SCHEDULE THE OPTIMIZER, NOT THE TEACHER: DIAGNOSING AND MITIGATING LONG-HORIZON COLLAPSE IN ON-POLICY SELF-DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy self-distillation (OPSD) lets a single language model teach itself. A frozen copy of the initial policy, conditioned on privileged ground-truth solutions, provides dense per-token supervision on the student's own rollouts. This delivers gains competitive with group relative policy optimization (GRPO) at a fraction of the sampling cost. We ask what the short budget hides. Training every frozen-teacher OPSD variant past 100 steps, we find a consistent trajectory: performance peaks, declines monotonically, and by step 300 falls *below the untrained base model*, at 1.7B, 4B, and 8B scales. The obvious suspect is the frozen teacher going stale, but the obvious remedy backfires. Refreshing the teacher to track the student collapses the model to a three-benchmark mean of 5.7 (8.9% on AIME24, vs. base 54.7%); its outputs degenerate into fluent but empty self-talk. Gentler anchor motion is no escape. An exponential moving average (EMA) teacher matches the best early score of the frozen-teacher family, yet ends the horizon below the frozen teacher's own endpoint at every rate we test (32.5–32.9 vs. 35.1 across EMA 0.99–0.9995). Only dose decay absorbs the penalty. The frozen teacher is not the bottleneck but the anchor. The collapse is instead *objective drift* under a constant learning rate, controlled primarily by the cumulative learning-rate dose. The one-sided clip that stabilizes early training explains the drift's visible signature, namely a training loss that descends below zero. It does not explain the drift itself. A horizon-aware cosine decay removes most of the collapse at all three scales, ending at or above the untrained base at 1.7B and 8B. Alongside the diagnosis, we make three supporting contributions. A difficulty-aware sample weighting brings small multi-seed gains and does not alter the long-horizon outcome. A numerically hardened hybrid of self-distillation and reinforcement learning with verifiable rewards trains stably where the naive form suffers a 10^{14} gradient explosion that poisons the optimizer state. A defect analysis of the official implementation clarifies which reported numbers are affected. Our results reframe frozen-teacher self-distillation as a short-horizon method: schedule the optimizer, not the teacher. Our code is available at the anonymous repository: `https://anonymous.4open.science/r/Schedule-the-Optimizer-Not-the-Teacher-ICLR-2027-DCB6`.

1 INTRODUCTION

Post-training on reasoning traces, whether by reinforcement learning (RL) or by self-distillation, now turns base language models into capable reasoners (Zelikman et al., 2022; Guo et al., 2025). This paper asks what happens when such a reasoner is trained *past* the point at which it has learned everything that its frozen teacher could possibly teach it.

The question matters because on-policy distillation (Agarwal et al., 2024; Gu et al., 2024; Ko et al., 2024) and its self-distilled variant OPSD (Zhao et al., 2026) are short-horizon methods by design. One shared frozen base model plays both roles (Figure 1, left). The student, a trainable low-rank adaptation (LoRA) adapter, sees only the problem x_i, while the teacher, the same base with the initial adapter frozen, scores that rollout under the reference y_i^*. Within roughly one hundred optimization steps this delivers gains matching group relative policy optimization (GRPO) at a fraction of the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

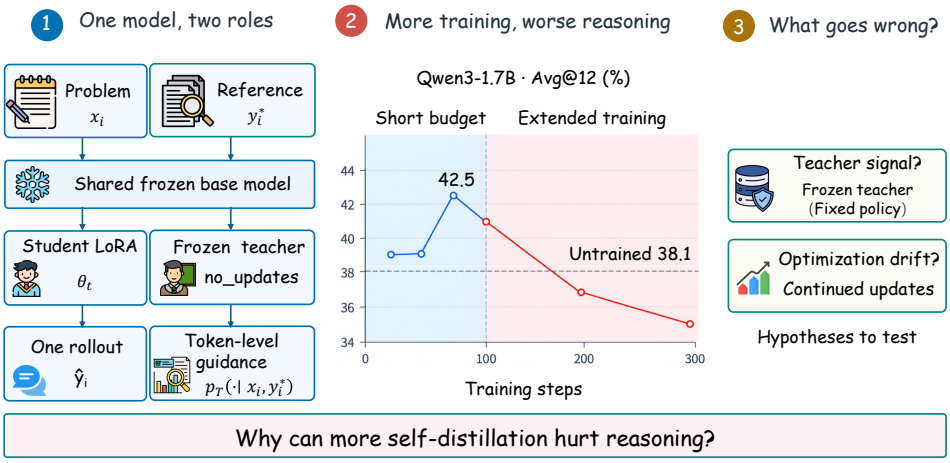


Figure 1: **The research problem in on-policy self-distillation.** *Left:* one shared frozen base: a trainable LoRA student and a frozen-adapter teacher scoring the rollout $\hat{y}_i$ under the reference y_i^*. *Middle:* Qwen3-1.7B baseline mean benchmark score peaks at 42.5 (step 75), then falls to 35.1 by step 300, below untrained 38.1. *Right:* two candidate explanations must be tested first.

sampling cost. Yet every published result stops at that same hundred-step mark, without a single exception.

We extended the training. The resulting curve (Figure 1) is not what the published results would suggest. The Qwen3-1.7B student rises from 38.1 to a peak of 42.5 at step 75, then declines to 35.1 by step 300, *below the model it started from*. The same shape appears for every variant and at 4B/8B scale (largest drop at 4B). In hindsight the warning was already in the original paper: its ablation shows AIME25 declining from 43.9 to 41.1 over steps 50–100. Inside the budget this reads as noise; on the longer horizon it is the first visible onset of the collapse.

Each of the three supervision families behind reasoning post-training has a documented failure mode, and each remedy inherits a blind spot. Reinforcement learning with verifiable rewards (RLVR) suffers reward sparsity (Yu et al., 2025), supervised fine-tuning (SFT) and off-policy distillation reintroduce train–inference mismatch (Agarwal et al., 2024), and on-policy distillation (Agarwal et al., 2024; Gu et al., 2024; Ko et al., 2024) resolves the mismatch at the price of a separate and usually larger teacher. OPSD removes that teacher by freezing a copy of the student, an elegant move that has never been stress-tested. *Every* reported result stops within a few hundred steps. The unexamined assumption is that supervision which helps there remains helpful at every later horizon.

We therefore take a diagnosis-first route and ask *where* the failure lives. The natural suspect is the frozen teacher: the student keeps improving while the teacher stays fixed, so its signal plausibly goes stale. The standard remedy is to let the teacher track the student (Tarvainen & Valpola, 2017; Grill et al., 2020; Furlanello et al., 2018). We tried the strongest version: resetting the teacher to the current student every 100 steps. The result is the worst 1.7B model in our study (three-benchmark mean 5.7; 8.9% on AIME24 vs. base 54.7%); its generations are fluent but *empty* (Appendix A). A fast exponential moving average (EMA) teacher (decay 0.99), the gentler version, matches our best early checkpoint (42.6, within the seed band). Yet by step 300 every anchor-mobility variant we carried to the horizon ends below the frozen teacher's 35.1. Both refresh pathologies replicate at 4B (41.2 at step 200 $\rightarrow$ 3.9 at step 300). The teacher is the *anchor*: it is the only term that ties the student to the initial policy. It is also the one ingredient that the collapse cannot possibly do without.

The real culprit is the optimizer. Under a constant learning rate (LR), the official clipped-divergence objective keeps pushing the student after the learnable signal is exhausted: once close to the teacher, the objective can still be reduced below zero by moving *away* from the teacher. That negative loss is the drift's visible signature, and the cumulative dose (the total learning rate spent over training) is its primary controlling variable. Decaying the learning rate to zero over the same horizon removes most of the collapse at all three scales (39.3/57.7/64.1 vs. 35.1/51.7/58.8 at 1.7B/4B/8B). Our central message is the opposite of that intuition: *The frozen teacher is not a stale signal source to be engineered away, but the anchor that keeps the student near the initial policy. The long-horizon*

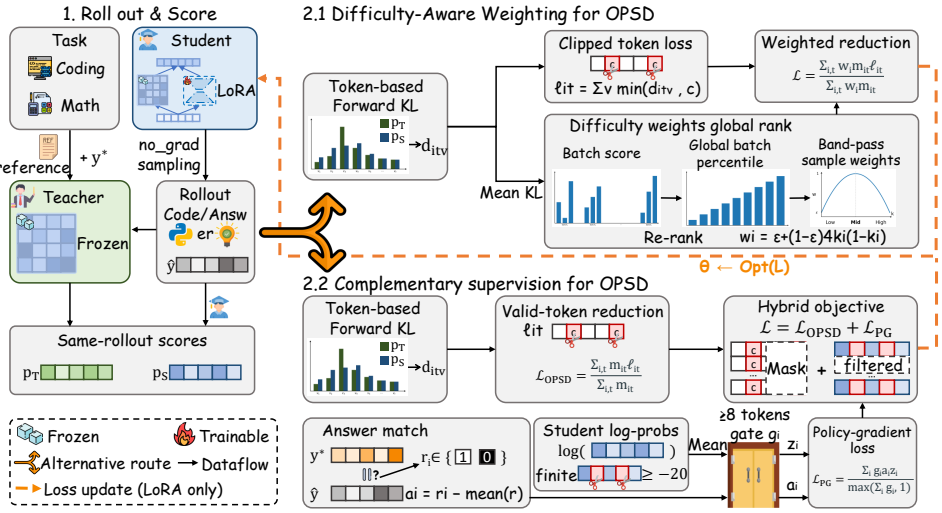


Figure 2: **Two complementary extensions of frozen-teacher OPSD on one shared rollout.** Upper route: difficulty-aware weighting; lower route: OPSD×RLVR hybrid (both Section 7; formal definitions in the text).

failure is objective drift under a constant learning rate, and the principled fix is to schedule the optimizer, not the teacher.

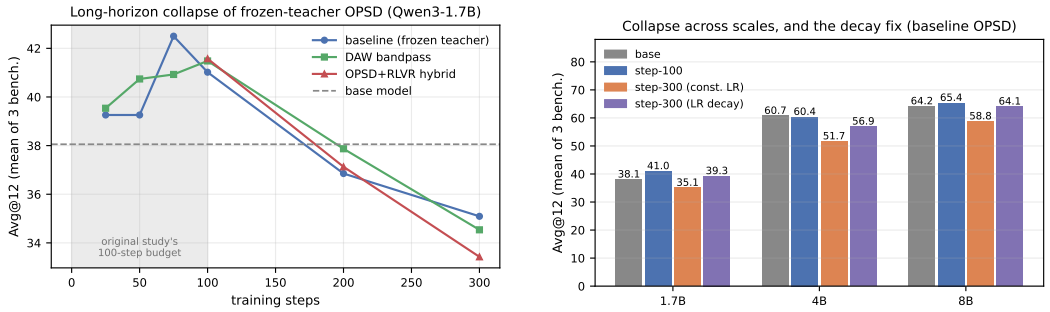
Beyond this diagnosis, the paper makes three supporting contributions. **(i)** A *difficulty-aware weighting* (DAW) scheme (upper route of Figure 2) reweighting samples by their teacher–student divergence percentile within the global batch: small multi-seed gains; the long-horizon collapse is untouched (Section 7). **(ii)** An *RLVR-hybrid objective* (lower route of Figure 2): the naive version explodes with a $\sim 10^{14}$ gradient norm that poisons the Adam second moment even when clipped. A hardened version trains stably (not better than baseline on three-seed means), a case study in why numerics are a research variable here (Section 7). **(iii)** A careful *defect analysis* of the official implementation (Section 3), covering bf16 divergence computation, a train/eval prompt mismatch, and a memory-inefficient loss reduction that we fix, clarifying which reported numbers are affected.

We begin with a faithful reproduction of the official implementation and its defects (Section 3). We then present the collapse itself (Section 4) and the experiments discriminating its cause (Section 5), and close with mitigations (Sections 6, 7). Ablations and sweeps are in Appendix B.

2 Related Work

Distillation: from off-policy to self-distillation. Knowledge distillation began as off-policy compression of a fixed teacher's outputs (Hinton et al., 2015). On-policy distillation repairs the resulting train–inference mismatch by scoring the student's own samples under a fixed teacher (Agarwal et al., 2024; Gu et al., 2024; Ko et al., 2024), a recipe that recently matched RL at a fraction of the sampling cost (Thinking Machines Lab, 2025). On-policy self-distillation removes the external teacher altogether: a frozen copy of the initial policy, primed with reference solutions, grades the student's own rollouts (Zhao et al., 2026; Jang et al., 2026). Yet these methods share one unexamined premise, that a fixed supervisor keeps helping indefinitely, and every published evaluation stops within a few hundred steps. We show the design breaks exactly there, and that it becomes removable once its controlling variable has been identified.

