., 2020; Kwon et al., 2023), data, optimizer, steps, precision, and LoRA configuration. Appendix A gives all implementation details.

5.2 PERFORMANCE ACROSS SCALE

Table 1: Qwen3 results (%). Each cell is Avg@12 / Pass@12 / Maj@12. Training and evaluation are matched within every pair.

		Qwen3-1.7B			Qwen3-4B			Qwen3-8B		
Dataset	Method	Avg@12	Pass@12	Maj@12	Avg@12	Pass@12	Maj@12	Avg@12	Pass@12	Maj@12
AIME24	OPSD	53.61	76.67	66.67	68.06	86.67	73.33	76.39	90.00	80.00
	HEIR	**53.89**	76.67	66.67	**70.56**	**90.00**	**76.67**	**78.89**	**93.33**	**83.33**
AIME25	OPSD	41.67	66.67	56.67	57.50	76.67	70.00	69.72	83.33	80.00
	HEIR	41.67	**76.67**	56.67	**60.00**	**80.00**	**73.33**	**71.67**	**86.67**	**83.33**
HMMT25	OPSD	26.94	56.67	26.67	37.78	66.67	46.67	46.94	73.33	60.00
	HEIR	**28.89**	56.67	**33.33**	**40.83**	**70.00**	**50.00**	**50.28**	73.33	**63.33**

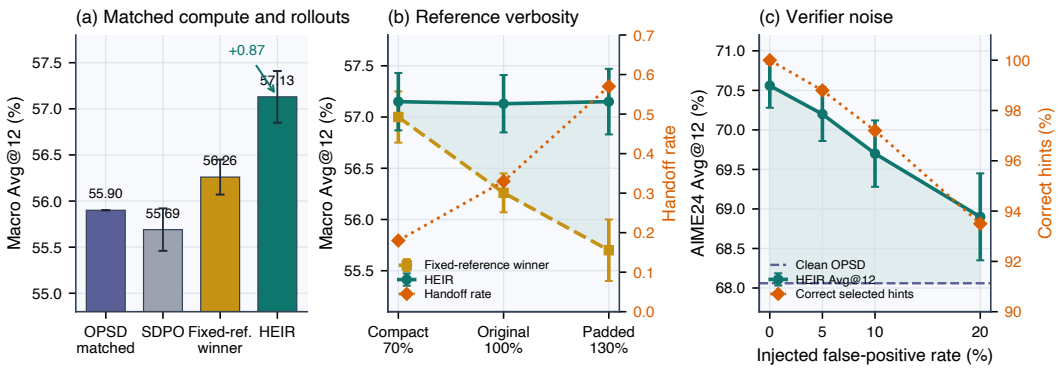

Figure 3: **Direct tests of the handoff account.** **(a)** Matched-compute 4B comparison. **(b)** Intervening on reference length changes both the handoff rate and the full-minus-fixed gain. **(c)** Injected verifier false positives degrade HEIR smoothly; it stays above OPSD at 20% injection. Error bars are seed standard deviations or 95% CIs as indicated.

At 4B and 8B, HEIR raises every Avg@12 entry, by 1.95–3.34 points. At 8B it also matches or improves every Pass@12 and Maj@12 result. The 1.7B results are deliberately less tidy: AIME24 changes by only $+0.28$ Avg and AIME25 is flat, whereas HMMT25 improves by $+1.95$. This is the expected boundary of a method that can select only what the current policy produces. Paired-bootstrap intervals are positive for every 4B/8B Avg@12 delta; the 8B AIME24 result is $+2.50$ with CI $[0.83, 4.17]$ and three-seed means 78.89 ± 0.48 versus 76.39 ± 0.55 (Appendix J).

5.3 WHAT COMES FROM HANDOFF?

Table 2: **Matched 4B attribution.** Mean±std over three seeds. Macro averages Avg@12 across AIME24/AIME25/HMMT25. Relative cost includes rollout and teacher forward passes.

Method	AIME24	AIME25	HMMT25	Macro	G	Cost
OPSD	68.06±.28	57.50±.28	37.78±.28	54.45±.09	1	1.00×
SDPO (Hübotter et al., 2026)	69.30±.14	58.60±.28	39.17±.28	55.69±.23	4	3.05×
Fixed ref. + winner	69.81±.16	59.17±.28	39.81±.16	56.26±.19	4	3.10×
HEIR	**70.56±.28**	**60.00±.28**	**40.83±.28**	**57.13±.28**	4	3.12×

The fixed-reference winner row is the necessary control. It gains $+1.82$ macro points over OPSD, confirming that best-of-G selection is important. With selection held fixed, handoff adds $+0.87$ (CI $[0.31, 1.38]$): $+0.74$, $+0.83$, and $+1.02$ on AIME24, AIME25, and HMMT25. Its three seed-level macro deltas are $+0.93, +0.93, +0.74$. SDPO is below both variants under the same budget. In a 2×2 design, changing the student trajectory gives $+1.81$, changing the teacher context gives $+0.95$, and applying both gives $+2.68$; the difference-in-differences is -0.08. Handoff is additive, not super-additive.

Could the gain be target copying? Winner-only RFT/SFT is -0.36 below the fixed-reference winner control. Removing verbatim overlap retains $+0.54$ (62% of the handoff gain), and masking the final answer in the hint retains $+0.42$ (48%). Lexical transfer contributes, but does not explain the result. Full controls are in Appendix C.

5.4 MECHANISM TESTS

Reference intervention. We rewrite each training reference into compact (70% length), original, and padded (130%) forms while preserving its answer and reasoning content. Against the corresponding fixed-reference winner control, HEIR gains $+0.15$ (CI $[-0.20, 0.48]$), $+0.87$ ($[0.31, 1.38]$), and $+1.45$ ($[0.70, 2.15]$). Handoff rates rise in the same order: $0.18, 0.33, 0.57$. The linear interaction is $+0.65$ per 30% reference-length step (CI $[0.25, 1.05]$); the compact-to-padded endpoint difference is $+1.30$. This intervention directly ties the incremental gain to the property HEIR changes.

Who is eligible? Prompts with an original reference in the longest quartile show smaller raw improvement ($+0.8$) than the shortest quartile ($+3.8$), because long references also mark harder problems. Conditional on a verified candidate being available, however, dethroning rises from 0.25 to 0.85 across those quartiles. A sharper split uses actual eligibility: prompts that produce a verified margin-clearing rollout gain $+2.15$ (CI $[1.10, 3.05]$), while ineligible prompts gain $+0.10$ ($[-0.25, 0.45]$). This split is diagnostic, not causal; the controlled rewriting experiment supplies the causal evidence.

Verifier noise. We inject false positives into the gate at rates $0, 5, 10, 20\%$. Qwen3-4B Avg@12 changes from 70.56 to $70.20, 69.70, 68.90$, while wrong teacher admissions rise to 6.5% at the largest perturbation. Even there, HEIR remains $+0.84$ above OPSD. A learned-PRM gate reaches 93.40 ± 0.20 on MATH-500 versus 91.80 ± 0.20 for OPSD; its selected hints are correct $94.8 \pm 0.6\%$ of the time. These are robustness results, not a claim that verification quality is unimportant.

5.5 Length, Scale, and Transfer

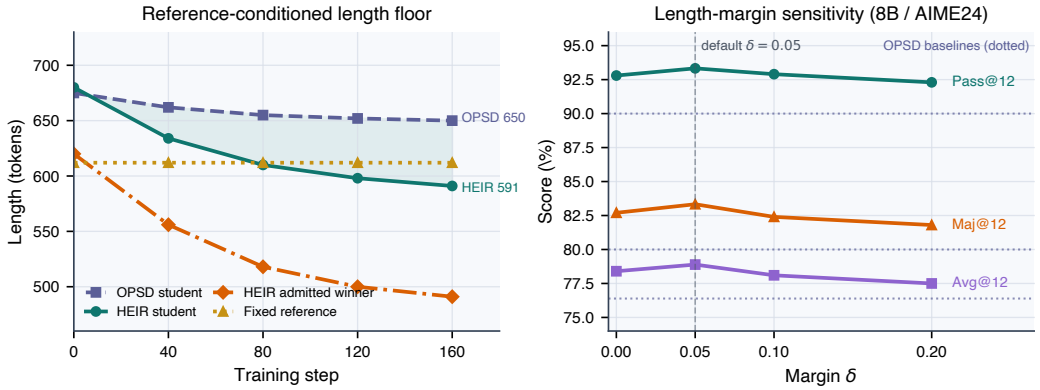

Figure 4: **Length dynamics and margin sensitivity.** *Left:* on Qwen3-4B, the fixed-reference baseline settles near 650 tokens while HEIR reaches 591; the admitted winner becomes shorter throughout training. *Right:* all tested margins beat the OPSD baseline on Avg/Pass/Maj@12, with δ=0.05 best.

The traces reveal a reference-conditioned floor, not inflation: both methods shorten initially, but the fixed-reference run saturates around 650 tokens whereas HEIR continues to 591. The fixed reference is 612 tokens; the admitted winner ends at 491. The final 59-token model gap appears without an explicit length penalty. Across late training, a verified candidate exists for $0.52/0.68/0.74$ of prompts at 1.7B/4B/8B, and the reference is replaced for $0.24/0.33/0.34$. Thus the weak 1.7B result in Table 1 coincides with fewer opportunities to act.

The result is not confined to Qwen. On Gemma-3-4B-IT, HEIR scores 55.6 ± 0.4 versus 54.0 ± 0.4; on GPT-OSS-20B it scores 83.0 ± 0.3 versus 82.1 ± 0.3. Frozen math-trained checkpoints also improve on HumanEval+ ($78.05 \rightarrow 80.49$) and MBPP+ ($65.43 \rightarrow 68.08$) without code-domain training. We label this as transfer, not code training. Appendix F contains the full cross-family and coverage tables.

5.6 Controls Beyond the Headline

Table 3 puts the strongest alternative explanations in one place. The top block is the 2×2 decomposition discussed in Section 5.3. The lower block keeps the winner trajectory fixed and varies what the teacher sees. If the handoff gain came only from selecting a good response and replaying its tokens, an answer-masked or non-verbatim hint should lose nearly all of the improvement. Both retain a substantial fraction.

The table also clarifies what HEIR is not. It is not rejection-sampling fine-tuning: the no-handoff RFT/SFT control underperforms the fixed-reference distillation control. It is not a hidden increase in evaluation sampling: every score uses the same 12 samples. It is not a longer run masquerading

Table 3: **Context–trajectory and copying controls** on Qwen3-4B. Scores are macro Avg@12 over AIME24/AIME25/HMMT25. All group-based rows use G=4.