Moving-teacher remedies. A complementary tradition keeps the supervisor in motion: born-again and snapshot distillation, Mean Teacher and momentum encoders, large language model (LLM) self-training and context distillation (Furlanello et al., 2018; Tarvainen & Valpola, 2017; Grill et al., 2020; He et al., 2020; Yang et al., 2019; Kim et al., 2021; Shen et al., 2022; Zelikman et al., 2022; Gulcehre et al., 2023; Snell et al., 2022). Built for semi-supervised consistency or iterative self-training, they have never been stress-tested as OPSD anchors. Transplanted into OPSD,



(a) All variants peak within 100 steps and end *below base* (dashed) by step 300; the published 100-step budget (gray zone) stops where the curves turn.

(b) Base, within-budget peak, and step-300 at 1.7B/4B/8B: constant LR falls below base at every scale; cosine decay removes most of it.

Figure 3: **The long-horizon collapse and its scale dependence.** (a) Mean Avg@12 over AIME24/25 and HMMT25 vs. training steps (Qwen3-1.7B) for the frozen-teacher baseline, DAW, and the RLVR-hybrid. (b) The same failure at three scales and its optimizer-side fix.

they fail informatively. A refreshed teacher is the student on its own distribution, degenerating into self-reinforcement with nothing to anchor it. An EMA teacher tracks the drift it should have stopped. The anchor matters, not its motion.

Optimization-side controls and concurrent work. RLVR's documented failures, namely entropy collapse, zero-variance groups, and length bias (Shao et al., 2024; Guo et al., 2025; Yu et al., 2025; Zheng et al., 2025), differ from ours. Annealing and weight averaging are classical (Loshchilov & Hutter, 2017; Izmailov et al., 2018), and Kullback–Leibler (KL) penalties anchor reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022), but neither has been tied to objective drift in self-distillation. Sample reweighting by learnability (Bengio et al., 2009; Kumar et al., 2010; Shrivastava et al., 2016; Paul et al., 2021) acts on loss magnitude, whereas our DAW acts on the batch-relative teacher–student divergence percentile. Concurrent 2026 analyses document on-policy distillation (OPD) failure phenomenology within the published budget (Li et al., 2026; Song & Zheng, 2026; Kim et al., 2026; Ichihara et al., 2026), and method-side variants extend the objective inside it (Yang et al., 2026; Liu et al., 2026; Shen et al., 2026). Our axis, the optimization horizon itself, is orthogonal to both, and it is where the transferable fix lives (Appendix C gives the full version of this discussion).

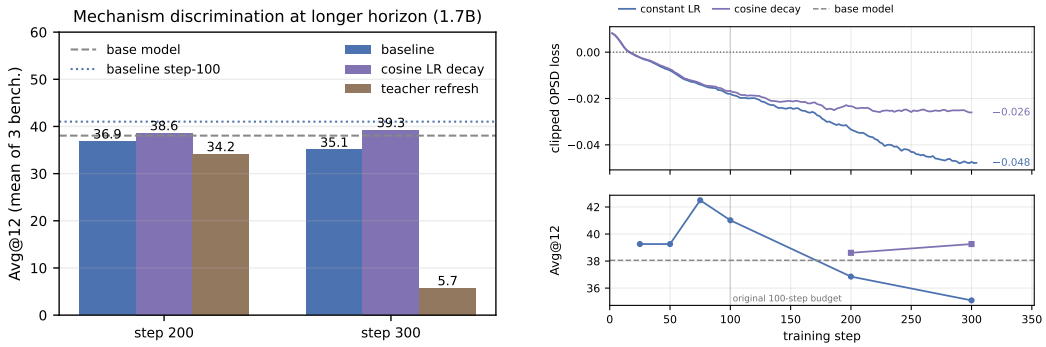
3 Background, Reproduction, and Defects

On-policy self-distillation. Given a dataset $\mathcal{S} = \{(x, y^*)\}$ of problems with verified reference solutions, OPSD instantiates two policies from one LLM: a student $p_S(\cdot \mid x)$ and a teacher $p_T(\cdot \mid x, y^*) \triangleq p_S(\cdot \mid x, y^*)$ at initialization. The student samples a rollout $\hat{y} \sim p_S(\cdot \mid x)$. Both policies score it token by token, and the student is updated to minimize the clipped per-token forward KL,

$$\mathcal{L}_{\mathrm{OPSD}}(\theta) = \mathbb{E}_{(x,y^*)\sim\mathcal{S}}\, \mathbb{E}_{\hat{y}\sim p_S}\Big[\frac{1}{|\hat{y}|}\sum_{n=1}^{|\hat{y}|}\sum_{v\in\mathcal{V}} \min\Big(t_{v,n}\big[\log t_{v,n} - \log s_{v,n}\big],\ \lambda\Big)\Big], \tag{1}$$

where $t_{v,n}$ and $s_{v,n}$ are the teacher's and student's token probabilities. The loss is the forward KL from temperature-scaled logits (τ=1.1), with each vocabulary item's contribution capped pointwise at λ=0.05 (not the token-summed KL), averaged over valid completion tokens. The teacher is frozen at the initial policy: gradients flow only through the student, and teacher/student share one base model. The teacher is the base weights with the student's LoRA adapter disabled.

Our platform. We reproduce the official codebase and extend it with configuration-gated switches, none of which alters default behavior. The switches add difficulty-aware weighting and an RLVR-hybrid term (the two routes of Figure 2), periodic teacher refresh, and a memory-efficient but mathematically identical loss reduction. Unless stated otherwise, every run uses the official hyperparameters and evaluation protocol exactly (12 samples per problem, Avg@12; "mean Avg@12"



(a) Mean Avg@12 at steps 200/300 (1.7B): cosine LR decay holds its gains above base (dashed); teacher refresh collapses (5.7).

(b) Clipped OPSD loss (top) and score (bottom): past $\sim$150 the annealed loss plateaus at -0.026 while the constant-LR loss reaches -0.048 as the score falls.

Figure 4: **Mechanism discrimination at longer horizon (1.7B).** (a) Annealing the optimizer arrests the collapse; moving the teacher does not. (b) The drift is visible in the training loss, which keeps descending below zero under constant LR while the benchmark score falls: the loss curve alone looks like steady improvement throughout training.

is the macro-average over AIME24, AIME25, and HMMT25). Full hardware/software pins are in Appendix D.

Reproduction. The platform reproduces the paper's central result class: step-100 training lifts the 1.7B mean above its base (41.0 vs. 38.1), within 1.5 points of the published mean (42.5; Table 5). Our three-seed replication puts the spread at ± 1.5 points (Figure 10). Throughout, gains are computed against *our own* base model; Appendix D gives the calibration details.

Defects in the official implementation. Because our later claims depend on implementation details, we audited the official code (Table 7, Appendix F). Two defects matter most: the KL/JSD (Jensen–Shannon divergence) loss is computed entirely in bf16, and training and evaluation use *different* prompt distributions. Neither drives the collapse (Appendix F.3). fp32 recomputation moves step 300 by $+0.3$ $[-2.1, +2.8]$, matched thinking-mode prompts still decline from 33.1 to 26.8, and alternative divergences end at the collapsed forward-KL level (reverse KL 36.0, JSD 35.3 vs. 35.1).

4 The Long-Horizon Collapse

Figure 3a shows the central phenomenon. Extended past the reported budget, all three frozen-teacher variants follow the same trajectory: rapid improvement, a peak between steps 75 and 100, a monotone decline, and terminal performance *below the untrained base model* at step 300 (all three variants vs. base 38.1; Table 1). Figure 3b shows the same failure across scales (step-300 drops of $-5.9/-8.7/-6.6$ at 1.7B/4B/8B), so it is not an artifact of one objective or one size. A paired bootstrap over the 90 problems confirms the deficits are not single-seed noise (Appendix G). Stratifying problems by base-model accuracy locates the damage in the mid-difficulty band (Appendix H), and the direction persists on MATH500's 500 problems at all three scales. Three properties make this failure mode worth isolating. It is *universal* across objective variants and scales, *monotone* rather than oscillatory, and *hidden* from the training dashboard. Within the reported budget every metric looks healthy, while the clipped training loss keeps descending even as the benchmark score falls (Figure 4b). The original paper's own ablation (its Table 3: AIME25 $43.9 \rightarrow 41.1$ between steps 50 and 100) is the first visible onset of this curve.

5 Mechanism Discrimination

Two hypotheses can explain the collapse. **H1 (teacher staleness):** the frozen teacher's conditional distribution becomes outdated as the student improves, so its supervision turns misleading. **H2**

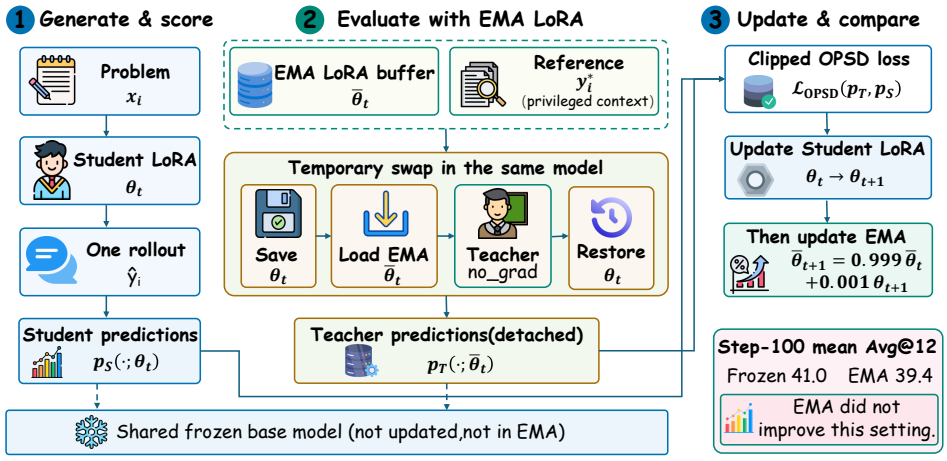


Figure 5: **The EMA-teacher variant we falsified.** The loop runs in three stages on one shared model: the student samples and scores a rollout; a temporary weight swap runs the EMA teacher's forward under the privileged reference without gradients; the student updates, then the EMA buffer follows ($\bar{\theta}_{t+1} = 0.999\,\bar{\theta}_t + 0.001\,\theta_{t+1}$). The fastest variant (decay 0.99) ties our best early checkpoint (42.6) but ends at 32.5, below the frozen teacher's 35.1: the frozen teacher's role is anchoring, not signal freshness. (Full three-stage description in Appendix F.)

(objective drift): under a constant learning rate, the clipped-divergence objective keeps moving the student away from the initial policy, and the drift itself is harmful once the learnable signal is exhausted. We designed one decisive experiment per hypothesis.

Against H1: refreshing the teacher is catastrophic. If staleness were the cause, letting the teacher track the student should help. Our refresh operator merges the student's LoRA delta into the base every 100 steps (the teacher becomes the current student) and re-initializes the adapter. The result is the most degraded model in our study: mean 5.7 at step 300 (base 38.1), already degraded at step 200 (34.2). Generations are fluent but *empty* (Appendix A). Gentler anchor motion does not avoid the outcome. A fast EMA teacher (decay 0.99, Figure 5) ties our best early checkpoint at step 100 (42.6, within the seed band), holds 39.0 at step 200, yet falls to 32.5 by step 300, significantly below frozen's 35.1 (-2.6 $[-5.3, -0.2]$). Slower EMAs (39.4/39.5 at step 100, 32.9/32.7 at step 300) and a 200-step refresh interval (32.6 at step 300) complete an anchor-mobility spectrum. Terminal performance is monotone in how far the anchor moved (frozen 35.1, one late refresh 32.6 $\approx$ EMA 32.5–32.9, three refreshes 5.7). Pairing the EMA anchor with cosine decay recovers the decay endpoint exactly (39.3 = 39.3) and tracks the pure-decay trajectory throughout (40.9/38.8 vs. 40.2/38.6 at steps 100/200). The anchor-mobility penalty is a full-dose phenomenon. The frozen teacher is load-bearing: removing the anchor turns the objective into self-reinforcement.

For H2: decaying the learning rate arrests the collapse. If the failure is drift under a constant step size, annealing should fix it. With a cosine decay to zero over the 300-step horizon (and nothing else changed), the collapse is largely removed at 1.7B: 39.3 vs. 35.1 at step 300, ending above the untrained base. At 4B decay retains 57.7 vs. 51.7/51.8 (Figure 11). At 8B it ends within 0.1 of base (64.1 vs. 64.2, against 58.8 for constant LR). The same ordering holds on MATH500 at all three scales. The trade-off: annealed runs peak lower mid-training (their early rate is already smaller) but avoid the terminal degradation entirely (Table 1).