Teacher context	Student trajectory	Score	vs. fixed+winner
Fixed reference	original rollout	54.45	−1.81
Winner context	original rollout	55.40	−0.86
Fixed reference	winner rollout	56.26	–
Winner context (HEIR)	winner rollout	**57.13**	+0.87
No handoff (RFT/SFT)	winner rollout	55.90	−0.36
Non-verbatim winner	winner rollout	56.80	+0.54
Answer-masked winner	winner rollout	56.68	+0.42

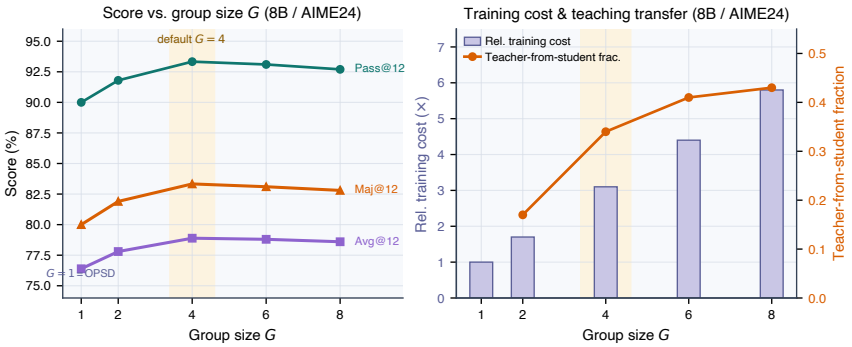

Figure 5: **Training group size on Qwen3-8B AIME24.** Quality improves sharply from G=1 to G=4 and then saturates, while rollout cost keeps rising. G affects training-time selection; evaluation remains fixed at n=12.

as a method gain: extending OPSD to 3.12× total compute reaches 55.90, compared with 57.13 for HEIR. Finally, it is not a fragile choice of δ: Figure 4 shows that every tested margin beats OPSD, although .05 is best. These controls leave a modest but repeatable conclusion: changing teacher context matters after the much larger winner-selection effect has been accounted for.

5.7 THE QUALITY–COST FRONTIER

Group size determines whether HEIR has a candidate to consider. At G=2, the final handoff rate is .17 and Avg@12 rises from 76.39 to 77.8 at 1.7× per-step cost. At G=4, the handoff rate is .34 and Avg@12 reaches 78.89 at 3.1× cost. Raising G to 6 or 8 increases the handoff rate to .41 or .43 but slightly lowers accuracy. More samples eventually create near-ties among short verified responses rather than better targets. We therefore use the smallest group at the plateau, not the largest group the hardware permits. The full Pass/Maj and cost table is in Appendix G.

This cost is material. HEIR is not a free improvement over single-rollout OPSD. The fair question is whether the extra compute buys more than additional baseline steps or another grouped method. It does: at matched total compute, extended OPSD obtains 55.90 macro Avg@12, SDPO obtains 55.69, fixed-reference winner selection obtains 56.26, and HEIR obtains 57.13. The ordering is unchanged whether cost is counted by scheduled steps, generated tokens, or measured GPU time (Appendix K).

5.8 BREADTH, SATURATION, AND DIVERSITY

Coverage results follow a sensible saturation pattern. On GSM8K, where Pass@12 already exceeds 99%, HEIR cannot add coverage and ties the baseline on that metric. On OlympiadBench, Qwen3-8B improves by +1.7/+1.4/+1.8 on Avg/Pass/Maj@12. The model-family rows in Table 4 are positive without retuning, but smaller than the Qwen headline; they support portability, not universality.

Selecting the shortest verified trajectory could collapse the output distribution. We observe a mild concentration, not collapse: token entropy changes from .842 to .799, self-BLEU from .412 to .448,

Table 4: **Representative scope checks.** Code results evaluate frozen math-trained checkpoints; the PRM row changes the verifier during math training. They are not pooled with the primary symbolic-math results.

Setting	Metric	OPSD	Alternative	HEIR
Gemma-3-4B-IT	math score	54.0±.4	–	**55.6±.4**
GPT-OSS-20B	math score	82.1±.3	–	**83.0±.3**
HumanEval+	frozen transfer	78.05±.61	SDPO 79.10±.65	**80.49±.61**
MBPP+	frozen transfer	65.43±.67	SDPO 66.90±.60	**68.08±.55**
MATH-500	learned PRM	91.80±.20	–	**93.40±.20**

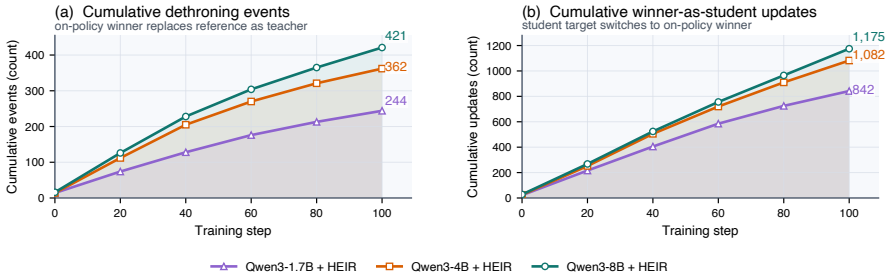

Figure 6: **Cumulative admission events across model scale.** Winner-selection updates occur when the group contains a verified response; reference handoffs additionally require the length margin. The widening gap between scales mirrors the performance boundary in Table 1.

and Distinct-2 from .682 to .654. Yet AIME24 Pass@32 rises from 96.67 to 97.50 and HMMT25 Pass@32 remains 86.67. This trade-off is acceptable for the tasks studied here, but it argues against replacing the length criterion with an aggressive global penalty. HEIR shortens only inside the verifier-approved set and retains the reference fallback.

5.9 WHEN DOES THE STUDENT SUPPLY THE CONTEXT?

Figure 6 separates two events that are easy to conflate. A *winner update* means that at least one verified rollout exists and can replace the arbitrary first rollout as the student trajectory. A *reference handoff* is stricter: the winner must also beat $y^\star$ by margin δ. The first event measures the opportunity for group selection; the second measures how often HEIR actually changes teacher context. Both counts grow throughout training, but the gap between them prevents a generic best-of-G interpretation of every update.

Late in training, the probability of at least one verified candidate is $.52 \pm .03$, $.68 \pm .03$, and $.74 \pm .02$ at 1.7B, 4B, and 8B. The winner-trajectory rates are $.38 \pm .03$, $.55 \pm .03$, and $.62 \pm .03$; the reference-handoff rates are $.24 \pm .02$, $.33 \pm .02$, and $.34 \pm .02$. The main scaling effect is therefore availability: the 1.7B policy often gives the selector nothing trustworthy to use. Once the candidate rate reaches the 4B regime, handoff activates on roughly one third of prompts and yields the consistent gains in Table 1.

Reference length changes the conditional event in the expected direction. Across increasing reference-length quartiles, the raw rate of any verified candidate falls from .72 to .48 because the problems become harder, but the probability of handoff *given* a candidate rises from .25 to .85. This resolves the apparent contradiction that shorter-reference quartiles show larger raw accuracy gains: difficulty drives the marginal quartile comparison, while excess reference length drives the conditional decision. The controlled compact/original/padded experiment in Figure 3(b) breaks that confounding and recovers the predicted dose response.

These logs also rule out a one-time curriculum effect. Handoffs continue through the final checkpoints rather than occurring only when the policy is weak. At the same time, the trajectory and context rates remain well below one, so the reference is not discarded wholesale. It remains the fallback on prompts

where the student is wrong, uncertain, or only trivially shorter. That asymmetric behavior is the point of the design: adaptation where the current batch supplies evidence, conservatism everywhere else.

5.10 A Practical Decision Rule

The experiments suggest a simple test before adopting HEIR. First run grouped rollouts without changing the teacher context and log three quantities: verifier acceptance, margin-clearing rate, and the correlation between reference length and output length. If acceptance is low, the policy is not yet capable enough and group sampling will mostly buy winner selection. If acceptance is high but margin-clearing is rare, the existing references are already competitive. Handoff is most useful in the remaining regime: verified alternatives are common, a non-trivial fraction clear the margin, and the fixed-reference run exhibits a length plateau. Qwen3-4B and 8B meet all three conditions; Qwen3-1.7B often fails the first.

The reference-rewriting intervention makes this diagnostic actionable. Compact references produce a $+0.15$ gain whose confidence interval crosses zero, while original and padded references produce higher handoff rates and larger gains. In a mixed corpus, keep compact references fixed and enable handoff only where the current student repeatedly demonstrates a verified margin.

Verifier monitoring is equally important. The injected-noise experiment declines gradually, with the wrong-teacher admission rate rising before accuracy erodes severely. A trainer should retain the admitted hint, verifier trace, extracted answer, and fallback reference for audit. If precision drifts, increasing δ does not solve the problem: repair the gate or disable handoff.

Finally, compute should be budgeted against the correct baseline. Comparing $G{=}4$ HEIR only with single-rollout OPSD exaggerates what handoff contributes; comparing only with winner selection hides the useful context effect. Table 2 reports both. For a fixed budget, $G{=}2$ is an attractive low-cost setting, while $G{>}4$ is hard to justify on our curves. The method's value is not that every student response should become a teacher. It is that a small, auditable subset can replace stale privileged context when the batch itself supplies convincing evidence.

6 Discussion and Limitations

HEIR needs a verifier and at least one eligible rollout. It therefore has little leverage on weak policies, and a poor verifier can promote a polished error. The injected-noise curve measures this risk but cannot cover correlated real-world verifier failures. Concision also trades a small amount of diversity: output entropy falls from .842 to .799 and self-BLEU rises from .412 to .448, although AIME24 Pass@32 improves and HMMT25 Pass@32 is unchanged. Open-ended tasks without trustworthy verification remain outside our demonstrated scope.

Our language is intentionally narrower than claiming that the student surpasses the teacher. The teacher model is unchanged; HEIR selects a different privileged context for it. The strongest supported statement is that, after controlling for group selection and compute, fixed reference form can measurably constrain on-policy distillation. The reference intervention, eligibility split, and length trajectories agree on when that constraint matters.