The combination fails: decay cannot save a moving anchor. Combining both interventions (refresh every 100 steps *plus* cosine decay) still collapses catastrophically: 14.4 mean at step 300 (4B replica: 1.4). Once the teacher stops anchoring, no optimizer-side remedy helps.

Dose–response: the clip threshold controls the drift rate. Our account predicts that loosening the pointwise clip should accelerate the drift even within the normal budget, and it does: λ=0.10 degrades step-100 to 35.8 (vs. 41.0), tightening to 0.03 leaves it unchanged (41.1), and the direction

replicates at 4B (59.0 vs. 60.4). The endpoint completes the curve. With the clip removed ($\lambda{=}\infty$, plain forward KL), step-100 falls to 33.2, decisively below both base (-4.8) and the clipped default (-7.8), and step-300 (33.7) ties the clipped baseline's own collapse, even though the unclipped loss is a true divergence and never goes negative (Appendix I, Q17). The clip thus rate-limits the drift itself, not merely style tokens.

Why the clipped objective drifts. The clip in Eq. 1 caps each vocabulary item's *positive* contribution to the forward KL at λ, while negative per-item terms ($p_T < p_S$) are unbounded below. Near the teacher the objective is no longer a true divergence: it goes negative and falls further as the student moves *away*. The asymmetry is quantitative (Proposition 1): with teacher $(0.25, 0.75)$ and $\lambda{=}0.05$, a one-hot on the 25% token costs -0.297, against 0 at the teacher. Symmetrizing the clip leaves the collapse intact (35.1 ± 0.5 vs. 34.6 ± 0.4), even though the symmetrized loss plateaus at -0.028, not -0.048. The asymmetry fixes the *direction and sign* of the drift; the cumulative learning-rate dose sets its pace (Section 6, Figure 6).

This reading has a measurable in-training signature (Figure 4b): the clipped loss crosses zero within twenty steps in every run, as predicted. Once the cosine schedule falls below half its peak, the annealed loss plateaus near -0.026 while the constant-LR loss descends to -0.048 as the benchmark score falls, a pattern that replicates across all six seed-schedule runs (Appendix J.1). The clip that stabilizes the first hundred steps thus converts the objective into a divergence-from-teacher reward in the low-signal regime, with the frozen teacher as the only brake until the constant learning rate overpowers it.

The drift also has a direct distance meter: the DAW statistics lane (Section 7) logs the unclipped per-sample teacher–student forward KL (Figure 12). Both schedules stop approaching the teacher within thirty steps. Thereafter the divergence grows with the learning rate spent: under constant LR it climbs monotonically to $5.6\times$ its minimum by step 300 ($1.6\times$/$1.9\times$/$3.6\times$ at steps 100/150/200); under cosine it settles near twice its minimum and never runs away.

6 Mitigations, and What Does Not Work

Table 1 collects the verdicts. Only optimizer-side interventions protect long-horizon performance, at all three scales, and the verdict survives three sanity checks. The decay-over-constant ordering holds under majority-vote and any-correct (pass@12) scoring at every scale (Appendix Table 9). It also holds seed by seed: the three-seed trajectories are indistinguishable at step 100 (40.8 ± 0.7 constant vs. 40.8 ± 0.5 cosine), separated by step 200 (37.0 ± 0.4 vs. 38.8 ± 0.4), and decisive at step 300 (34.6 ± 0.4 vs. 39.4 ± 1.0, non-overlapping). A paired bootstrap further places the decay advantage at $+4.2$/$+5.3$/$+5.3$ (1.7B/4B/8B), every interval excluding zero (Appendix Table 10). Everything that touches the teacher ends the horizon worse (refresh -29.4 at 1.7B, -47.8 at 4B, EMA-0.99 -2.6 below frozen), while DAW and the hardened hybrid tie the baseline at step 300 (Appendix Table 10).

Two schedule probes pin down what the fix requires. Linear decay to zero serves as well as cosine (39.2 vs. 39.3 at step 300; -0.1): the anneal's shape is immaterial. Nor is annealing required at all: a *constant* half-rate run (300 steps, exactly the cosine-300 dose) ends at 38.8, statistically tied with cosine (-0.5). What the fix buys is a smaller cumulative dose, not a decaying shape. Stretching the horizon is not free (cosine-500 still falls to 35.0), and annealing within the 100-step budget lowers the peak ($39.7 < 41.0$). Decay is insurance, not a peak-raiser.

These probes admit a compact summary (Figure 6). At fixed step count, terminal performance declines with the cumulative dose $B = \int \eta \, dt$ across schedule families, to within seed noise: 41.0 at $B{=}100$, 39.3/39.2/38.8 at $B{\approx}150$ (three families, one dose), 37.2 at $B{=}226$, and 35.0/35.1 at $B{=}250$/300. Below $B{\approx}100$ the ordering inverts (learning-limited regime). The converse cell qualifies the law. A constant full-rate run delivering the same dose $B{=}150$ in half the steps ends at 36.8, 2.0–2.5 points below all three $B{\approx}150$ families, on every benchmark. Terminal performance is thus governed by the dose together with the per-step rate: decay protects by halving the dose *and* by lowering the rate late in training. Each point remains a single seed: eight points, not a fitted law. The midpoints now filled in Table 1 sit where the account puts them. Every schedule family's step-100 value lies within about a point of the constant family's 41.0, inside the ± 1.5 seed band, and the dose-matched constant run scores 41.1 at $B{=}100$, matching the full-rate 41.0.

Table 1: What works and what does not at longer horizon (mean Avg@12, 1.7B unless noted). †: single seed; ±: s.d. over three seeds. Full per-benchmark values: Appendix K. The base model (no training) variant is step-invariant, with its single value spanning the three step columns.

Variant	step 100	step 200	step 300
Base model (no training)		38.1	
Frozen teacher (baseline)	41.0	36.9	35.1
DAW band-pass	41.5	37.9	34.5
RLVR-hybrid (hardened)†	41.6	37.1	33.4
Cosine LR decay to zero	40.2	38.6	**39.3**
Cosine LR decay, 3 seeds	40.8±0.5	38.8±0.4	39.4±1.0
Constant LR, 3 seeds	40.8±0.7	37.0±0.4	34.6±0.4
Constant 2.5×10^{-6} (dose-matched)	40.9	41.1	38.8
Linear LR decay to zero	41.0	39.9	39.2
Cosine decay over 500 steps	41.1	40.7	37.2
Teacher refresh every 100 steps	41.0	34.2	**5.7**
Teacher refresh every 200 steps	41.9	40.3	32.6
Refresh + LR decay (combo)	39.3	36.9	14.4
EMA teacher (decay 0.99)	42.6	39.0	32.5
EMA + cosine decay	40.9	38.8	39.3
DAW + cosine decay, 3 seeds	41.5±0.8	38.8±0.7	39.6±1.0
Symmetric clip (constant LR), 3 seeds	40.9±0.2	36.5±0.6	35.1±0.5
Clip removed ($\lambda=\infty$, constant LR)†	33.2	32.6	33.7
fp32 KL computation (constant LR)	41.0	37.8	35.4
Matched thinking-mode prompts (constant LR)	33.1	31.0	26.8
4B frozen teacher	60.4	56.4	51.7
4B DAW band-pass	61.1	56.6	51.8
4B cosine LR decay, 3 seeds	60.7	57.8	**57.7±0.5**
4B teacher refresh every 100 steps	60.4	41.2	**3.9**
4B refresh + LR decay (combo)	58.9	44.2	1.4
8B frozen teacher	65.4	60.3	58.8
8B DAW band-pass	65.3	61.4	58.0
8B cosine LR decay	64.2	63.7	**64.1**

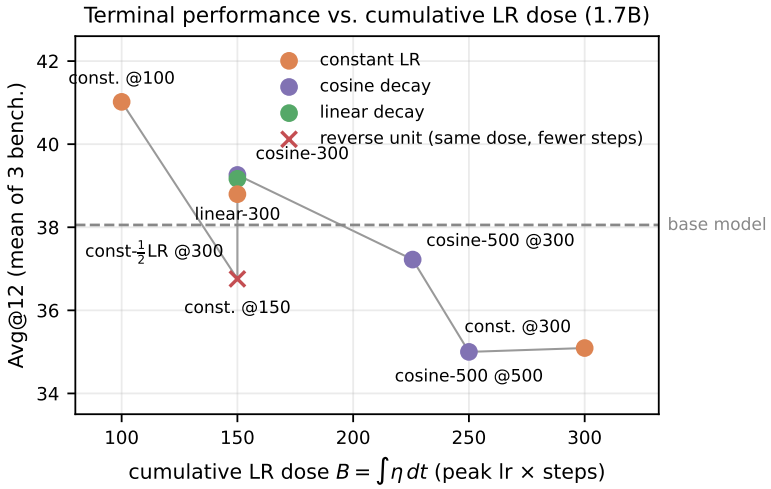


Figure 6: **Terminal performance vs. cumulative LR dose (1.7B).** Terminal mean Avg@12 against $B = \int \eta\, dt$; the × cell falls below the declining curve.

7 Secondary Studies

Difficulty-aware weighting. In the upper route of Figure 2, each sample is weighted by the percentile rank k_i of its mean unclipped teacher–student KL within the *global* batch (gathered across ranks, detached), through a band-pass profile $w_i = \varepsilon + (1-\varepsilon)\,4k_i(1-k_i)$ (ε=0.1, normalized to mean 1). Mastered samples ($k_i \to 0$) carry no signal, while samples far beyond the student ($k_i \to 1$)

inject noise the teacher cannot rationalize. The weighted reduction $\mathcal{L}_{\mathrm{DAW}} = \frac{\sum_i w_i \sum_t \ell_{i,t}}{\sum_i w_i T_i}$ (T_i: valid tokens of sample i, $\ell_{i,t}$ as in Eq. 1) keeps the global learning-rate scale unchanged.

Multi-seed results (Figure 10) show a consistent direction. Band-pass DAW matches or slightly exceeds the baseline at every seed (41.2±0.6 vs. 40.4±0.6, overlapping spreads) and is robust to its only hyperparameter (ε=0.3 yields 41.1). Ablations (Appendix B): down-weighting the hardest decile *hurts* (-2.1); the easiest is neutral ($+0.1$): hard samples carry growth signal even when imperfectly rationalized. DAW does not alter the long-horizon trajectory, collapsing with the baseline at step 300 (-0.6 $[-3.2, +2.0]$ at 1.7B, -0.8 $[-3.3, +1.7]$ at 8B; Appendix Table 10). The collapse is an optimization-time phenomenon, not a data-weighting one.

OPSD×RLVR hybrid: a numerics cautionary tale. The lower route of Figure 2 combines the two on one rollout (OPSD never uses answer correctness, and RLVR uses nothing else): $\mathcal{L} = \mathcal{L}_{\mathrm{OPSD}} + \alpha\,\mathcal{L}_{\mathrm{PG}}$, $\mathcal{L}_{\mathrm{PG}} = -(r_i - \bar{r})\,\frac{1}{T_i}\sum_t \log p_S(y_t \mid \cdot)$, with $r_i \in \{0, 1\}$ the math-verify correctness of the boxed answer, $\bar{r}$ the batch mean reward, a minimum-valid-token gate ($T_i \geq 8$, $\max(\cdot, 1)$-guarded), and NaN-guarded clamped log-probabilities.

The naive version produced a gradient norm of 2.7×10^{14} at step 35 (Figure 8). Although gradient clipping capped the update, the spike poisoned Adam's second-moment estimate, and the run ended far below baseline (36.3). The cause was unhardened numerics (bf16 underflow, unbounded log-probabilities, near-empty completions). The hardened term (clamping, NaN guards, minimum-length gate, α 0.1→0.02) trains stably (41.6 at step 100 on seed 42) with the largest seed variance of our variants; its three-seed mean (39.7±1.6) sits below the baseline (40.4±0.6) and DAW (41.2±0.6). A hardened probe at α=0.05 also trains stably (41.2), bracketing the stability boundary between 0.05 and the explosive 0.1. The single-seed 4B hybrid holds the highest 4B step-100 entry (62.7, inside the seed band; no scaling conclusion). All three step-100 checkpoints preserve their MATH500 ability essentially intact (92.1/91.8/91.9).

Component ablations and hyperparameter sweeps. Appendix B gathers both methods' ablations and knob sweeps: every removal is neutral or negative; every knob reads as a stability boundary with the default on the safe side.

8 Discussion and Limitations

When is OPSD the right tool? Frozen-teacher OPSD is a *short-horizon* method: gains arrive within the first hundred steps, and continuing past them is harmful unless the learning rate is annealed. The practical rule is to stop near the peak of a known horizon, or to decay to zero over it. Decay protects the endpoint rather than raising the peak (39.3 vs. 42.5); at 4B/8B it is the only endpoint near base (57.7 vs. 51.7; 64.1 vs. 58.8).

Limitations. (i) Protection is concentrated where the claims need it: the central decay-over-constant ordering holds over three seeds with non-overlapping ranges and bootstrap support, and the step-100 secondary comparisons and the symmetric-clip null replicate over three seeds; the remaining probes are single-seed, labeled P3 in Appendix L, and carry no load alone. Differences within $\sim$2 points are read as noise. (ii) The evidence covers mathematical reasoning at 1.7B–8B, LoRA students with 1024-token completions, and 30-problem benchmarks (MATH500 near saturation); headline numbers inherit the official train/eval prompt mismatch, quantified in Section 3. (iii) The drift account rests on training-log signatures and dose–response probes, not a formal analysis of the clipped objective's fixed points; it characterizes how fast drift proceeds, not how it degrades competence, and pure GRPO and KL-regularized RLHF remain untested (Appendix F.6).

9 Conclusion

Frozen-teacher on-policy self-distillation has a hidden expiry date: extended training collapses it below the base model, across variants and scales. The cause is not the teacher going stale: refreshing it is catastrophic, and gentler anchor motion only delays the fall. The collapse is objective drift under a constant learning rate, controlled by the cumulative dose. A horizon-aware decay to zero removes most of the collapse and ends above the untrained base. *Schedule the optimizer, not the teacher.*

Use of Generative AI.
LLM assistants were used only to polish English wording and to find candidate related work; all technical content is the authors' own. Every number was recomputed from the stored evaluation outputs, every cited paper verified against its source.

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 *ICLR*, 2024.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *ICML*, 2009.

Tommaso Furlanello, Zachary C Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born-again neural networks. In *ICML*, 2018.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS*, 2020.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *ICLR*, 2024.

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, et al. Rest: Reinforced self-training for language modeling. *arXiv preprint arXiv:2308.08998*, 2023.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

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

Jonas Hübotter. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026.

Yuki Ichihara, Naoto Iwase, Mohammad Atif Quamar, and Junpei Komiyama. Privileged solutions or context-induced teacher behavior? Dissecting on-policy self-distillation. *arXiv preprint arXiv:2608.09228*, 2026.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *UAI*, 2018.

Ijun Jang, Jewon Yeom, Juan Yeo, Hyunggu Lim, and Taesup Kim. Stable on-policy distillation through adaptive target reformulation. In *ACL*, 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.

Kyungyul Kim, ByeongMoon Ji, Doyoung Yoon, and Sangjun Hwang. Self-knowledge distillation with progressive refinement of targets. In *ICCV*, 2021.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. Distillm: Towards streamlined distillation for large language models. In *ICML*, 2024.

M Pawan Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. In *NeurIPS*, 2010.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huanang Gao, Wenkai Yang, Zhiyuan Liu, and Ning Ding. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*, 2026.

Zehao Liu, Yuanpu Cao, Jinghui Chen, and Vasant G. Honavar. Restoring the sweet spot: Pass-rate weighted self-distillation for llm reasoning. *arXiv preprint arXiv:2605.27765*, 2026.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *ICLR*, 2017.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, et al. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. In *NeurIPS*, 2021.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Guobin Shen, Xiang Cheng, Chenxiao Zhao, Lei Huang, Jindong Li, Dongcheng Zhao, and Xing Yu. Anti-self-distillation for reasoning rl via pointwise mutual information. *arXiv preprint arXiv:2605.11609*, 2026.

Yiqing Shen, Liwu Xu, Yuzhe Yang, Yaqian Li, and Yandong Guo. Self-distillation from the last mini-batch for consistency regularization. In *CVPR*, 2022.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. *arXiv preprint arXiv:2601.19897*, 2026.

Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *CVPR*, 2016.

Charlie Snell, Dan Klein, and Ruiqi Zhong. Learning by distilling context. In *ALR*, 2022.

Mingyang Song and Mao Zheng. A survey of on-policy distillation for large language models. *arXiv preprint arXiv:2604.00626*, 2026.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017.

Thinking Machines Lab. On-policy distillation, 2025. Thinking Machines Lab blog post.

Chenglin Yang, Lingxi Xie, Siyuan Qiao, and Alan L Yuille. Snapshot distillation: Teacher-student optimization in one generation. In *CVPR*, 2019.

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. Learning beyond teacher: Generalized on-policy distillation with reward extrapolation. *arXiv preprint arXiv:2602.12125*, 2026.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. In *NeurIPS*, 2022.

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

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

A Failure Case Gallery

Guide: this section shows what the refresh failure looks like in the model's actual outputs, and quantifies it in generation statistics, chiefly response length.

Generations from the step-300 teacher-refresh model are fluent but content-free. Example (truncated): *"First let me recap what was in my mind initially and think again how this approach differs... Starting over again... Maybe trying different angles would help clarify things? Wait but looking again—my original approach here involves translating everything step-wise..."* The model never commits to a computation across the full thinking budget. This behavioral signature, grammatical self-referential planning without progress, is distinct from the reasoning style of the base model and of the frozen-teacher checkpoints. It is consistent with our self-reinforcement account of the refresh failure, which the next paragraph quantifies.

The degeneration is quantifiable in the generation statistics. Mean response length (characters, averaged over 90 problems and 12 samples) stays between 51k and 64k for the untrained base and every non-refresh step-300 variant across all three scales; over all steps the band widens to 51–66k (Appendix Q). Under refresh it grows to 89k at step 200 and 102k at step 300 at 1.7B, to 140k for the refresh+decay combination, and to 162k (214k for the combination) at 4B, up to four times the base length. The refresh failure is thus runaway deliberation: more text, less content.

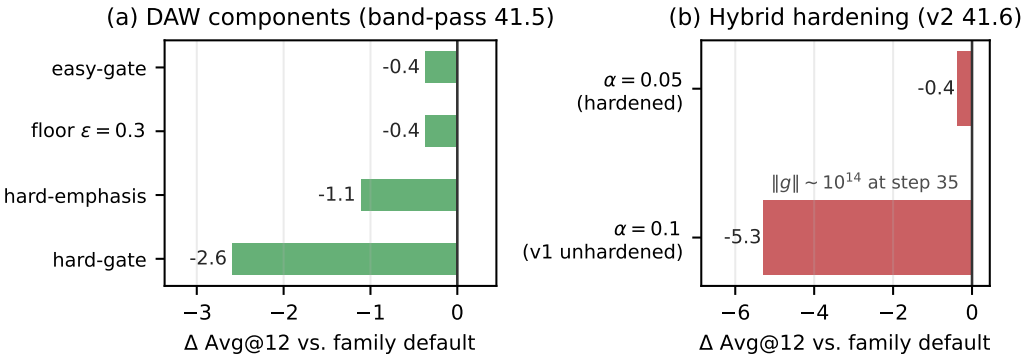


Figure 7: **Component ablations at step 100 (1.7B, seed 42).** Each bar is the change in mean Avg@12 relative to the family default (zero line): DAW band-pass (41.5) in (a), the hardened hybrid v2 (41.6) in (b). Every removal or deformation is neutral or negative: band-pass weighting is a local optimum. The hard-gate (-2.6) is the largest DAW cost; the unhardened hybrid v1 (-5.3) pays for the $\|g\|{\sim}10^{14}$ gradient explosion at step 35. Per-benchmark values in Appendix Table 13.

B Ablation and Hyperparameter Analyses

Guide: this section gathers the component ablations and hyperparameter sweeps referenced throughout the main text; results that anchor an earlier claim reappear here as display only.

Sections 5–7 deployed component removals as evidence at the point where each argument needed them; this section gathers them in one place. Results that anchor an earlier claim reappear here as display only; the interpretations stay where they were made.

Figure 7 and Table 2 decompose the two secondary methods (all deltas in the figure are relative to the family default). On the DAW side, replacing the band-pass profile with a single-tail gate removes the gain. Gating the easiest decile is neutral (-0.4), emphasizing the hardest costs 1.1, and gating the hardest costs 2.6: the hardest samples carry growth signal even when the teacher cannot rationalize them (Section 7). Raising the floor weight to $\varepsilon{=}0.3$ costs only 0.4, so the profile is insensitive to its one hyperparameter. On the hybrid side the hardening is essential rather than cosmetic: at $\alpha{=}0.1$ the unhardened objective explodes (gradient norm $\sim 10^{14}$ at step 35) and ends 5.3 below the hardened v2, while the hardened probe at $\alpha{=}0.05$ sits 0.4 below v2. The boundary between 0.05 and 0.1 is a joint one, since the 0.1 point is also the unhardened one (Section 7).

Table 2: Ablation family at step 100 (1.7B, mean Avg@12, seed 42).

Variant	Mean Avg@12
Frozen-teacher baseline	41.0
DAW band-pass	**41.5**
DAW easy-gate (down-weight easiest 10%)	41.1
DAW hard-emphasis (up-weight hardest)	40.4
DAW hard-gate (down-weight hardest 10%)	38.9
DAW band-pass with ε=0.3	41.1
EMA teacher (decay 0.999)	39.4
RLVR-hybrid v1 (unhardened, $\alpha = 0.1$)	36.3
RLVR-hybrid v2 dose probe ($\alpha = 0.05$)	41.2
Clip λ=0.03 (tighter)	41.1
Clip λ=0.10 (looser)	35.8
Clip removed (λ=∞)	33.2

B.1 Hyperparameter Sweeps

Beyond component removal, the continuous knobs (Figure 9) read as stability boundaries rather than tuning opportunities: each has a safe side and a cliff, and the defaults sit on the safe side. The two optimizer-side quantities (clip threshold and learning-rate schedule) are already established as mechanism levers (Sections 5 and 6). Table 3 reproduces the schedule probes in dose coordinates, for completeness: family, shape, and stopping step matter only through $B = \int \eta \, dt$ (Figure 6). Nothing else survives the reduction to a single curve.

Table 3: Schedule probes from Table 1, reproduced in dose coordinates (terminal mean Avg@12, 1.7B, seed 42). Display only; the interpretation is in Section 6.

Run	$B = \int \eta \, dt$	terminal score
Constant LR, stopped at step 100	100	41.0
Cosine decay over 300 steps	150	39.3
Linear decay over 300 steps	150	39.2
Constant $\frac{1}{2}$LR over 300 steps	150	38.8
Cosine over 500, stopped at step 300	226	37.2
Cosine over 500, fully annealed	250	35.0
Constant LR over 300 steps	300	35.1

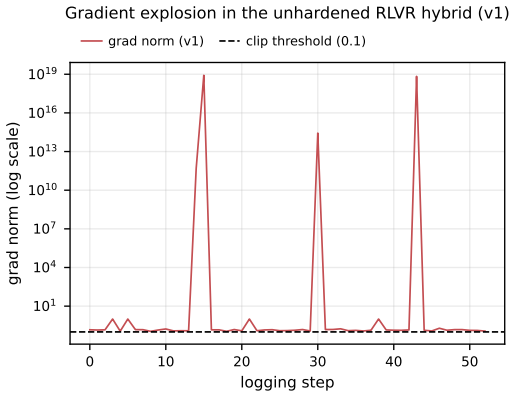


Figure 8: **Gradient explosion in the unhardened RLVR-hybrid objective.** Gradient norm (log scale) over training steps for the naive hybrid (v1). The $\sim 10^{14}$ spike is clipped by `max_grad_norm`=0.1, but the Adam second moment is already poisoned, and the run never recovers. The hardened objective (v2) removes the explosion entirely.