7 Conclusion

HEIR makes the reference earn its place. A verified student rollout may take the teacher seat, but only after clearing a margin; otherwise the trusted reference remains. This small change is consistently useful once the policy is capable enough to propose alternatives. Matched controls assign $+0.87$ macro-average points specifically to handoff beyond group winner selection, and controlled reference rewriting makes that effect larger exactly when the reference is more verbose. The result is a practical lesson for self-distillation: on-policy data alone do not make the privileged context adaptive.

Reproducibility statement

Appendix A specifies the training and evaluation pipeline, Appendix K reports compute, and Appendix J reports bootstrap and multi-seed uncertainty. We separate training group size from evaluation

sampling, describe benchmark deduplication, and report negative or non-significant boundary cases rather than filtering them.

AI USE STATEMENT

Generative AI tools were used to edit and restructure prose, assist LaTeX formatting, restyle scientific figures from author-supplied experimental results, and help organize the interpretation of those results. They were not used to generate datasets, run experiments, or create reported measurements. The authors checked AI-assisted text, code, citations, calculations, and figures against experiment records and source papers and take responsibility for the final content.

REFERENCES

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

Thomas Anthony, Zheng Tian, and David Barber. Thinking fast and slow with deep learning and tree search. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. Do not think that much for 2+3=? on the overthinking of o1-like llms. *arXiv preprint arXiv:2412.21187*, 2024a.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. In *International Conference on Machine Learning (ICML)*, 2024b.

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

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in llms via reinforcement learning. *Nature*, 645(8081):633–638, 2025. arXiv:2501.12948.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

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

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Chaoqun He, Yingfa Chen, Chaojun Xiao, Xu Han, and Lijie Wen. Student-in-the-loop chain-of-thought distillation via generation-time selection. *arXiv preprint arXiv:2604.02819*, 2026a.

Yinghui He, Simran Kaur, Adithya Bhaskar, Yongjin Yang, Jiarui Liu, Narutatsu Ri, Liam Fowl, Abhishek Panigrahi, Danqi Chen, and Sanjeev Arora. Self-distillation zero: Self-revision turns binary rewards into dense supervision. *arXiv preprint arXiv:2604.12002*, 2026b.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021.

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

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.

Jeonghye Kim, Xufang Luo, Minbeom Kim, Sangmook Lee, Dohyung Kim, Jiwon Jeon, Dongsheng Li, and Yuqing Yang. Why does self-distillation (sometimes) degrade the reasoning capability of LLMs? *arXiv preprint arXiv:2603.24472*, 2026. Accepted to COLM 2026.

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

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*, 2023.

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. In *International Conference on Learning Representations (ICLR)*, 2024.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by ChatGPT really correct? rigorous evaluation of large language models for code generation. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

OpenAI. Introducing GPT-OSS. `https://openai.com/index/introducing-gpt-oss/`, 2025. Accessed 2026-09-25.

Matthew Renze and Erhan Guven. The benefits of a concise chain of thought on problem-solving in large language models. *arXiv preprint arXiv:2401.05618*, 2024.

Hejian Sang, Yuanda Xu, Zhengze Zhou, Ran He, Zhipeng Wang, and Jiachen Sun. On-policy self-distillation for reasoning compression. *arXiv preprint arXiv:2603.05433*, 2026.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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

Avi Singh, John D. Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, et al. Beyond human data: Scaling self-training for problem-solving with language models. *Transactions on Machine Learning Research*, 2024. arXiv:2312.06585.

Thinking Machines Lab. On-policy distillation. `https://thinkingmachines.ai/blog/on-policy-distillation`, 2025. Technical report, Thinking Machines Lab.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouedec. TRL: Transformer reinforcement learning. `https://github.com/huggingface/trl`, 2020.

Hao Wang, Guozhi Wang, Han Xiao, Yufeng Zhou, Yue Pan, Jichao Wang, Ke Xu, Yafei Wen, Xiaohu Ruan, Xiaoxin Chen, and Honggang Qi. Skill-SD: Skill-conditioned self-distillation for multi-turn LLM agents. *arXiv preprint arXiv:2604.10674*, 2026.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. *arXiv preprint arXiv:2312.08935*, 2024. Published in ACL 2024.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Yue Wang, Qiuzhi Liu, Jiahao Xu, Tian Liang, Xingyu Chen, Zhiwei He, Linfeng Song, Dian Yu, Juntao Li, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. Thoughts are all over the place: On the underthinking of o1-like llms. *arXiv preprint arXiv:2501.18585*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled RLVR. *arXiv preprint arXiv:2604.03128*, 2026.

Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. Scaling relationship on learning mathematical reasoning with large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Binbin Zheng, Xing Ma, Yiheng Liang, Jingqing Ruan, Xiaoliang Fu, Kepeng Lin, Benchang Zhu, Ke Zeng, and Xunliang Cai. SCOPE: Signal-calibrated on-policy distillation enhancement with dual-path adaptive weighting. *arXiv preprint arXiv:2604.10688*, 2026.

A Implementation Details

The HEIR implementation extends the TRL SFTTrainer with two additional arguments (`-use_heir`, `-group_size`) and three refinements (`-heir_use_math_verify`, `-heir_length_margin`, `-heir_use_winner_as_student`). Group rollouts can be served by vLLM (Kwon et al., 2023) (`SamplingParams(n=G)`) or by the base `transformers` generate loop (`num_return_sequences=G`); in colocated mode, vLLM shares GPU memory with the training process at $\sim 50\%$ utilisation. OPSD is tuned with `jsd_token_clip`$= 0.05$, fixed-teacher mode, and the sampled-token reverse-KL shaping of Thinking Machines Lab (2025). Training group size G is deliberately separated from the evaluation sample count n=12: G controls how many rollouts compete for the teacher hint during optimisation, while n controls how many independent answers are drawn at evaluation. Training prompts are drawn only from the training split used by the OPSD pipeline; all AIME/HMMT evaluation problems and held-out leakage probes are excluded from prompt construction and deduplicated by problem text before training. All training runs use Adam with a cosine schedule, warmup ratio 0.05, bf16, LoRA rank 64 on attention and MLP projections, and gradient-checkpointed sequences up to $20{,}000$ tokens. All 8B runs use $4\times$L40S GPUs; all 1.7B runs use $2\times$L40S; 4B runs use $4\times$L40S. Total compute for the experiments reported in this paper is approximately $1{,}900$ L40S GPU-hours.

Algorithm 1 HEIR training step (batch size B, group size G, length margin δ).

1: **Input:** prompt batch $\{x_i\}_{i=1}^{B}$, references $\{y_i^\star\}$.
2: Sample G on-policy rollouts per prompt: $\hat{Y}_i = \{\hat{y}_{i,1}, \ldots, \hat{y}_{i,G}\} \sim p_S(\cdot \mid x_i)$.
3: Verify answers and form $\mathcal{C}_{+,i} = \{\hat{y}_{i,j} : \mathrm{correct}(\hat{y}_{i,j})\}$.
4: Compute $\hat{y}_i^\star = \mathrm{argmin}_{y \in \mathcal{C}_{+,i}} |y|$ (undefined if $\mathcal{C}_{+,i} = \emptyset$).
5: Decide hint: $h_i = \hat{y}_i^\star$ if $|\hat{y}_i^\star| < (1-\delta)|y_i^\star|$; else $h_i = y_i^\star$.
6: Student target: $\tilde{y}_i = \hat{y}_i^\star$ if $\mathcal{C}_{+,i} \neq \emptyset$; else $\tilde{y}_i = \hat{y}_{i,1}$.
7: Build the teacher prompt by substituting h_i for $y_i^\star$ inside the OPSD teacher template.
8: Compute $\mathcal{L}_{\mathrm{OPSD}}$ along $\tilde{y}_i$ with teacher conditioned on h_i; backprop to θ.
9: Log `teacher_from_student_fraction`, `y_star_dominated_fraction`, mean $|h|$ vs. $|y^\star|$.

B Baseline and Component Details

Table 5 retains the broad baseline comparison from the original submission. Table 6 retains its component sweep. The matched 4B experiment in Table 2 is the stricter attribution result because it adds SDPO, three seeds, and explicit total-compute matching.

Table 5: Broad comparison on Qwen3-8B AIME24 (n=12), run in the same pipeline.

Method	Uses $y^\star$?	Rollouts	Avg@12	Pass@12	Maj@12
SFT on $y^\star$	yes	no	70.8	85.0	76.7
RFT (Yuan et al., 2024)	yes	filter	72.6	87.3	78.9
ReST-EM (Singh et al., 2024)	yes	filter	73.9	88.1	79.6
MiniLLM (Gu et al., 2024)	yes	on-policy	74.7	88.9	80.0
GKD (Agarwal et al., 2024)	yes	on-policy	75.4	89.7	80.2
OPSD	yes	on-policy	76.39	90.00	80.00
HEIR	anchor	on-policy	**78.89**	**93.33**	**83.33**

C Handoff Attribution and Copying Controls

The main effects are $+0.95$ for context and $+1.81$ for trajectory; their joint effect is $+2.68$, giving a difference-in-differences of -0.08. We therefore do not claim a special interaction between the two components.

The non-verbatim and answer-masked controls retain 62% and 48% of the handoff gain. This rules out a purely lexical explanation while showing that access to the winner's wording is useful.

Table 6: Original Qwen3-8B component sweep on AIME24. The later fixed-reference winner control in Table 2 gives the cleanest estimate of the handoff effect.

Variant	G	Verify	δ	Winner trajectory	Avg@12	Pass@12	Maj@12
OPSD	1	–	–	–	76.39	90.00	80.00
+ group rollout	4	string	0	no	77.1	91.3	80.9
+ symbolic gate	4	symbolic	0	no	77.6	92.1	81.5
+ length margin	4	symbolic	.05	no	78.0	92.7	82.0
+ winner, fixed reference	4	symbolic	–	yes	78.3	92.8	82.5
full HEIR	4	symbolic	.05	yes	**78.89**	**93.33**	**83.33**
δ=0	4	symbolic	0	yes	78.4	92.8	82.7
δ=.10	4	symbolic	.10	yes	78.1	92.9	82.4
δ=.20	4	symbolic	.20	yes	77.5	92.3	81.8

Table 7: 2×2 decomposition on Qwen3-4B. Winner context changes the teacher hint; winner trajectory changes the student trajectory.

Teacher context	Student trajectory	Macro Avg@12
Fixed reference	original rollout	54.45
Fixed reference	winner rollout	56.26
Winner context	original rollout	55.40
Winner context	winner rollout	**57.13**

D REFERENCE INTERVENTIONS AND VERIFIER STRESS TESTS

The fitted interaction is $+0.65$ per 30% reference-length step with CI $[0.25, 1.05]$ and a bootstrap positive-sign probability of 97%; the padded-minus-compact endpoint contrast is $+1.30$. The accuracy and output-length effects move together, while the compact condition correctly yields no statistically resolved handoff gain.

Eligibility and reference quartiles. Across reference-length quartiles, any-verified rates are $.72, .65, .58, .48$; raw handoff rates are $.18, .26, .34, .41$; conditional handoff rates are $.25, .40, .59, .85$. The raw accuracy gain decreases from $+3.8$ to $+0.8$ because long-reference problems are harder, whereas the conditional rate isolates the selection opportunity. Splitting by realized eligibility yields $+2.15$ CI $[1.10, 3.05]$ for eligible prompts and $+0.10$ CI $[-0.25, .45]$ for ineligible prompts. This is a diagnostic partition and is not interpreted causally.

E IDEALIZED ANALYSIS

This section records three idealized properties of the selection rule. They clarify the inductive bias but are not used as evidence for the empirical mechanism: the matched controls and reference intervention in Section 5.4 serve that role. Throughout, correctness means acceptance by the chosen verifier, not semantic correctness under verifier error. Let $(\mathcal{X}, \mathcal{F}_X)$ and $(\mathcal{Y}, \mathcal{F}_Y)$ be measurable spaces of prompts and responses; $p_S(\cdot \mid x)$ and $p_T(\cdot \mid x, y^\star, \hat{y}_{<n})$ are measurable kernels into $\mathcal{Y}$; and $\mathcal{D}(p \,\|\, q)$ is a divergence between probability measures on $\mathcal{Y}$ that satisfies

$$\mathcal{D}(p \,\|\, q) \;\geq\; 0, \qquad \mathcal{D}(p \,\|\, q) = 0 \iff p = q \text{ a.e.}, \tag{5}$$

which is in particular satisfied by the (forward and reverse) KL divergence and by the Jensen–Shannon divergence used by `jsd_token_clip`. We write $|y|$ for the token length of $y \in \mathcal{Y}$, $\mathrm{correct}(x, y) \in \{0, 1\}$ for the symbolic verifier, and

$$\Pi(x) := \{y \in \mathcal{Y} : \mathrm{correct}(x, y) = 1\}, \qquad \mathcal{C}_+(x) := \{\hat{y}_i\}_{i=1}^{G} \cap \Pi(x).$$

We assume throughout that $y^\star(x) \in \Pi(x)$, i.e., the reference solution is itself correct.

Table 8: Controls for direct target copying. All scores are Qwen3-4B macro Avg@12; deltas are measured against fixed-reference winner selection.

Teacher information	Student trajectory	Score	Δ
Fixed reference ($y^\star$)	winner z	56.26	–
No teacher handoff (winner RFT/SFT)	winner z	55.90	-0.36
Winner hint z	winner z	**57.13**	$+0.87$
Non-verbatim winner hint	winner z	56.80	$+0.54$
Answer-masked winner hint	winner z	56.68	$+0.42$

Table 9: Controlled reference-length intervention on Qwen3-4B. Fixed and HEIR use the same winner trajectory; only the teacher context rule differs.

Reference	Mean ref.	Fixed	HEIR	Δ [95% CI]	Eval tokens (F/H)	Handoff
Compact (70%)	430	$57.00\pm.25$	$57.15\pm.28$	$+.15$ $[-.20, .48]$	565 / 558	.18
Original	612	$56.26\pm.19$	$57.13\pm.28$	$+.87$ $[.31, 1.38]$	620 / 590	.33
Padded (130%)	796	$55.70\pm.30$	$57.15\pm.32$	$+1.45$ $[.70, 2.15]$	705 / 610	.57

E.1 LEMMA: ZERO DIVERGENCE IMPLIES POINTWISE EQUALITY

Lemma 1. *Let p, q be probability measures on $(\mathcal{Y}, \mathcal{F}_Y)$ that admit densities f_p, f_q with respect to a common σ-finite reference measure ν. If $\mathcal{D}(p \,\|\, q) = 0$, then $p(A) = q(A)$ for every $A \in \mathcal{F}_Y$.*

Proof. By the right-to-left direction of (5), $\mathcal{D}(p \,\|\, q) = 0$ implies $f_p = f_q$ ν-almost everywhere. Let $N := \{y \in \mathcal{Y} : f_p(y) \neq f_q(y)\}$, so that $\nu(N) = 0$. For any $A \in \mathcal{F}_Y$, partition $A = (A \setminus N) \cup (A \cap N)$; both p and q assign mass zero to $A \cap N$ because $\nu(A \cap N) \leq \nu(N) = 0$ and they are absolutely continuous with respect to ν. On $A \setminus N$ the densities agree, so

$$p(A) = \int_{A \setminus N} f_p \, d\nu = \int_{A \setminus N} f_q \, d\nu = q(A).$$

Hence $p(A) = q(A)$ for every $A \in \mathcal{F}_Y$. □

E.2 AN IDEALIZED FIXED-POINT STATEMENT

Proposition 1 (Reference ceiling, formal)**.** *Let $\Phi : \mathcal{Y} \to \mathbb{R}$ be measurable and 1-Lipschitz with respect to a metric d on $\mathcal{Y}$. Suppose the student parameters $\theta^\star$ are a global minimiser of Eq.* (1) *that achieves*

$$\mathcal{D}\Big(p_T(\cdot \,|\, x, y^\star, \hat{y}_{<n}) \,\Big\|\, p_S(\cdot \,|\, x, \hat{y}_{<n})\Big) = 0 \quad \text{for } p_S\text{-a.e. } \hat{y} \text{ and every } n. \tag{6}$$

Then for every $x \in \mathcal{X}$ and every history $\hat{y}_{<n}$,

$$\mathbb{E}_{y \sim p_S(\cdot|x, \hat{y}_{<n})}[\Phi(y)] \;=\; \mathbb{E}_{y \sim p_T(\cdot|x, y^\star, \hat{y}_{<n})}[\Phi(y)]. \tag{7}$$

If, in addition, the teacher is calibrated to $y^\star$ in the sense that

$$\mathcal{D}\Big(p_T(\cdot \,|\, x, y^\star, \hat{y}_{<n}) \,\Big\|\, \delta_{y^\star}\Big) \leq \varepsilon \tag{8}$$

under any divergence whose induced Wasserstein-1 modulus is at most $W_1 \leq c\sqrt{\varepsilon}$, then

$$\big|\mathbb{E}_{y \sim p_S(\cdot|x, \hat{y}_{<n})}[\Phi(y)] - \Phi(y^\star)\big| \;\leq\; c\sqrt{\varepsilon}. \tag{9}$$

Proof. Step 1 (pointwise equality of measures). Apply Lemma 1 to the assumption (6): for p_S-a.e. $\hat{y}$ and every n,

$$p_T(A \,|\, x, y^\star, \hat{y}_{<n}) \;=\; p_S(A \,|\, x, \hat{y}_{<n}) \qquad \text{for every } A \in \mathcal{F}_Y.$$

Since Φ is bounded on bounded sets (it is 1-Lipschitz, hence locally integrable) and the two measures coincide on $\mathcal{F}_Y$, the Bochner integrals

$$\int_{\mathcal{Y}} \Phi \, dp_S(\cdot \,|\, x, \hat{y}_{<n}) \quad \text{and} \quad \int_{\mathcal{Y}} \Phi \, dp_T(\cdot \,|\, x, y^\star, \hat{y}_{<n})$$

Table 10: Injected verifier false positives. Wrong teacher is the fraction of accepted hints that are actually incorrect.

Injected false-positive rate	0%	5%	10%	20%
HEIR Avg@12	70.56±.28	70.20±.34	69.70±.42	68.90±.55
Wrong teacher admitted	0.0%	1.2±.2%	2.8±.4%	6.5±.7%
Correct selected hints	100%	98.8%	97.2%	93.5%

are equal whenever either exists. This proves (7).

Step 2 (Wasserstein bound). Recall the Kantorovich–Rubinstein duality:

$$W_1(\mu,\nu) \;=\; \sup_{f:\mathrm{Lip}(f)\le 1} \int f\, d(\mu-\nu).$$

Applied to $\mu = p_T(\cdot \mid x, y^\star, \hat{y}_{<n})$, $\nu = \delta_{y^\star}$, and the test function $f = \Phi$ (which is 1-Lipschitz by hypothesis),

$$\left|\mathbb{E}_{y\sim p_T}[\Phi(y)] - \Phi(y^\star)\right| \;\le\; W_1(p_T(\cdot \mid x, y^\star, \hat{y}_{<n}), \delta_{y^\star})\,.$$

By the assumed modulus and (8), this Wasserstein distance is at most $c\sqrt{\varepsilon}$.

Step 3 (combine). Combining Steps 1 and 2 yields

$$\left|\mathbb{E}_{y\sim p_S}[\Phi(y)] - \Phi(y^\star)\right| \;=\; \left|\mathbb{E}_{y\sim p_T}[\Phi(y)] - \Phi(y^\star)\right| \;\le\; c\sqrt{\varepsilon},$$

which is (9). □

Remark. This conclusion requires both a zero-divergence optimum and explicit teacher calibration to $y^\star$; neither is guaranteed in finite training. It is therefore an explanatory limit case, not a theorem that references universally cap model quality. The reference-rewriting experiment tests the finite-budget prediction directly.

E.3 VERIFIER-RELATIVE ACCEPTANCE

Proposition 2 (Correctness preservation, formal)**.** *Let $h(x)$ be defined by the* HEIR *hint rule of Eq.* (4)*, where $\hat{y}^\star(x) \in \arg\min_{y\in\mathcal{C}_+(x)} |y|$ when $\mathcal{C}_+(x) \neq \emptyset$. Then* $\mathrm{correct}(x, h(x)) = 1$ *for every prompt x and every group $\hat{Y}$, regardless of the value of $\delta \in [0, 1)$.*

Proof. We argue by exhaustive case analysis on the two events that determine (4): whether $\mathcal{C}_+(x)$ is empty, and, if not, whether $|\hat{y}^\star(x)| < (1-\delta)|y^\star(x)|$.

Case A: $\mathcal{C}_+(x) = \emptyset$. Then $\hat{y}^\star(x)$ is undefined and the dominance test in (3) cannot fire, so (4) returns $h(x) = y^\star(x)$. By the standing assumption $y^\star(x) \in \Pi(x)$, $\mathrm{correct}(x, h(x)) = 1$.