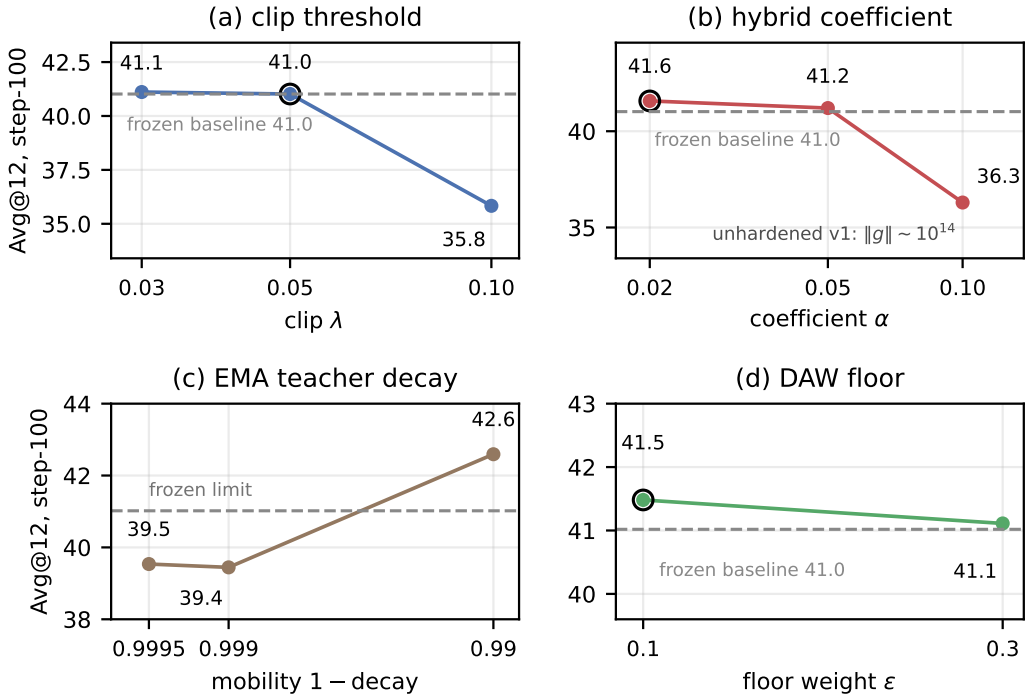


Figure 9: **Hyperparameter sweeps at step 100 (1.7B, seed 42).** Each panel varies one knob on a log axis; the dashed line is the frozen-teacher baseline (41.0), and open markers mark the defaults. (a) Tightening the clip to λ=0.03 changes nothing (+0.1); loosening to 0.10 costs 5.2 (the dose–response lever of Section 5; at 4B the same direction, 59.0 vs. 60.4). (b) The hybrid coefficient is flat up to α=0.05 and explodes at 0.1 (unhardened v1). (c) EMA teacher decay, plotted against anchor mobility $1 - \text{decay}$: the fastest anchor (0.99) reaches 42.6, tying the best step-100 checkpoint in our study (within the $\pm$1.5 seed band), while both slower ones fall below frozen (39.4, 39.5), the anchor spectrum of Section 5. (d) The DAW floor ε is insensitive (41.5 $\rightarrow$ 41.1). Per-benchmark values in Appendix Table 13, for every panel above.

C Extended Related Work

Guide: this section collects the full literature discussion behind Section 2, organized by which object each line of work keeps fixed and which it lets move.

On-policy distillation: fix the teacher, move the data. Classical knowledge distillation (KD) trains a student to match a teacher on a fixed dataset (Hinton et al., 2015), which for auto-regressive models creates a train–inference distribution mismatch. On-policy distillation fixes the data source instead: generalized knowledge distillation (GKD) trains the student on its own generations with flexible divergences (Agarwal et al., 2024); MiniLLM replaces forward KL with a policy-gradient reverse-KL objective to become mode-seeking (Gu et al., 2024); DistiLLM redesigns the divergence (skew-KL) and the sampling schedule to cut rollout cost (Ko et al., 2024); the Thinking Machines recipe (Thinking Machines Lab, 2025) demonstrates per-token teacher log-ratio rewards at scale; and Veto stabilizes on-policy KD by constructing an intermediate target distribution in logit space (Jang et al., 2026). All of these keep the teacher fixed and ask how the *data* or the *divergence* should move. None asks what happens when training continues far past the point where the student has absorbed the signal. Our work shows this is where the fixed-teacher design breaks, and what it takes to repair it.

Self-supervision and self-distillation: move the teacher. A complementary family lets the teacher evolve: Born-Again Networks retrain identical architectures sequentially (Furlanello et al., 2018); Mean Teacher uses an EMA of student weights as the consistency target (Tarvainen &

Valpola, 2017), an idea central to BYOL and momentum-encoder self-supervision (Grill et al., 2020; He et al., 2020); snapshot and last-epoch/batch variants distill from the model's own history (Yang et al., 2019; Kim et al., 2021; Shen et al., 2022). In LLM self-training, STaR and ReST bootstrap rationales from answer-checked self-generations (Zelikman et al., 2022; Gulcehre et al., 2023). Context distillation (Snell et al., 2022) uses the same model with and without privileged context; it is the direct philosophical ancestor of OPSD, but off-policy. Our teacher-refresh and EMA experiments are exactly these "move the teacher" remedies transplanted into OPSD, and their failure is informative. In vision/self-supervised learning (SSL) the student chases a slowly improving target on a fixed input distribution. In OPSD the refreshed teacher is *the student itself* on its own on-policy distribution, so the objective collapses to self-reinforcement, with nothing left to anchor the student.

RLVR and its documented failure modes. GRPO and its successors (DAPO, GSPO, Dr. GRPO) drive reasoning improvements with verifiable rewards (Shao et al., 2024; Guo et al., 2025; Yu et al., 2025; Zheng et al., 2025), and their failure modes are well catalogued: entropy collapse, zero-variance (all-correct/all-wrong) groups, and length bias. OPSD was motivated as an answer to the second of these: dense per-token signal where GRPO's group baseline vanishes. Our collapse is a *different* failure mode of the same family: not reward sparsity, but objective drift under prolonged optimization. The RLVR-hybrid experiment additionally documents a numerics-driven catastrophe (a $\sim 10^{14}$ gradient spike poisoning Adam) that this literature's stability analyses do not cover (Section 7).

The 2026 self-distillation wave and its failure analyses. Self-distillation has exploded into a crowded frontier: SDPO turns textual or environment feedback into dense self-teacher signals (Hübotter, 2026), SDFT uses demonstration-conditioned self-teaching to reduce forgetting in continual learning (Shenfeld et al., 2026), and a long tail of variants explores long context, reasoning compression, safety, and multilingual transfer. Three concurrent works are our closest neighbors, and the contrast with them clarifies our contribution. *(i) Analyses of when OPD fails* (Li et al., 2026; Song & Zheng, 2026) catalog failure phenomenology across objectives and teacher–student pairs, but treat the teacher–student pair as given and evaluate within standard budgets. Our object is orthogonal, namely the *optimization horizon* itself, and we show failure can be induced in a configuration their axes consider healthy. *(ii) A concurrent study of self-distillation degradation* (Kim et al., 2026) attributes failures to suppression of epistemic verbalization, i.e., a change in *what* the model says. Ours is a change in *where the optimization pressure points* once the student approaches the teacher: there the objective keeps descending by moving away from it (Section 5). The two accounts are complementary rather than competing. *(iii) OP^{2}SD* (Ichihara et al., 2026), an independent multi-seed reproduction of OPSD, corroborates the within-budget results we also reproduce, while neither study extends the horizon. Finally, on the method side, ExOPD (Yang et al., 2026) casts OPD as dense KL-constrained RL and scales the implicit reward *past* the teacher, SC-SDPO (Liu et al., 2026) reweights self-distillation by pass rate to restore a difficulty sweet spot, and AntiSD (Shen et al., 2026) reads the OPSD token reward as a pointwise mutual information that can suppress deliberation tokens. Our RLVR hybrid is the reward-grounded cousin of ExOPD with the KL anchor kept. Our DAW shares SC-SDPO's motivation but uses the batch-relative *teacher–student divergence percentile* rather than pass rate as the difficulty signal. AntiSD's token-level analysis is consistent with our finding that the clipped token objective, not the teacher, is where the long-horizon failure lives (Figure 12 records it directly).

Sample weighting and curriculum. Reweighting training samples by learnability has a long history: curriculum learning orders easy-to-hard (Bengio et al., 2009), self-paced learning uses the loss itself as the curriculum signal (Kumar et al., 2010), and hard-example mining or loss-based pruning emphasizes the other tail (Shrivastava et al., 2016; Paul et al., 2021). DAW differs in the signal it uses: not loss magnitude against ground truth, but the *teacher–student divergence percentile within the batch*, a measure of how rationalizable a sample currently is for the frozen teacher. Our finding that down-weighting the hardest decile *hurts* while band-pass weighting helps slightly complicates the standard curriculum intuition (Appendix B). The hardest samples carry growth signal even when the teacher cannot fully rationalize them.

Stability and schedules. Finally, our mitigation connects to a classical observation: averaging or annealing the optimizer stabilizes training (Izmailov et al., 2018; Loshchilov & Hutter, 2017).

In post-training, KL penalties to a frozen reference play a related anchoring role in RLHF/RLVR (Ouyang et al., 2022). We show that in self-distillation the anchor is not a separate KL term. It is the frozen teacher itself, and the horizon over which the learning rate stays high determines when the anchor is overpowered and the collapse begins.

D Experimental Details

Guide: this section records the hardware and software environment and the exact training, evaluation, and teacher-refresh configurations needed to reproduce every run in the paper.

Hardware. All experiments ran on NVIDIA A800-SXM4-80GB nodes (8 GPUs per node, NVLink-interconnected, compute capability 8.0; Ubuntu 22.04.4 LTS, kernel 5.15, dual-socket Intel Xeon Platinum 8374C with 144 cores and 1 TiB RAM per node). Training and evaluation together use the node's 8 A800 GPUs, with runs scheduled in parallel lanes across them; evaluation serves vLLM with tensor parallelism 4. Model weights, checkpoints, and per-problem evaluation outputs all reside on the nodes' local storage.

Software environment. Training and evaluation share one conda environment, pinned as in Table 4: PyTorch (bf16, FlashAttention-2) drives training under DeepSpeed ZeRO-2 with CPU optimizer offload; vLLM in colocate mode generates all training rollouts and also serves every evaluation; PEFT provides the LoRA adapters; TRL is the trainer backbone of the official OPSD codebase we extend (Section 3); and math-verify scores answers. The system CUDA toolkit is 12.4 (`/usr/local/cuda`) while the PyTorch wheels ship CUDA 12.8 runtime libraries (an API-compatible pairing, as the runtime logs record), and the wheel-bundled NCCL 2.27.3 backs the collectives. Runs are launched with Accelerate, and experiment tracking uses Weights & Biases. The machine-readable environment file (the exact `environment.yml` the runs were created from) accompanies the code release, along with the canonical bootstrap script.

Table 4: Software environment (one conda environment, all runs; versions as installed on the training nodes at experiment time, exactly as pinned in the released `environment.yml`).

Component	Version
Python	3.10
NVIDIA driver	580.126.09
CUDA (toolkit / PyTorch build)	12.4 / 12.8
NCCL	2.27.3
PyTorch	2.8.0
Transformers	4.57.1
TRL	0.26.0
Accelerate	1.11.0
PEFT	0.17.1
DeepSpeed	0.18.2
vLLM	0.11.0
xFormers	0.0.32.post1
FlashAttention-2	2.8.3
Triton	3.4.0
Datasets	3.6.0
NumPy	2.2.6
Weights & Biases	0.22.3
math-verify	0.8.0

Training. All runs use the official OPSD configuration unless noted: Qwen3-1.7B / Qwen3-4B / Qwen3-8B base models; LoRA $r = 64$, $\alpha = 128$ on q,k,v,o,gate,up,down projections; global batch 32 (per-device micro-batch and gradient accumulation configured so the global batch is invariant); AdamW, lr 5×10^{-6}, max grad norm 0.1; cosine or default linear schedule as indicated; temperature 1.1, top-p 0.95, top-k 20 for rollouts; completion length 1024; forward KL (β=0) with pointwise clip 0.05; bf16, FlashAttention-2, gradient checkpointing, DeepSpeed ZeRO-2 with CPU optimizer offload, vLLM colocate (GPU memory fraction 0.4–0.5). Each experiment runs on the node's 8

Table 5: Reproduction check (Qwen3-1.7B, Avg@12, temperature 1.0, thinking mode). "Paper" numbers are from Zhao et al. (2026); "ours" is the seed-42 reproduction on the official codebase and configuration, with all platform switches off (Section 3).

	AIME24	AIME25	HMMT25	Mean
Base (paper)	51.5	36.7	23.1	37.1
Base (ours)	54.7	36.4	23.1	38.1
OPSD step-100 (paper)	57.2	41.1	29.2	42.5
OPSD step-100 (ours)	54.2	40.0	28.9	41.0

A800-80GB GPUs, in parallel lanes. Long-horizon runs keep the official 30-epoch schedule (approximately constant lr) unless labeled cosine decay.

Evaluation. Thinking mode enabled; temperature 1.0, top-p 0.95, no top-k, min-p 0, presence penalty 0; 12 samples per problem (Avg@12 = mean accuracy); max 38,912 new tokens; vLLM with tensor parallel 4. Answers are extracted from the last `\boxed{}` and verified by math-verify (5s timeout, string-match fallback). Benchmarks: AIME 2024, AIME 2025, HMMT 2025 (30 problems each); MATH500 with val_n=4 to bound evaluation cost.

Teacher refresh operator. At each refresh step: merge the LoRA delta into the base (teacher becomes the current student), re-initialize the adapter ($A \sim$ kaiming, $B = 0$), and unmerge the (now zero) delta (a no-op on the weights that resets PEFT's merge-state bookkeeping) so the base retains the merged weights, then save a cleaned full model for evaluation. Checkpoint-based evaluation is invalid because PEFT checkpoints store only the adapter, never the merged full model that evaluation must load.