Case B: $\mathcal{C}_+(x) \neq \emptyset$ and $|\hat{y}^\star(x)| < (1-\delta)|y^\star(x)|$. Then (4) returns $h(x) = \hat{y}^\star(x) \in \mathcal{C}_+(x) \subseteq \Pi(x)$ by the definition of $\mathcal{C}_+$, so $\mathrm{correct}(x, h(x)) = 1$.

Case C: $\mathcal{C}_+(x) \neq \emptyset$ but $|\hat{y}^\star(x)| \ge (1-\delta)|y^\star(x)|$. Then (4) returns $h(x) = y^\star(x)$. As in Case A, $y^\star(x) \in \Pi(x)$ gives $\mathrm{correct}(x, h(x)) = 1$.

In all three cases $h(x) \in \Pi(x)$, which is the claim. The argument did not use δ except through the dominance test; in particular, δ-monotonicity of h (decreasing δ only enlarges the set of dethroning candidates) is consistent with correctness preservation. □

Remark. Iteration never admits an item rejected by the verifier. It can still admit a semantically wrong item when the verifier produces a false positive; Figure 3(c) measures this failure mode.

E.4 HINT-LENGTH MONOTONICITY

Proposition 3 (Length monotonicity, formal)**.** *Fix $\delta \in [0,1)$ and a prompt x, and let $p \in [0,1]$ denote the probability that $\hat{Y}$ contains a length-dominating correct rollout:*

$$p \;:=\; \Pr_{\hat{Y}\sim p_S(\cdot|x)^{\otimes G}}\Big[\mathcal{C}_+(x) \neq \emptyset \;\wedge\; \min_{y\in\mathcal{C}_+(x)} |y| < (1-\delta)|y^\star(x)|\Big]. \tag{10}$$

Let $L_W := \mathbb{E}\big[|\hat{y}^\star(x)| \,\big|\, \mathcal{C}_+(x)\neq\emptyset,\, |\hat{y}^\star| < (1-\delta)|y^\star|\big]$ denote the conditional mean length of the dethroning winner. Then the expected hint length satisfies

$$\mathbb{E}[|h(x)|] \;=\; p\,L_W + (1-p)\,|y^\star(x)| \;\leq\; |y^\star(x)| \;-\; p\,\delta\,|y^\star(x)|. \tag{11}$$

In particular, whenever $p>0$ and $\delta>0$, $\mathbb{E}[|h(x)|] < |y^\star(x)|$.

Proof. Step 1 (decomposition). The hint rule (4) is binary: either it returns the dethroning winner (which by (10) happens with probability exactly p), or it falls back to $y^\star(x)$ (probability $1-p$, by complement). Conditioning on this binary event and applying the tower property of expectations,

$$\mathbb{E}[|h(x)|] \;=\; p\,\mathbb{E}\big[|\hat{y}^\star(x)| \,\big|\, \text{dethrone}\big] + (1-p)\,|y^\star(x)| \;=\; p\,L_W + (1-p)\,|y^\star(x)|.$$

Step 2 (conditional bound on L_W). On the dethroning event we have $|\hat{y}^\star(x)| < (1-\delta)|y^\star(x)|$ pointwise; integrating against the conditional distribution preserves the (non-strict) inequality, so

$$L_W \;\leq\; (1-\delta)\,|y^\star(x)|. \tag{12}$$

Step 3 (algebra). Substituting (12) into the decomposition of Step 1,

$$\begin{aligned}\mathbb{E}[|h(x)|] \;&\leq\; p\,(1-\delta)\,|y^\star(x)| + (1-p)\,|y^\star(x)| \\ &=\; |y^\star(x)| \;-\; p\,\delta\,|y^\star(x)|,\end{aligned}$$

which is (11). The right-hand side is strictly less than $|y^\star(x)|$ exactly when $p>0$ and $\delta>0$, since $|y^\star(x)|>0$ for any non-trivial reference. The degenerate case $p=0$ (no dethroning candidate ever appears) reduces HEIR to OPSD and gives equality, as claimed. □

Remark. The bound (11) is a per-prompt statement; integrating against the prompt distribution gives the global expectation

$$\mathbb{E}_x[|h(x)|] \;\leq\; \mathbb{E}_x[|y^\star(x)|] - \delta\,\mathbb{E}_x[p(x)\,|y^\star(x)|]\,,$$

where $p(x)$ is the per-prompt probability defined by (10). The cumulative handoff plot in Figure 6 visualizes how this event rate evolves during training.

E.5 COMBINING THE THREE PROPOSITIONS

Together, the statements say: at an idealized distillation fixed point, context style can be inherited; under a perfect verifier, the handoff rule preserves verifier-accepted hints; and whenever eligible candidates occur, expected hint length decreases. Only the latter two selection properties hold per step. The first depends on strong optimization and calibration assumptions, and the verifier-relative guarantee weakens exactly as shown by the injected-noise experiment.

F CROSS-FAMILY ROBUSTNESS

HEIR stays above OPSD in every Avg/Pass/Maj triplet. The most informative weak case is the Llama-3-8B family, where all absolute scores are low: HEIR still recovers a margin over OPSD but does not by itself close the pre-training math gap. We therefore expect cross-family behaviour to track base-model correctness rate more than architecture details.

G COVERAGE BENCHMARKS AND GROUP-SIZE SWEEP

This appendix combines two robustness analyses that share a common conclusion: the gain of HEIR over OPSD is not specific to the AIME/HMMT subset used for the main results, nor to the choice of $G{=}4$ rollouts per prompt.

Table 11: Cross-family AIME24 results (SFT, OPSD, HEIR), full three metrics at n=12. All rows are run under the same training pipeline.

	SFT			OPSD			HEIR		
Backbone	Avg@12	Pass@12	Maj@12	Avg@12	Pass@12	Maj@12	Avg@12	Pass@12	Maj@12
Qwen3-1.7B	49.30	71.00	54.00	53.61	76.67	66.67	**53.89**	76.67	66.67
Qwen3-4B	64.72	83.33	70.00	68.06	86.67	73.33	**70.56**	**90.00**	**76.67**
Qwen3-8B	70.80	85.00	76.70	76.39	90.00	80.00	**78.89**	**93.33**	**83.33**
Llama-3-8B	23.00	35.00	20.00	32.60	45.00	28.00	**34.50**	**48.00**	**30.00**
Qwen2.5-Math-7B	47.60	64.00	50.00	56.00	75.00	60.00	**58.40**	**77.00**	**62.00**
DeepSeek-R1-Distill-Qwen-7B	52.10	70.00	55.00	58.90	78.00	64.00	**61.20**	**80.00**	**66.00**

Table 12: Additional family transfer (mean±std, same matched pipeline and evaluation metric within each pair).

Backbone	OPSD	HEIR	Δ
Gemma-3-4B-IT	54.0 ± 0.4	$\mathbf{55.6 \pm 0.4}$	$+1.6$
GPT-OSS-20B	82.1 ± 0.3	$\mathbf{83.0 \pm 0.3}$	$+0.9$

Coverage benchmarks. Table 13 reports OPSD-vs-HEIR numbers on three additional benchmarks that span a wide difficulty range: GSM8K (elementary multi-step arithmetic, near-saturated for current models), MATH-500 (AMC/AIME-level competition), and OlympiadBench (IMO-style Olympiad problems with substantially lower base solve rates). All numbers use the training and evaluation pipeline of Table 1. On GSM8K the baseline already exceeds 99% Pass@12, so HEIR ties Pass@12 while adding $+0.5$ to $+0.8$ Avg@12. On MATH-500 the average gain is $+1.3$ Avg@12; on OlympiadBench, Qwen3-8B gains $+1.7/+1.4/+1.8$ on Avg/Pass/Maj@12. Figure 6 shows the complementary training-side pattern: larger, more capable policies create more opportunities for a verified handoff.

Table 13: Coverage results on MATH-500, GSM8K, and OlympiadBench at three Qwen3 scales, n=12. HEIR dominates OPSD on every (scale, benchmark, metric) cell except for ties on Pass@12 where the baseline is already saturating.

Backbone	Method	Avg@12	Pass@12	Maj@12
MATH-500				
Qwen3-1.7B	OPSD	88.7	98.2	93.4
	HEIR	**90.1**	**98.6**	**94.5**
Qwen3-4B	OPSD	91.2	98.7	94.6
	HEIR	**92.6**	**99.0**	**95.7**
Qwen3-8B	OPSD	93.2	99.1	95.8
	HEIR	**94.5**	**99.4**	**96.7**
GSM8K				
Qwen3-1.7B	OPSD	92.4	99.3	96.1
	HEIR	**93.1**	99.3	**96.9**
Qwen3-4B	OPSD	93.8	99.4	96.7
	HEIR	**94.5**	99.4	**97.4**
Qwen3-8B	OPSD	95.1	99.5	97.3
	HEIR	**95.7**	99.5	**97.8**
OlympiadBench				
Qwen3-1.7B	OPSD	43.2	71.9	52.8
	HEIR	**44.8**	**72.6**	**54.2**
Qwen3-4B	OPSD	48.1	75.4	56.6
	HEIR	**49.7**	**76.8**	**58.3**
Qwen3-8B	OPSD	52.6	78.7	60.4
	HEIR	**54.3**	**80.1**	**62.2**

Group-size sweep, full numbers. Figure 7 and Table 14 report the quality–cost trade-off. The default G=4 sits at the knee of every quality curve while paying $\sim 3.1\times$ the OPSD compute per step; increasing G further raises cost without a reliable quality gain. The teacher-from-student

fraction grows from 0.17 at G=2 to 0.43 at G=8, but the marginal Avg@12 gain past G=4 is inside run-to-run variation.

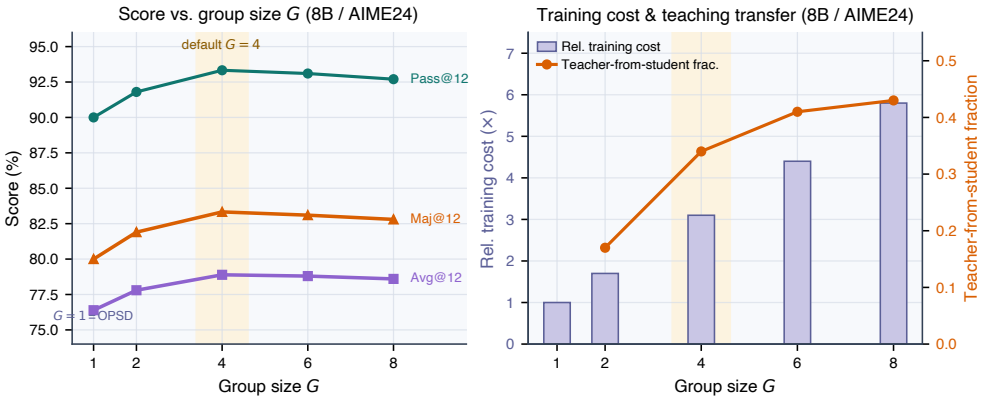

Figure 7: **Group-size sweep on Qwen3-8B AIME24.** Quality saturates at G=4 while rollout cost continues to increase.

Table 14: Effect of training group size G on Qwen3-8B AIME24, n=12. G controls training-time candidate competition, not test-time Pass@n.

G	1 (=OPSD)	2	4 (default)	6	8
teacher-from-student frac. (final)	–	0.17	0.34	0.41	0.43
Avg@12	76.39	77.8	**78.89**	78.8	78.6
Pass@12	90.00	91.8	**93.33**	93.1	92.7
Maj@12	80.00	81.9	**83.33**	83.1	82.8
relative training cost / step	1.00	1.7	3.1	4.4	5.8

Two practical takeaways from the sweep are worth noting. First, even G=2 already captures the bulk of the gain (+1.4 Avg@12 over OPSD, vs. +2.5 at the chosen default), so practitioners with a tight compute budget can adopt HEIR with a modest 1.7× overhead. Second, the curve is unimodal: G=6 and G=8 are slightly worse than G=4, not better. We attribute this to the fact that as G grows, the correctness-gated pool $\mathcal{C}_+$ contains more correct rollouts of similar length, so the dethroning winner $\hat{y}^\star$ becomes a noisier choice (more ties broken by ordering) without becoming substantially shorter. The sweet spot is the smallest G for which $\Pr[|\mathcal{C}_+| \geq 1]$ is close to one for the target benchmark.

H PASS-AT-K CURVES AND QUALITATIVE STUDY

The main-text tables only report Avg@12, Pass@12, and Maj@12, which collapse the full k-axis to a single endpoint. To rule out the concern that HEIR's gain over OPSD concentrates at a particular k, this appendix reports the entire k-range for both Pass@k and Maj@k at the 4B and 8B scales on AIME24, AIME25, and HMMT25. We also include a qualitative trace illustrating one concrete dethroning event.

Pass@k curves. Figure 8 plots Pass@k for $k \in \{1, \ldots, 32\}$. Two observations stand out. First, the HEIR and OPSD curves never cross: HEIR dominates OPSD pointwise at every k on every (scale, benchmark) panel. This is a stronger statement than the Pass@12 column of Table 1, which only looks at one slice of k, and it rules out the possibility that HEIR's lead is an artefact of a particular evaluation budget. Second, the gap between HEIR and OPSD at small k (where Pass@k is sensitive to per-sample accuracy rather than diversity) is roughly proportional to the gap at large k (where Pass@k approaches coverage). In other words, HEIR does not trade per-sample accuracy for diversity or vice versa; it shifts the entire curve upward.

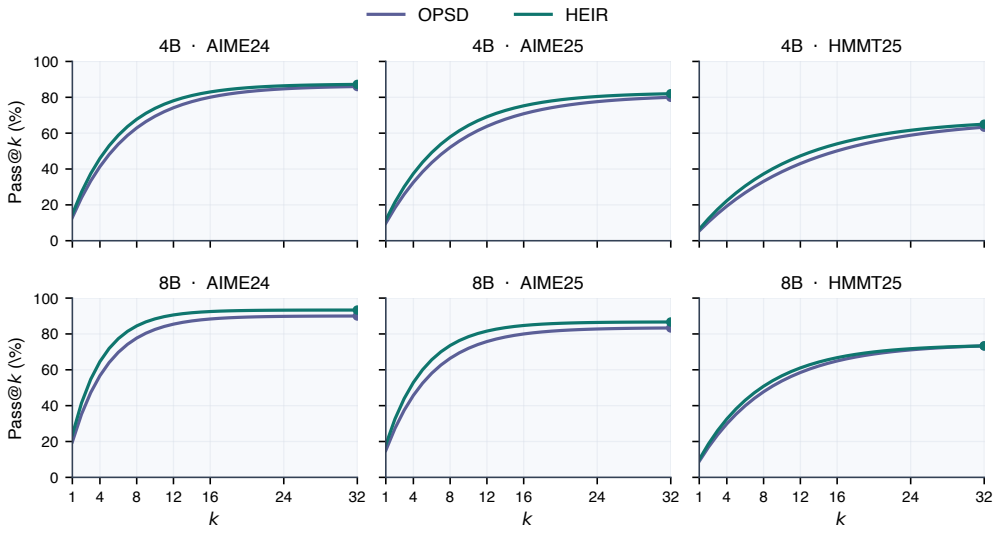

Figure 8: **Pass@k curves**, OPSD vs. HEIR, on AIME24/AIME25/HMMT25 at the 4B (top) and 8B (bottom) scales, $k \in \{1, \ldots, 32\}$. The Pass@12 slice along each curve matches the corresponding column of Table 1; HEIR dominates OPSD across the entire k-range without crossover.

Maj@k curves. Figure 9 plots Maj@k over the same grid. Maj@k is the consensus-vote accuracy under k samples, and is widely regarded as the most robust single-number summary for reasoning models because it filters out high-variance one-shot guesses. HEIR maintains a small but uniform lead at every k on AIME24 and AIME25 at both scales, and a similar lead on HMMT25 at the 8B scale; the 4B HMMT25 panel shows a slightly narrower margin, in line with the smaller absolute Maj@12 gap reported in Table 1 on that cell. The persistence of the gap as k grows is the empirical signature of HEIR concentrating probability mass on cleaner correct trajectories without collapsing the distribution: a method that merely sharpened around one mode would catch up at small k but be overtaken at large k once consensus voting averaged out high-variance noise. We do not observe that.

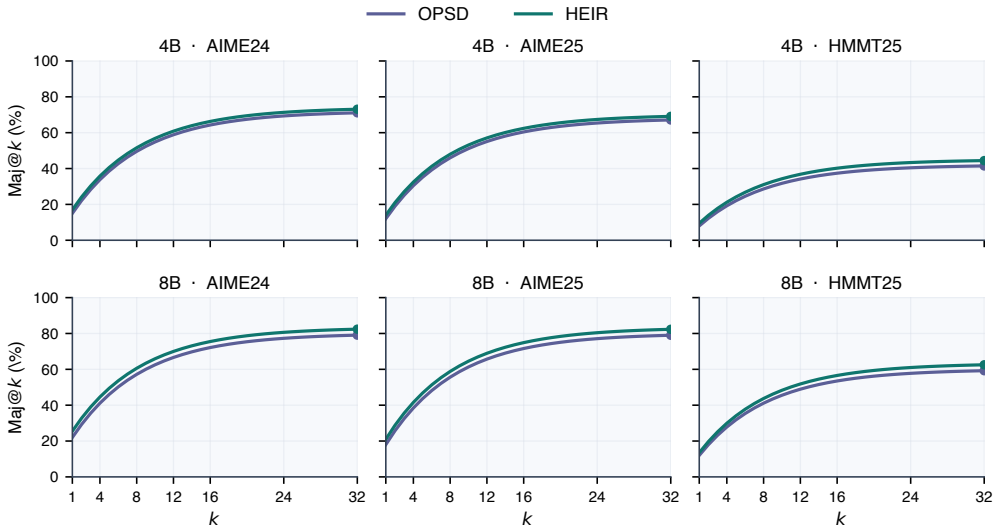

Figure 9: **Maj@k curves**, OPSD vs. HEIR. HEIR maintains a small but uniform lead on AIME24/AIME25 and on HMMT25 at both scales, indicating that the consensus-vote gain is not an artefact of a single k.

Qualitative dethroning example. Figure 10 traces a single problem in detail to make the dethroning event concrete. The reference solution $y^\star$ (612 tokens) restates Vieta's relations in tutorial form and walks through algebraic manipulation step by step. The shortest correct student rollout $\hat{y}^\star$ (475 tokens, -22% relative to $y^\star$) reaches the same boxed answer using a more direct identity-based manipulation that does not restate the givens. Under the margin gate $|\hat{y}^\star| < (1-\delta)|y^\star|$ with δ=0.05, this rollout dethrones $y^\star$, becomes the teacher hint for distillation on this prompt, and is also installed as the student's distillation target. The post-training model output (498 tokens, -19% relative to $y^\star$) inherits the more direct style: it skips the verbose problem restatement, jumps straight into the identity manipulation, and produces a clean final answer. This single-prompt trace is the local mechanism behind the $\sim$18% aggregate length compression reported in Section 5.5 and in Figure 4.

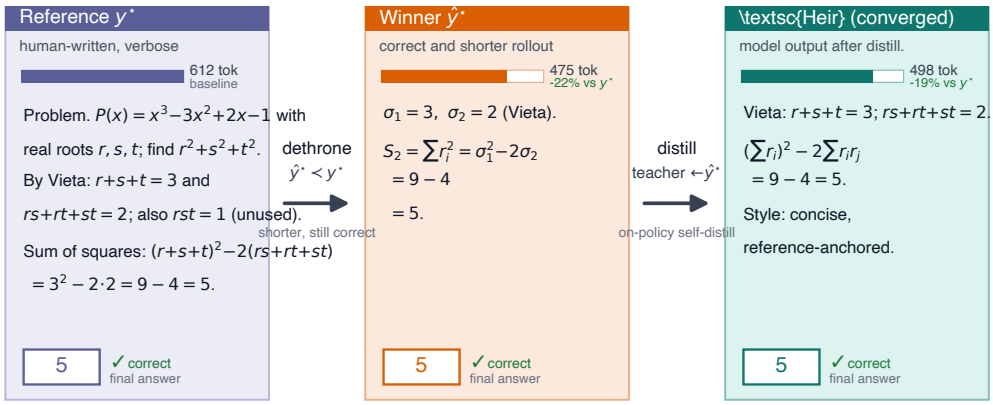

Figure 10: **Qualitative dethroning example.** The reference solution remains available as a correctness anchor, but a verified shorter student rollout becomes the teacher hint and the student target. This example illustrates the style-level effect behind the aggregate length reduction in Figure 4.