Seeds. seed 42 for all curve/mechanism runs; multi-seed step-100 runs use seeds $\{42, 1, 2\}$ for the baseline, DAW band-pass, and hybrid v2, respectively.

Reproduction calibration details. Two calibration facts matter throughout. (i) Our base-model numbers exceed the paper's on AIME24 (54.7 vs. 51.5), so gains are computed against *our own* base; per-benchmark gaps between our step-100 reproduction and the published numbers are 0.3–3.0 points (Table 5). (ii) The published step-100 numbers are single-seed, while our three-seed replication puts the spread at ± 1.5 points (Figure 10). Training and evaluation also use different prompt distributions in the official codebase ("Problem:"-prefixed non-thinking rollouts capped at 1024 tokens vs. thinking-mode evaluation at 38,912), so every published number carries a train/eval mismatch that our reproduction inherits. Appendix F.3 quantifies its effect.

Defect details. Of the defects in Table 7, two matter most. The KL/JSD loss is computed entirely in bf16: only $\sim$3 significant digits survive the cancellation terms, plausibly the hidden reason the clip is needed at all. We also fixed a memory-inefficient $[B, T, V]$ boolean-index copy in the loss reduction that OOMs at per-device batch 8; the replacement is memory-efficient but mathematically identical. The probes isolating each defect from the long-horizon outcome are in Appendix F.3: fp32 recomputation moves step 300 by $+0.3$ $[-2.1, +2.8]$; matched thinking-mode prompts still decline from 33.1 to 26.8 (-6.3 vs. the baseline's -5.9); reverse KL (β=1) climbs to a plateau ($32.6 \to 36.7 \to 36.0$) and JSD β=0.5 (in fp32; $\beta \in (0, 1)$ is unstable in bf16) is flat ($34.9 \to 35.3$); both end at the collapsed forward-KL level (36.0/35.3 vs. 35.1).

Code and artifact availability. The anonymous repository (`https://anonymous.4open.science/r/Schedule-the-Optimizer-Not-the-Teacher-ICLR-2027-DCB6`) contains the complete artifact set behind the paper. It is a fork of the official OPSD code (commit `7448751`) whose default code path is unmodified; every extension used in this paper is gated behind a flag that is default-off (Section 3). The set contains the extended training code with all switches, the evaluation pipeline, per-problem correctness for all 128 evaluation files (`per_problem_scores.json`), and the append-only results ledger with 313 records (`compact_results.json`), to which every number in the paper traces. It further contains the canonical bootstrap script, which regenerates every interval in Table 10 bit-for-bit under

its fixed seed, the figure-generation script, the 20 training logs behind the loss and divergence figures, and CPU-runnable unit tests for the reward parsing and the refresh checkpointing. A `REPRODUCING.md` file maps each paper claim to the exact command that produces it, mirroring the run registry of Appendix N, and the pinned environment file matches Table 4 entry for entry.

E E-SERIES SUPPLEMENTARY RUNS

Guide: per-benchmark values of the supplementary probe families referenced in Sections 5 and 6: the EMA anchor spectrum, the dose family (including the reverse unit), and the divergence family.

Table 6: E-series supplementary runs at 1.7B, per-benchmark Avg@12 (%). Source: `compact_results.json`.

Variant	step	AIME24	AIME25	HMMT25	Mean
EMA anchor spectrum					
EMA teacher (0.99)	100	55.6	42.5	29.7	42.6
EMA teacher (0.99)	300	43.3	33.3	20.8	32.5
EMA teacher (0.999)	100	52.2	38.1	28.1	39.4
EMA teacher (0.999)	300	42.2	33.3	23.1	32.9
EMA teacher (0.9995)	100	54.7	38.6	25.3	39.5
EMA teacher (0.9995)	300	44.7	33.3	20.0	32.7
EMA + cosine decay	300	53.6	38.9	25.3	39.3
Dose family (schedule shapes and the reverse unit)					
Cosine LR decay to zero	200	52.2	38.1	25.6	38.6
Cosine LR decay to zero	300	53.3	40.3	24.2	39.3
Linear LR decay to zero	200	52.5	37.5	29.7	39.9
Linear LR decay to zero	300	52.2	40.0	25.3	39.2
Constant 2.5×10^{-6} (dose-matched)	300	51.1	39.2	26.1	38.8
Constant LR, 150 steps (reverse unit)	150	48.3	38.6	23.3	36.8
Cosine decay over 500 steps	300	50.6	36.9	24.2	37.2
Cosine decay over 500 steps	500	45.6	34.7	24.7	35.0
Cosine decay within 100-step budget	100	56.1	38.3	24.7	39.7
Divergence family					
Reverse KL (β=1)	100	44.4	35.0	18.3	32.6
Reverse KL (β=1)	200	46.7	37.2	26.1	36.7
Reverse KL (β=1)	300	47.2	37.8	23.1	36.0
JSD (β=0.5)	100	45.8	36.1	22.8	34.9
JSD (β=0.5)	300	48.3	34.2	23.3	35.3

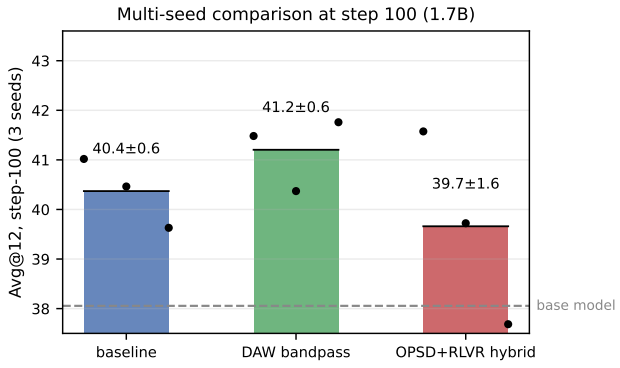


Figure 10: **Multi-seed comparison at step 100 (1.7B).** Mean Avg@12 over the three competition benchmarks for three independent seeds (points) and their mean (bar).

The EMA spectrum's terminal levels (32.5/32.9/32.7) sit within 0.4 of each other across benchmarks: anchor mobility, not anchor rate, sets the endpoint. The reverse unit (constant full rate stopped at step 150, B=150) lands below all three same-dose families on every benchmark. The dose law is therefore dose-primary with a per-step-rate correction.

F ANTICIPATED CONCERNS

Guide: this section assembles, in one place, the analyses behind the objections a skeptical reader is most likely to raise. Each subsection uses only experiments already reported or data already collected, and traces every number to the tables of this appendix.

The EMA-teacher loop in detail. The loop runs in three stages on one shared model. *(1) Generate and score:* the student (LoRA θ_t on the shared frozen base, which is never updated and never enters the EMA) samples one rollout $\hat{y}_i$ without gradients and rescores it with gradients, $p_S(\cdot;\theta_t)$. *(2) Evaluate with EMA LoRA:* a temporary swap inside the same model saves θ_t, loads the EMA buffer $\bar{\theta}_t$, runs the teacher forward with the privileged reference y_i^* under no_grad, and restores θ_t before any update. The teacher predictions $p_T(\cdot;\bar{\theta}_t)$ stay detached. *(3) Update and compare:* the clipped OPSD loss $\mathcal{L}_{\mathrm{OPSD}}(p_T, p_S)$ drives the student update $\theta_t \to \theta_{t+1}$, and only then the buffer follows, $\bar{\theta}_{t+1} = 0.999\,\bar{\theta}_t + 0.001\,\theta_{t+1}$: the EMA tracks the LoRA weights as a slow average, not a second full model. Despite the smoothly tracking teacher, step-100 mean Avg@12 drops from 41.0 (frozen) to 39.4 (EMA 0.999). The fastest variant we scanned (decay 0.99) seals the pattern: 42.6 at step 100, tying the best early checkpoint in our study within the seed band, but 32.5 at step 300, below the frozen teacher's 35.1.

F.1 WHY THESE HORIZONS?

The 300-step horizon is three times the published budget, chosen to be far past the peak without entering a regime no practitioner would run. The decline is already decisive at 300 under the bootstrap (-7.4 from the peak), and at 8B it is visible by step 200 (65.4 $\to$ 60.3). Stretching further does not help (the cosine-500 probe, fully annealed, lands at 35.0 at step 500, on par with the constant-LR collapse), so the phenomenon is not an artifact of stopping at one particular horizon. Shorter horizons are covered by the dose–response relation (Section 6): below $B \approx 100$ the ordering inverts because the score is learning-limited, so nothing in our account predicts harm inside the published budget.

F.2 IS THE REFRESH FAILURE AN IMPLEMENTATION ARTIFACT?

Five independent checks argue it is a real property of removing the anchor. (i) It replicates at 4B, more abruptly (41.2 at step 200, 3.9 at step 300; combo 1.4). (ii) It appears across the whole anchor-mobility spectrum (EMA at three decay rates and a 200-step refresh interval all end below the frozen teacher), so it does not depend on the discrete merge-and-reset operator. (iii) Combining refresh with LR decay does not rescue it (14.4 and 1.4), ruling out "too much learning rate" as the explanation. (iv) The failure has a quantitative signature independent of scoring: response length grows to 89k–102k characters at 1.7B and up to 214k at 4B, two to four times every non-refresh variant (Appendix A). (v) The evaluation protocol is valid for this variant specifically because we evaluate the saved merged model rather than adapter checkpoints (Appendix D). Two further observations close the remaining confounds. The refresh operator is function-preserving (the merged base plus a zeroed adapter reproduces the student's logits exactly), but it does re-initialize the adapter and reset optimizer state at each merge. That reset, however, is common to refresh alone: the EMA variant involves no merge and no reset yet fails in the same direction, and failure is monotone across the whole mobility spectrum, so a mechanism present at one point of the spectrum cannot explain the spectrum. Nor is the failure optimization instability: after each refresh the training loss descends smoothly (Figure 13b) with bounded gradient norms. The failure signature is instead distributional: response lengths at two to four times base and, at 4B, a boxed-answer rate collapsing to 24.5%. We add one process note in the same spirit: an initial implementation of the refresh operator (v1) zeroed the adapter's B matrix and thereby froze the student. We caught it because both checkpoints produced numerically identical evaluations (36.3/36.3), discarded the run, and re-implemented the

operator (v2) before drawing any conclusion. A run that cannot change the model cannot produce the refresh collapse; the reported failure required a correctly moving anchor.

Table 7: Implementation-level defects identified in the official OPSD code (severity H/M/L), and which results they touch; Section 3 quantifies each of them.

Defect	Sev.	Consequence
KL/JSD computed in bf16 without fp32 upcast	H	Noisy divergence estimates; plausibly the hidden reason clipping is required at all
Train/eval prompt mismatch ("Problem:" prefix; non-thinking rollouts vs. thinking-mode eval)	M	Train–eval distribution shift present in all published numbers
Non-thinking scripts clip at 10^{-6}–10^{-7}	M	Near-zero gradients by construction; we confirmed empirically that the shipped non-thinking config performs no learning at all (6.4 mean, statistically indistinguishable from a sane-clip non-thinking run at 6.9 and from the non-thinking base); published non-thinking results are irreproducible with the shipped configs
Loss reduction ignores `num_items_in_batch`	M	Micro-batch equal-weighting under gradient accumulation
$[B, T, V]$ boolean-index copy in loss reduction	M	Out of memory at per-device batch 8 (fixed in our platform)
`lmbda` is a dead parameter; README file tree inaccurate	L	Misleading GKD mixing option

F.3 Do the flagged defects drive the collapse?

Each defect from Table 7 has a probe isolating it from the long-horizon outcome. Recomputing the divergence in fp32 moves step-300 performance by $+0.3$ $[-2.1, +2.8]$. Training with thinking-mode rollout prompts matched to the evaluation format still declines (33.1 at step 100 to 26.8 at step 300, -6.3 versus the baseline's -5.9), with matched runs starting lower because 1024-token thinking rollouts truncate before committing to an answer. This compares absolute declines across two regimes with different starting points: it rules out the mismatch as the driver of the decline, not as a contributor to its size. Symmetrizing the clip leaves the collapse intact (35.1 ± 0.5 vs. 34.6 ± 0.4) even though the training loss plateaus (-0.028 vs. -0.048), which separates the boundedness of the objective from the drift itself. The shipped non-thinking configurations perform no learning at all under their near-zero clips (6.4 and 6.9, indistinguishable from the non-thinking base), so those published numbers are unaffected by any collapse mechanism. The defects are real and worth fixing (Section 3); none of them produces, or removes, the collapse.

F.4 Scope: LoRA, 1024 tokens, three scales, mathematics only

All runs use the official configuration: LoRA rank 64/α 128 on all linear projections, 1024-token completion budget, mathematical reasoning only, at 1.7B–8B. The collapse's size is not monotone in scale (-5.9/-8.7/-6.6 at 1.7B/4B/8B), so we make no claim that it grows with capacity. Full-parameter students, longer completion budgets, and non-math domains are untested; the mechanism account (drift under cumulative learning rate in a clipped objective that stops being a divergence) makes predictions there, but verifying them is future work (Section 8, limitations iii–iv).

F.5 Isn't this just overfitting?

The peak-then-decline shape resembles the textbook overfitting curve, and the question deserves a direct answer. Four facts separate the two. First, the entire decline unfolds *within a fraction of one epoch*. With 29,434 training examples and global batch 32, one epoch is $\sim$920 steps, so our step-300 horizon sits at epoch 0.33 (as the training logs record) and even the 500-step probe reaches only 0.54.

The student collapses before finishing a single pass over the training set, having seen about a third of it once. Second, the training loss falls *below zero*, under its value at the teacher distribution (0, by Proposition 1i). Fitting the data better cannot take a divergence below its value at the target, while moving away from the teacher can, so the late loss decrease reflects drift away from the teacher, not improved fit. Third, the attribution points away from a data-side account. Reweighting the data (DAW) leaves the terminal score unchanged (-0.6 $[-3.2, +2.0]$), while purely optimizer-side changes (annealing, halving the rate) determine the outcome. Fourth, three seeds (three data orders and adapter initializations) land within a 0.8-point range of each other at step 300 ($35.1/34.4/34.3$; s.d. 0.4). A data-order-specific pathology would not do that. We conclude the phenomenon lives on the optimization-dynamics side. Classic overfitting builds on repeated exposure to the data, and the horizon here provides less than a single pass.

F.6 What would GRPO do at the same horizon?

We did not run pure GRPO to 300 steps. It is a different training stack, and our compute went into three-scale replication and seed protection of the phenomenon we set out to diagnose. Two pieces of evidence bound the answer, though. The hardened RLVR-hybrid, which adds a verifiable-reward policy-gradient term to the *same* rollouts, is the closest in-family probe of outcome grounding. It collapses alongside the baseline (33.4 at step 300; -1.7 $[-4.4, +0.9]$ vs. constant LR), so outcome grounding alone does not protect a horizon. The RLVR literature itself documents its own failure modes (entropy collapse, zero-variance groups, length bias; Section 2), so "switch to GRPO" is not known to be a horizon fix. Our verdict is deliberately scoped. The horizon pathology of frozen-teacher OPSD is real, diagnosed, and dose-controlled; whether other objective families share it is an open question our hybrid result makes sharper, not one we answer.

F.7 Is the step-300 drop an evaluation artifact?

Three checks say the collapse is a loss of correctness, not a failure of extraction. First, the boxed-answer format rate barely moves. Across base, step-100, and step-300 checkpoints of every non-refresh variant at all three scales, 94.4–100% of the 12 samples per problem still yield a parseable `\boxed{}` answer (Table 8). The collapsed model still commits to an answer, but the answer is more often wrong. Second, the drop propagates to majority-vote and any-correct scoring in the same order (Table 9), which an extraction failure sensitive to sampling would not do. Third, the damage is concentrated in the mid-difficulty band rather than spread uniformly (Appendix H), the opposite of a formatting or truncation failure. The refresh failure is the informative contrast: at 4B it alone breaks extraction itself. The format rate falls from 97.7% (base) to 74.4% at step 200 and 24.5% at step 300 while response length inflates to two to four times base (Appendix A). At 1.7B the format rate holds near 94% while scores collapse to 5.7, i.e., the model still commits to a boxed answer that is simply wrong. The two failure modes have quantitatively distinct signatures, and neither one contaminates the constant-LR collapse that carries our central claim.

F.8 Why is 4B the worst case, and why doesn't decay fully rescue it?

4B is the sore thumb of our own tables: the largest constant-LR drop (-8.7 from step 100 to 300) and the only scale where the decayed run stays reliably below base (-3.8 $[-7.0, -0.6]$). Decomposing the residual with the stratified analysis (Table 11) locates it precisely. The 4B decay run *restores* the mid band above base (55.0 vs. 53.3) and the 1–4 band above it (23.1 vs. 21.8), and its entire residual deficit sits in the easy band (85.7 vs. 92.8, -7.1). This is the same easy-band erosion seen in milder form at 1.7B (-3.4) and 8B (-2.4): the consolidation cost of annealing, amplified. Two observed facts then explain why 4B fares worst overall. Within the published budget, 4B gains the least of the three scales (step-100 mean 60.4 vs. base 60.7, within the ± 1.5 seed band, where 1.7B gains $+2.9$ and 8B $+1.2$), so there is less protected gain to keep. It also drifts most per unit dose, as its dose curve shows: B=100 (its step-100 checkpoint) already ties base at 60.4, B=150 (cosine-300) lands at 56.9, and B=300 (constant) at 51.7. This is the same dose–response relation as 1.7B but shifted down, so that no dose in our measured grid nets above base at 4B. The decay schedule still recovers the same absolute amount as at the other scales ($+5.3$, confidence interval (CI) $[+2.7, +8.1]$). What it cannot do is create within-budget gains that the objective did not deliver.

Table 8: Boxed-answer format rate (% of generations from which a parseable \boxed{} answer is extracted; 30 problems × 12 samples per benchmark). Non-refresh variants keep extraction intact at every scale and step; the 4B refresh run alone breaks it, deepening with dose. Source: the per-problem evaluation records retained for every run in this paper.

Run	AIME24	AIME25	HMMT25	Mean
1.7B base	98.6	99.7	99.7	99.3
1.7B step-100 (constant LR)	98.3	98.1	99.2	98.5
1.7B step-300 constant LR	98.9	98.6	100.0	99.2
1.7B step-300 cosine decay	98.9	98.6	98.9	98.8
1.7B refresh step-300	95.3	91.9	95.6	94.3
4B base	98.9	96.7	97.5	97.7
4B step-100 (constant LR)	98.1	94.7	94.4	95.7
4B step-300 constant LR	99.2	99.2	98.9	99.1
4B step-300 cosine decay	98.1	96.4	96.1	96.9
4B refresh step-200	74.7	71.4	77.2	74.4
4B refresh step-300	25.6	16.1	31.7	24.5
8B base	98.9	94.7	95.6	96.4
8B step-300 constant LR	98.6	97.5	98.9	98.3
8B step-300 cosine decay	98.1	93.6	93.9	95.2

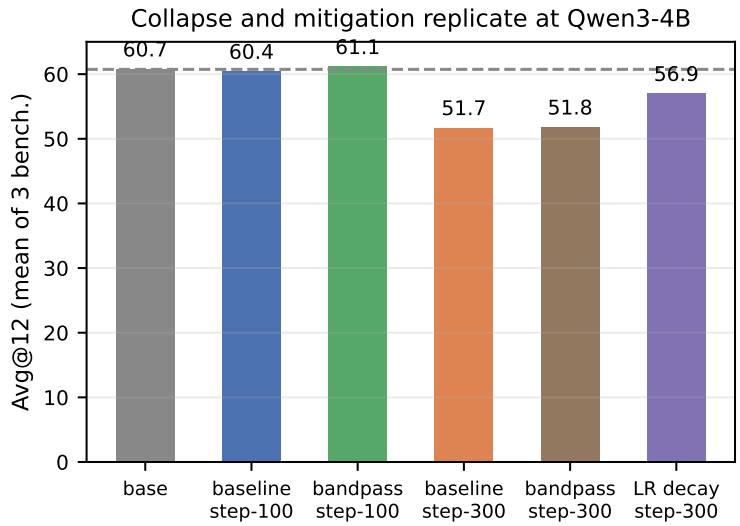


Figure 11: **The collapse and its mitigation replicate at Qwen3-4B.** Mean Avg@12 for base, step-100, and step-300 variants. Constant-LR training at 300 steps falls far below base (∼52), while cosine LR decay retains 57.7±0.5 over three seeds (the bar shows the seed-42 run, 56.9). The failure mode and its cause are not scale-specific artifacts of the 1.7B setting; the same dose-controlled fix works at this scale too.

G Scoring-Rule Robustness

Guide: this section answers whether the key step-300 orderings survive changes of scoring rule (Table 9) and quantifies, for every comparison quoted in the text, what a paired bootstrap over the underlying problems can and cannot order (Table 10).

Table 9 recomputes the key step-300 comparisons under the three scoring rules available from our 12-sample protocol. The ordering produced by mean accuracy (decay above constant LR, both above refresh) holds unchanged under majority-vote and any-correct scoring at every scale. The 8B

decay run exceeds the base model under majority vote (73.3 vs. 70.0) despite tying it under mean accuracy, so the near-recovery at 8B is not an artifact of the averaging rule.

Table 9: Scoring-rule robustness for the key step-300 comparisons (mean over the three competition benchmarks; 12 samples per problem). Maj = majority vote, Pass = any correct. All rows are seed-42 runs; the 4B cosine decay mean over three seeds is 57.7 ± 0.5 (Table 1).

Scale	Run	Avg@12	Maj	Pass
1.7B	Base (no training)	38.1	48.9	65.6
	Baseline peak (step 75)	42.5	51.1	68.9
	Constant LR, step 300	35.1	42.2	61.1
	Cosine decay, step 300	39.3	50.0	65.6
	Teacher refresh, step 300	5.7	13.3	23.3
4B	Base (no training)	60.7	66.7	78.9
	Constant LR, step 300	51.7	61.1	73.3
	Cosine decay, step 300	56.9	66.7	77.8
8B	Base (no training)	64.2	70.0	77.8
	Constant LR, step 300	58.8	70.0	78.9
	Cosine decay, step 300	64.1	73.3	78.9

Table 10 completes the statistical picture with paired-bootstrap confidence intervals for every step-300 comparison quoted in the text. Each pair of runs is aligned problem by problem over the 90 problems. Entries are mean differences in Avg@12 (first-named run minus second), and intervals are percentile CIs over 10,000 resamples of the problems. Intervals that contain zero mark differences our data cannot reliably order. Examples include the decay-versus-base comparison at 1.7B and 8B (the recovery is statistically complete there) and the DAW/hybrid-versus-baseline comparisons (the secondary variants track the baseline's collapse indistinguishably). At 4B, by contrast, the decay run still sits reliably below the base (-3.8 $[-7.0, -0.6]$). Among the newer probes, the dose-matched constant run, the symmetric clip, fp32, and DAW+decay are all indistinguishable from their comparators. The EMA-0.99 step-300 deficit is clear (-2.6, interval excluding zero), and its slower rates confirm it without bootstrap (EMA 0.999/0.9995 end at 32.9/32.7, both below frozen's 35.1 and within 0.4 of the 0.99 terminal). EMA+cosine decay lands exactly on the decay endpoint ($39.3 = 39.3$), confirming the mobility penalty as a full-dose phenomenon. The 200-step-refresh deficit is marginal (-2.5, interval containing zero). The fully unclipped run ends statistically tied with the clipped baseline (-1.4) yet reliably below base (-4.4), the measured endpoint of the clip dose–response (Section 5), completing the curve at its far end.

Two properties of this machinery deserve disclosure. The intervals are not adjusted for multiplicity, and several graze zero (the 1.7B decay advantage has lower bound $+0.9$; the EMA-0.99 deficit, -0.2). Across this many comparisons one or two borderline exclusions are expected, so no single interval is asked to carry a claim. The headline ordering rests on the three-scale replication and the seed-wise separation of Section 6. The resampling unit is the problem, pooled across the three benchmarks. The per-problem records needed for the stricter stratified variant are retained with our release materials, so the stricter analysis can be rerun.

H Where the Collapse Lives

Guide: this section locates the damage in the difficulty distribution (and what each intervention does to each difficulty band), separating "loses skill at the margin" from "loses the ability to commit to an answer at all" (the two failure modes of our tables).

Table 11 stratifies the 90 competition problems into four buckets by how many of the 12 base-model samples are correct (0, 1–4, 5–8, 9–12) and reports each key run's mean accuracy within a bucket. Three facts stand out. *(i)* No run at any scale makes material progress on problems the base never solves (at most 3.2 points in the entire table). OPSD's gains and its collapse both live on problems the base could already partially solve. *(ii)* The long-horizon decline is concentrated in the mid-difficulty band (base-solvable problems the model is in the process of consolidating): under constant LR, $52.4 \to 42.3$ at 1.7B, $53.3 \to 33.3$ at 4B, and $61.1 \to 48.6$ at 8B, with a shallower erosion of the easy

Table 10: All paired-bootstrap comparisons at step 300 (points of mean Avg@12, first-named run minus second; 95% percentile CIs over 10,000 problem resamples).