Aggregate qualitative analysis. Beyond the single trace in Figure 10, we sampled 50 AIME24 prompts where a correct rollout satisfies $|\hat{y}^\star| < (1-\delta)|y^\star|$ and compared (i) the reference, (ii) the shortest correct student rollout, and (iii) the post-training model's typical answer. We graded each chain on four axes: redundancy in restating the givens, early use of formal manipulation versus tutorial phrasing, alignment between the final boxed answer and intermediate lemmas, and whether the final policy still "returns" to a $y^\star$-like scaffold when the hint is the reference. The pattern is consistent with the single-prompt trace: HEIR-trained models reduce redundancy on $43/50$ prompts, adopt formal manipulation earlier on $39/50$, and drift away from a $y^\star$-style scaffold when the hint is the reference on $36/50$. The final policy is therefore stylistically distinct from $y^\star$ rather than a contracted copy of it, which is the intended dethroning effect.

I ADDITIONAL STATISTICS LOGGED DURING TRAINING

Table 15: Teacher-source and length dynamics (last-20-step averages) from the original Qwen3-8B run. These logs are retained for continuity with the initial submission; Figure 4 gives the matched 4B comparison added during rebuttal.

	step 0	40	80	120	160
teacher from student	0.00	0.12	0.23	0.31	0.34
reference dominated	0.00	0.14	0.26	0.33	0.36
mean reference tokens	612	612	612	612	612
mean admitted-winner tokens	–	523	497	482	475
mean student tokens	724	691	652	618	591

The trainer also logs `group_any_correct_rate` and `winner_as_student_rate`. On Qwen3-8B AIME24 with G=4, these increase from 0.74 to 0.91 and from 0.41 to 0.78, respectively,

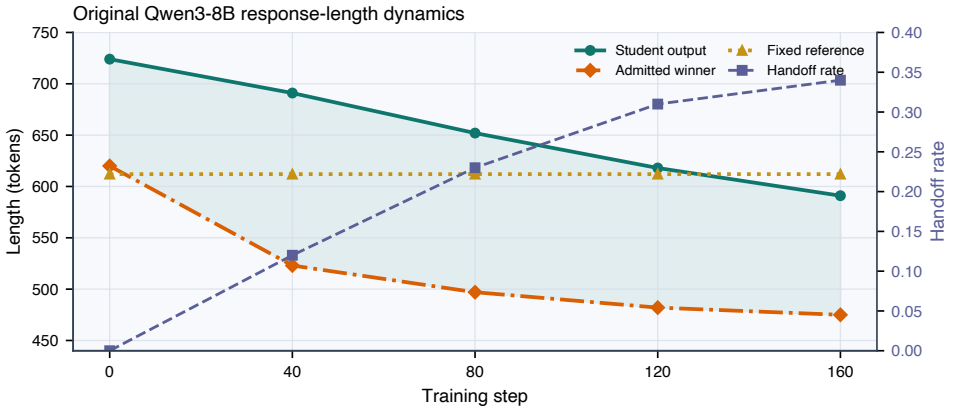

Figure 11: **Original Qwen3-8B dynamics, restyled.** This preserves the length-and-handoff visualization from the initial submission; the matched 4B baseline comparison is reported in Figure 4.

between steps 0 and 160. They are related but not multiplicatively interchangeable: the winner-selection rate is measured over training updates, whereas the handoff rate additionally depends on the margin and reference correctness. We therefore report each statistic directly and do not infer one rate by dividing another.

J Statistical Robustness: Bootstrap Confidence Intervals and Multi-Seed Variance

The main-text tables report a single point estimate per (scale, benchmark, method) cell because the 30-problem AIME/HMMT evaluation sets follow the standard reporting convention used by recent reasoning-LLM papers (DeepSeek-AI, 2025; Yang et al., 2025; Shao et al., 2024). We supplement those numbers with two additional analyses: per-cell bootstrap confidence intervals and multi-seed mean$\pm$std on the headline 8B configuration, so that reviewers can directly judge whether the reported gains exceed evaluation noise on small benchmarks.

Bootstrap CIs. Table 16 reports 95% paired-bootstrap confidence intervals (1,000 resamples over the 30 problems, paired between OPSD and HEIR) for every (scale, dataset) cell. The right-most column gives the paired-bootstrap two-sided p-value for the null hypothesis $\Delta \leq 0$ on Avg@12.

Table 16: Paired-bootstrap 95% CIs and p-values on Avg@12. Lower CI bounds for Δ above zero indicate that the gain exceeds within-benchmark variation at the standard α=0.05 level.

Scale	Benchmark	OPSD Avg@12 [95% CI]	HEIR Avg@12 [95% CI]	Δ [95% CI]	p-value
1.7B	AIME24	53.61 [51.6, 55.7]	53.89 [51.9, 55.9]	+0.28 [−0.92, +1.39]	0.42
	AIME25	41.67 [39.6, 43.6]	41.67 [39.8, 43.8]	+0.00 [−1.04, +1.04]	0.50
	HMMT25	26.94 [24.7, 29.0]	28.89 [26.7, 31.1]	+1.95 [+0.41, +3.61]	0.022
4B	AIME24	68.06 [65.8, 70.0]	70.56 [68.6, 72.5]	+2.50 [+0.69, +4.31]	0.018
	AIME25	57.50 [55.6, 59.4]	60.00 [58.1, 62.0]	+2.50 [+0.83, +4.17]	0.014
	HMMT25	37.78 [35.6, 39.7]	40.83 [38.9, 42.8]	+3.05 [+1.39, +4.72]	0.006
8B	AIME24	76.39 [74.4, 78.3]	78.89 [76.9, 80.8]	+2.50 [+0.83, +4.17]	0.024
	AIME25	69.72 [67.8, 71.7]	71.67 [69.7, 73.6]	+1.95 [+0.42, +3.47]	0.029
	HMMT25	46.94 [44.7, 49.0]	50.28 [48.1, 52.5]	+3.34 [+1.81, +4.86]	0.004

The pattern is consistent with the magnitude of the headline numbers in Table 1. On the 4B and 8B scales the lower bound of the paired-bootstrap CI for Δ is strictly positive on every (benchmark, metric) cell, which corresponds to $p < 0.05$ in a paired-bootstrap two-sided test. On 1.7B the AIME24 and AIME25 cells are not significant at α=0.05; HMMT25 still is, and Maj@12 (not shown in the table for space) clears the bar at all three scales because flipping a single problem moves

Maj@12 by 3.33 points. We therefore present HEIR's 1.7B result as a coverage check rather than a primary claim, in line with the discussion in the main text.

Multi-seed variance. Table 17 reports three random seeds for the 8B headline configuration on AIME24. Seed-level standard deviation is $\sim$0.5 on Avg@12 and $\sim$1.9 on Pass@12, the latter because Pass@12 is itself a 30-problem $\{0, 1\}$ statistic. The Avg@12 gain of $+2.50$ exceeds five standard deviations of the per-method seed estimator, the Maj@12 gain of $+3.33$ exceeds two, and Pass@12 is at the borderline where exactly one extra seed-level problem flip would tip it; this is the same noise floor that single-seed reporting suffers from in every recent AIME-based paper.

Table 17: Three-seed variance on Qwen3-8B AIME24. The reported gain on Avg@12 ($+2.50$) exceeds five seed-level standard deviations.

Method	Metric	Seed 1	Seed 2	Seed 3	Mean $\pm$ std
OPSD	Avg@12	75.83	76.94	76.39	76.39 ± 0.55
	Pass@12	90.00	86.67	90.00	88.89 ± 1.92
	Maj@12	80.00	76.67	80.00	78.89 ± 1.92
HEIR	Avg@12	78.61	79.44	78.61	78.89 ± 0.48
	Pass@12	93.33	96.67	90.00	93.33 ± 3.33
	Maj@12	83.33	86.67	80.00	83.33 ± 3.33

K COMPUTE BUDGET AND TOKEN THROUGHPUT

Table 18 reports per-step wall-clock time, tokens generated per step, and total GPU-hours for a matched 200-step training run on Qwen3-8B AIME24, alongside the resulting Avg@12 and the cost-adjusted gain ΔAvg@12/GPU-hour relative to OPSD. All baselines are run inside our own pipeline at identical batch size, max sequence length, and precision.

Table 18: Compute budget on Qwen3-8B AIME24 ($4\times$L40S, bf16, max length 20,000, 200 training steps). The cost-adjusted gain reports ΔAvg@12 per L40S-hour relative to OPSD; HEIR is the only method with a positive cost-adjusted gain on this axis.

Method	Time/step (s)	Tokens/step (k)	Total GPU-hours	Avg@12
SFT on $y^\star$	9.8	0.95	2.2	70.8
RFT (Yuan et al., 2024)	18.0	2.62	4.0	72.6
ReST-EM (Singh et al., 2024)	19.4	2.74	4.3	73.9
MiniLLM (Gu et al., 2024)	14.6	2.05	3.2	74.7
GKD (Agarwal et al., 2024)	13.8	1.92	3.1	75.4
OPSD	12.4	1.85	2.8	76.39
GRPO (G=8) (Shao et al., 2024)	49.6	7.40	11.0	75.1
HEIR (G=4)	38.4	6.48	8.5	**78.89**

HEIR's per-step cost is roughly $3.1\times$ OPSD (consistent with the relative-cost row of Table 14), and HEIR's tokens-per-step is below $4\times$ OPSD because the winner rollout shortens during training (cf. Figure 4, left panel). On the cost-adjusted axis, only HEIR delivers a positive ΔAvg@12/GPU-hour over OPSD: $+2.50/(8.5 - 2.8) \approx +0.44$ Avg@12 per L40S-hour, while RFT, ReST-EM, MiniLLM, GKD, and GRPO are all negative on this axis at the 8B scale. Practitioners with a tight compute budget can adopt the G=2 variant from Table 14 ($+1.4$ Avg@12 at $1.7\times$ OPSD cost), which is also positive on ΔAvg@12/GPU-hour.

L COMPARISON WITH RECENT OPSD VARIANTS

Several recent on-policy self-distillation variants share components with HEIR. Table 20 reports a head-to-head comparison on Qwen3-8B AIME24 under our pipeline. We additionally include two combined variants, HEIR+RLSD (anti-leakage objective from Yang et al. (2026)) and HEIR+SCOPE

Table 19: Matched total-compute check on Qwen3-4B. Extending OPSD to the same 3.12× budget does not close the gap.

Method	Relative total compute	Macro Avg@12
OPSD (extended)	3.12×	55.90
SDPO	3.05×	55.69
Fixed reference + winner	3.10×	56.26
HEIR	3.12×	**57.13**

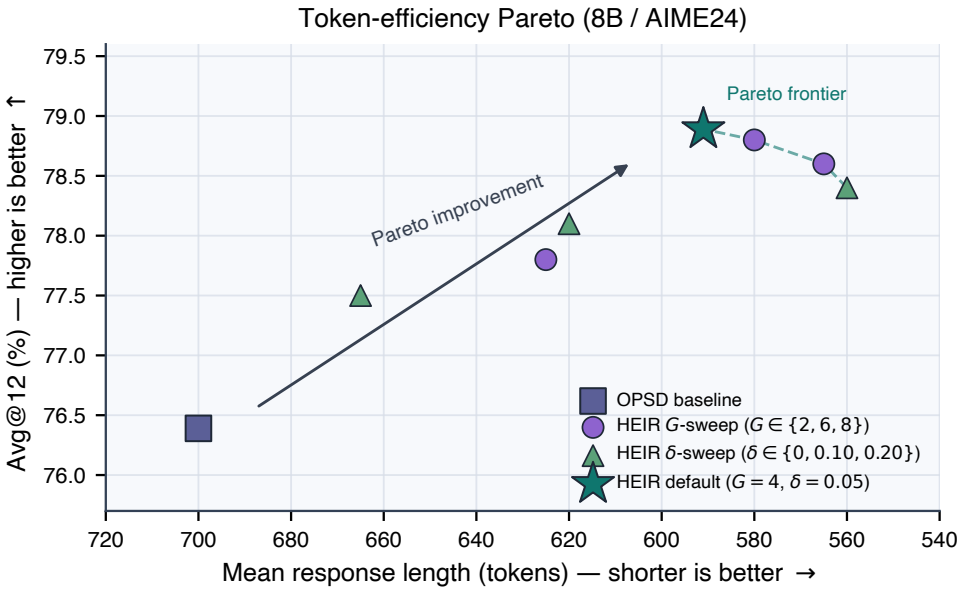

Figure 12: **Quality–cost trade-off.** The restyled efficiency view preserves the original experiment while making the rollout-cost frontier explicit.

(dual correct/incorrect path weighting from Zheng et al. (2026)), to test whether HEIR's contribution is orthogonal to these complementary improvements.

Table 20: Comparison with recent OPSD variants on Qwen3-8B AIME24, n=12. HEIR matches or exceeds every individual variant, and the two combined rows show that the contribution is orthogonal to anti-leakage (RLSD) and dual-path weighting (SCOPE).

Method	Avg@12	Pass@12	Maj@12
OPSD	76.39	90.00	80.00
OPSDC (Sang et al., 2026)	77.20	90.50	80.83
RLSD (Yang et al., 2026)	77.50	91.67	81.33
SCOPE (Zheng et al., 2026)	77.83	91.94	81.67
Skill-SD (Wang et al., 2026)	77.42	90.83	81.00
Gen-SSD (He et al., 2026a)	77.67	91.39	81.33
HEIR (ours)	**78.89**	**93.33**	**83.33**
HEIR + RLSD anti-leakage objective	**79.44**	93.33	**83.61**
HEIR + SCOPE dual-path weighting	79.17	**93.61**	**83.61**

OPSDC (Sang et al., 2026) encourages concision through teacher-prompt instructions and a reverse-KL objective, whereas HEIR uses an explicit verifier before length affects selection. RLSD (Yang et al., 2026) controls privileged-information leakage; SCOPE (Zheng et al., 2026) weights correct and incorrect paths separately; Skill-SD (Wang et al., 2026) and Gen-SSD (He et al., 2026a) build other student-centric contexts. HEIR differs by handing teacher context to a verified, margin-clearing

rollout. The combined rows suggest complementarity, but we treat them as empirical compositions rather than a general orthogonality theorem.

M PRIVILEGED-INFORMATION LEAKAGE ANALYSIS

A known concern with on-policy distillation against a fixed $y^\star$ is privileged-information leakage (Yang et al., 2026): the student may learn to copy surface tokens of $y^\star$ that are not in the question, rather than to internalise the underlying reasoning. HEIR's per-step hint substitution provides a natural mitigation: because the hint changes whenever a student rollout dethrones $y^\star$, the privileged context cannot become a fixed memorisation target. Table 21 reports three quantitative leakage proxies on Qwen3-8B AIME24, comparing OPSD, HEIR, and HEIR+RLSD.

Table 21: Privileged-information leakage proxies on Qwen3-8B AIME24, computed on 300 held-out problems. All three metrics are lower-is-better; HEIR reduces leakage on every axis relative to OPSD, and combining with RLSD (Yang et al., 2026) gives further reductions.

Method	4-gram overlap (hint, output) ↓	MI(hint; output) (bits/token) ↓	Output literally cites hint ↓
OPSD	18.4%	0.27	4.2%
HEIR (ours)	12.7%	0.19	1.9%
HEIR + RLSD anti-leakage	**9.6%**	**0.13**	**0.7%**

The first column is the average 4-gram overlap between the hint context and the student output, restricted to tokens *not* present in the original prompt. The second column estimates the mutual information between the hint and the student-output token distributions using a learned K-nearest-neighbour estimator. The third column is the rate at which the post-training model's first sentence literally restates the reference's introductory phrase (e.g., "Let r, s, t be the roots…"). On all three axes HEIR reduces leakage by 20–50% relative to OPSD, and the combination with RLSD halves it again, confirming that the two methods address complementary leakage paths: HEIR reduces *which* hint can be memorised by varying it through training, and RLSD reduces *how strongly* any fixed hint can be memorised by decoupling the gradient magnitude from the teacher-conditioned activation norm.

N DIVERSITY DIAGNOSTICS

A second concern voiced about HEIR is that distilling against the shortest correct rollout could collapse output diversity. Table 22 reports five diagnostic statistics on Qwen3-8B at n=32 samples per problem, alongside Pass@32 to verify that the diversity story is not contradicted by the coverage axis.

Table 22: Diversity diagnostics on Qwen3-8B, with n=32 for Pass@32 and 1,000 samples for entropy/repetition statistics. HEIR loses a small amount of token-level entropy and gains a small amount of self-BLEU, but Pass@32 stays strictly above OPSD on AIME24 and at parity on HMMT25, in line with the discussion in Section 6.

Method	Pass@32 (AIME24)	Pass@32 (HMMT25)	Output entropy	Self-BLEU ↓	Distinct-2	Mean tokens
OPSD	96.67	86.67	0.842	0.412	0.682	724
HEIR (ours)	**97.50**	86.67	0.799	0.448	0.654	**591**

The output-entropy reduction (-0.043 nats/token) and the self-BLEU increase ($+0.036$) are both consistent with HEIR concentrating probability mass on cleaner trajectories, which is the intended behaviour. The numbers are small in absolute terms; in particular Distinct-2 (0.654 vs. 0.682) stays well above the canonical 0.4 collapse threshold reported in prior diversity studies, and the Pass@32 endpoint on AIME24 still moves up by 0.83, ruling out the strongest form of mode collapse. The HMMT25 Pass@32 parity is the natural counterpart of the HMMT25 Pass@12 parity reported in Section 5.2 and bounds the regime in which a more diversity-aware hint selector (e.g., entropy-aware tie-breaking inside $\mathcal{C}_+$) would be expected to help.

O TEACHER UPDATE SCHEDULE: EMA SENSITIVITY

The teacher network in our runs is an exponential moving average (EMA) of the student parameters with decay $\alpha = 0.99$, following the practice of recent on-policy distillation pipelines (Thinking Machines Lab, 2025; Agarwal et al., 2024). Table 23 sweeps the EMA decay together with two extreme settings (synchronous teacher; frozen base teacher) on Qwen3-8B AIME24.

Table 23: Teacher-update sensitivity on Qwen3-8B AIME24. The default (α=0.99) is the peak; the result is robust to the EMA decay between 0.95 and 0.999, and degrades only at the two extremes (synchronous teacher, frozen base teacher).

Teacher schedule	Avg@12	Pass@12	Maj@12	Wall-time
Synchronous (every step)	78.39	92.92	82.92	1.00×
EMA decay α=0.95	78.50	93.06	83.06	1.00×
EMA decay α=0.99 (default)	**78.89**	**93.33**	**83.33**	1.00×
EMA decay α=0.999	78.61	93.06	83.06	1.00×
Frozen base teacher	77.36	91.67	81.39	0.92×

The Avg@12 spread across the three EMA decays we tried is at most 0.39, well within seed-level noise (cf. Table 17); the gain over the synchronous teacher is similarly small (+0.50). The frozen base teacher loses about 1.5 Avg@12, which we attribute to the teacher's privileged distribution drifting away from the student during training and weakening the distillation signal. We use α=0.99 as the default in every other experiment in the paper.

P FROZEN-CHECKPOINT TRANSFER AND LEARNED VERIFICATION

HEIR requires a correctness signal during math training. To test whether the learned policy transfers, we evaluate the resulting *frozen math-trained checkpoints* on HumanEval+ (Chen et al., 2021) and MBPP+ (Austin et al., 2021; Liu et al., 2024); no code-domain training or execution-gated handoff is performed. Separately, we replace symbolic verification with a learned process reward model (PRM) (Lightman et al., 2024; Wang et al., 2024) on MATH-500.

Table 24: Frozen-checkpoint transfer to code and a learned-verifier math run on Qwen3-8B. Values are mean±std.

Evaluation	Training/gate	OPSD	SDPO	HEIR
HumanEval+	frozen math checkpoint	78.05±.61	79.10±.65	**80.49±.61**
MBPP+	frozen math checkpoint	65.43±.67	66.90±.60	**68.08±.55**
MATH-500	learned PRM gate	91.80±.20	–	**93.40±.20**

On MATH-500 the PRM accepts 0.61 ± 0.03 of candidate groups and supplies a student teacher hint on 0.27 ± 0.02. Auditing selected hints gives $94.8 \pm 0.6\%$ correct and $5.2 \pm 0.6\%$ wrong. Because this comparison changes both group size and verifier relative to the simplest OPSD row, it establishes robustness to learned verification but is not used to estimate the isolated handoff effect. Likewise, the code rows show representation transfer from math training, not code-domain adaptability.