Scale	Comparison	Δ	95% CI
1.7B	Constant LR vs. base	-3.0	$[-6.0, 0.0]$
	Constant LR vs. peak (step 75)	-7.4	$[-10.7, -4.4]$
	Cosine decay vs. constant LR	$+4.2$	$[+0.9, +7.6]$
	Cosine decay vs. base	$+1.2$	$[-1.7, +4.1]$
	Linear vs. cosine decay	-0.1	$[-2.6, +2.6]$
	DAW vs. constant LR	-0.6	$[-3.2, +2.0]$
	Hybrid vs. constant LR	-1.7	$[-4.4, +0.9]$
	Refresh vs. constant LR	-29.4	$[-36.4, -22.6]$
	Constant $\frac{1}{2}$LR vs. cosine decay	-0.5	$[-2.8, +1.9]$
	DAW+decay vs. cosine decay	$+1.8$	$[-0.4, +3.9]$
	Symmetric clip vs. constant LR	$+0.7$	$[-1.8, +3.1]$
	EMA 0.99 vs. constant LR	-2.6	$[-5.3, -0.2]$
	Refresh-200 vs. constant LR	-2.5	$[-5.5, +0.4]$
	fp32 vs. constant LR	$+0.3$	$[-2.1, +2.8]$
	Clip removed ($\lambda{=}\infty$) vs. constant LR	-1.4	$[-3.9, +1.0]$
	Clip removed ($\lambda{=}\infty$) vs. base	-4.4	$[-7.1, -1.8]$
4B	Constant LR vs. base	-9.1	$[-12.7, -5.5]$
	Cosine decay vs. constant LR	$+5.3$	$[+2.7, +8.1]$
	Cosine decay vs. base	-3.8	$[-7.0, -0.6]$
	Refresh vs. constant LR	-47.8	$[-55.5, -39.9]$
8B	Constant LR vs. base	-5.4	$[-8.6, -2.3]$
	Cosine decay vs. constant LR	$+5.3$	$[+2.4, +8.3]$
	Cosine decay vs. base	-0.1	$[-2.5, +2.4]$
	DAW vs. constant LR	-0.8	$[-3.3, +1.7]$

band ($92.8 \to 80.8$ at 4B). *(iii)* Cosine decay restores the mid band to or above the base level at 1.7B and 4B (53.6 and 55.0) and recovers most of it at 8B (58.3 vs. 48.6 under constant LR). Teacher refresh, by contrast, destroys *every* band including the easy one ($94.7 \to 18.7$ at 1.7B, $92.8 \to 6.1$ at 4B, $95.0 \to 6.5$ at 8B). Under refresh the model does not lose skill at the margin; it loses the ability to commit to any answer at all (Appendix A shows the generations).

Table 11: Difficulty stratification: mean accuracy within buckets defined by the number of base-model solves out of 12. Peak is the step-75 checkpoint at 1.7B and step-100 at 4B and 8B. Refresh was run at all three scales.

Scale	Bucket (base solves)	n	Base	Peak	Const. 300	Decay 300	Refresh 300
1.7B	0	31	0.0	3.2	1.6	1.6	0.3
	1–4	20	16.2	27.5	13.8	22.5	0.4
	5–8	14	52.4	58.9	42.3	53.6	2.4
	9–12	25	94.7	94.0	89.7	91.3	18.7
4B	0	19	0.0	0.9	0.4	0.4	0.0
	1–4	13	21.8	23.7	14.7	23.1	0.0
	5–8	5	53.3	46.7	33.3	55.0	5.0
	9–12	53	92.8	92.0	80.8	85.7	6.1
8B	0	20	0.0	3.0	2.9	2.9	0.2
	1–4	9	20.4	31.4	21.3	29.6	0.0
	5–8	6	61.1	70.8	48.6	58.3	7.0
	9–12	55	95.0	95.6	86.4	92.6	6.5

I Frequently Asked Questions

Guide: compact answers with pointers; each claim traces to the sections cited.

Q1. The benchmarks have 30 problems each. How much statistical power do the comparisons have? Every headline comparison is a paired bootstrap over the 90 pooled problems (Table 10); the same orderings hold under majority-vote and any-correct scoring (Table 9) and in direction on MATH500's 500 problems. Gaps our intervals cannot order are reported as ties, and single-checkpoint differences within $\sim$2 points are treated as seed noise (Appendix L).

Q2. Why train for 300–500 steps? 300 steps is three times the published budget, the decline is decisive there (-7.4 $[-10.7, -4.4]$ from the peak), visible by step 200 at 8B, and stretching to 500 does not recover (35.0). Shorter horizons fall in the learning-limited regime below $B \approx 100$, where we predict and observe no harm (Section 6; Appendix F.1).

Q3. Isn't LR decay just early stopping in disguise? No: a constant run at half the learning rate for 300 steps, never annealed, matches the decayed runs (38.8 vs. 39.3, CI $[-2.8, +1.9]$); what matters is the cumulative dose $B = \int \eta \, dt$, not the shape. Conversely, annealing *within* the 100-step budget lowers the peak (39.7 vs. 41.0), so decay is insurance against an unknown horizon, not a free win (Section 6, Figure 6). It buys safety, not points.

Q4. Many runs are single-seed. Why trust them? Appendix L maps every verdict to its protection level. The central ordering (decay above constant LR at step 300) has three seeds with non-overlapping ranges (39.4 ± 1.0 vs. 34.6 ± 0.4) and bootstrap intervals excluding zero at all three scales; the single-seed probes are labeled P3 and carry no load alone.

Q5. Couldn't the refresh failure be an implementation bug? It replicates at 4B, appears across the EMA spectrum and the refresh+decay combo, has a scoring-independent signature (response length two to four times base), and was validated against exactly this concern, including a caught-and-discarded buggy first implementation whose failure mode was numerically identical checkpoints (Appendix F.2 gives the full five-check account).

Q6. Doesn't the train/eval prompt mismatch (or bf16) explain the collapse? Matched-prompt runs decline by -6.3 versus the baseline's -5.9, fp32 changes step 300 by $+0.3$ $[-2.1, +2.8]$, and the symmetric-clip null separates objective boundedness from drift (Appendix F.3).

Q7. The proposition covers a two-token marginal. Does it extend to the full vocabulary? We claim the sign and direction of the bias and an elementary construction, not global optimality over the simplex for general V; Part (ii) gives a profitable drift direction for every token pairing, and the symmetric-clip experiment tests the asymmetry claim empirically (Proposition 1; Section 5).

Q8. Did you try teacher refresh together with LR decay? Yes. The combination collapses (14.4 at 1.7B, 1.4 at 4B): once the anchor moves, no optimizer-side remedy helps (Section 5).

Q9. Why does the dose–response relation invert below $B \approx 100$? Two regimes: learning-limited at small doses (more dose, more score: 39.7 at B=50 vs. 41.0 at B=100) and drift-limited above $B \approx 100$, where the terminal score declines with dose, to within seed noise. The relation is stated for the drift regime (Section 6).

Q10. LoRA-only students, 1024-token completions: does any of this transfer? Those are the official configuration and stated limitations (iii)–(iv); we make no transfer claim and the mechanism account's predictions there are untested (Appendix F.4).

Q11. What did this cost? Roughly 1,850 A800 GPU-hours in total, of which about 480 are analog-estimated for runs replicated on an external machine; just over 80% of the total is evaluation under the 12-sample thinking-mode protocol (Appendix M, Table 17).

Q12. How does this relate to concurrent 2026 analyses of self-distillation failures? Our axis (optimization horizon) is orthogonal to failure-phenomenology catalogs; the closest account of degradation (suppressed epistemic verbalization) is complementary (a change in what the model says versus whether the objective remains a divergence), and an independent multi-seed reproduction corroborates our within-budget numbers (Section 2).

Q13. Isn't the peak-then-decline curve just overfitting? No. The decline reaches below-base performance at epoch 0.33, before one pass through the 29,434-example training set, so the repeated exposure overfitting requires never occurs. The training loss falls below its value at the teacher (0), which improved fitting cannot produce; the negative mass is reachable only by moving away from the teacher. Attribution agrees: data-side reweighting leaves the terminal score unchanged (-0.6 $[-3.2, +2.0]$) while optimizer-side changes determine the outcome. Three seeds (independent data orders) end within a 0.8-point range (s.d. 0.4). Appendix F.5 gives the four-part argument.

Q14. Would GRPO at the same horizon avoid the collapse? We did not run pure GRPO. It is a different training stack, and our compute went into three-scale replication and seed protection of the phenomenon we set out to diagnose. The closest in-family probe, the hardened hybrid (which adds a verifiable-reward policy-gradient term to the same rollouts), collapses alongside the baseline (33.4 at step 300, -1.7 $[-4.4, +0.9]$), so outcome grounding alone does not protect a horizon. The RLVR literature documents its own failure modes (entropy collapse, zero-variance groups, length bias), so switching families is not known to be a horizon fix. Our verdict is deliberately scoped to OPSD's horizon; whether other objective families share the pathology is an open question (Appendix F.6).

Q15. Could the step-300 drop be an evaluation artifact: answers that no longer parse, or generations cut off by the token cap? No. Every non-refresh variant keeps a 94.4–100% boxed-answer format rate at every scale and step (Table 8), and an extraction or truncation failure would suppress exactly this rate. The drop propagates to majority-vote and any-correct scoring in the same order (Table 9) and concentrates in the mid-difficulty band rather than spreading uniformly (Appendix H), both inconsistent with a formatting artifact. The one run that does break extraction is the 4B refresh (74.4% at step 200, 24.5% at step 300, with response length two to four times base), a signature distinct from the constant-LR collapse that carries our central claim (Appendix F.7).

Q16. Why is 4B the worst scale, and why does decay leave it below base? The 4B decay run restores the mid-difficulty band above base (55.0 vs. 53.3); its entire residual deficit sits in the easy band (85.7 vs. 92.8), the consolidation cost of annealing in amplified form (-3.4 at 1.7B, -2.4 at 8B). The recovery over constant LR is the same absolute size as at the other scales ($+5.3$ $[+2.7, +8.1]$). Two observed facts make 4B the worst case: it gains least within the budget (step-100 60.4 vs. base 60.7), and its dose curve is shifted down (60.4/56.9/51.7 at $B{=}100/150/300$), so no dose in our measured grid nets above base there. We state this as a limitation rather than explain it away (Appendix F.8 quantifies the easy-band residual deficit).

Q17. The per-item clip c (Eq. 1): is it just a truncation, and would removing it help? It is a pointwise truncation, capping each vocabulary item's *positive* contribution to the forward KL at c ($\lambda{=}0.05$ nats; the token-summed KL is not capped), and it is essential at every point of the sweep: loosening to $\lambda{=}0.10$ costs 5.2 points within the budget ($41.0 \to 35.8$ at step 100, same direction at 4B), tightening to $\lambda{=}0.03$ is free ($+0.1$), and it plausibly suppresses exactly the bf16 divergence noise our defect analysis flags (Section 3).

We also trained the fully unclipped endpoint ($\lambda{=}\infty$, the plain forward KL; 1.7B, 300 steps, seed 42). Removing the clip does not help: it hurts in both regimes. Within the budget it destroys the gains outright: step-100 mean 33.2, a decisive -7.8 $[-11.1, -4.5]$ below the clipped default and already below the untrained base (-4.8 $[-7.6, -2.1]$). Notably, the unclipped loss is a true divergence and never goes negative (it descends from 0.15 toward $\sim$0.06 and stays there), so this damage needs no negative-loss channel at all. At the horizon there is no rescue either: step-300 lands at 33.7, reliably below base (-4.4 $[-7.1, -1.8]$) and statistically tied with the clipped baseline's own collapse (35.1; -1.4 $[-3.9, +1.0]$). The symmetric-clip null is the mirror image: capping the negative side as well plateaus the training loss at -0.028 yet leaves step-300 performance unchanged (35.1 ± 0.5 vs. 34.6 ± 0.4). Both halves of the account are now measured. The clip protects the within-budget regime, and neither bounding nor unbounding the objective changes the long-horizon outcome, which is controlled by the cumulative learning-rate dose rather than by the clip (Section 5).

J THE CLIPPED OBJECTIVE IS NOT A DIVERGENCE

Guide: this section states and proves the formal claim behind the drift's signature and direction (the clipped objective can be decreased by moving away from the teacher) and delimits exactly what the proposition does and does not assert. Its final subsection collects the loss-trajectory evidence (Figure 13) behind the magnitudes quoted in the remarks that follow the proof.

Write the per-position objective of Eq. 1 as

$$\ell(p_S) \;=\; \sum_{v\in\mathcal{V}} \min\big\{\, t_v\big(\ln t_v - \ln s_v\big),\; \lambda \,\big\}, \qquad t_v \triangleq p_{T,v},\; s_v \triangleq p_{S,v}, \tag{2}$$

a function on the simplex Δ^{V-1} with per-item clip $\lambda > 0$ ($\lambda = 0.05$, nats, in all our runs; the teacher logits are temperature-scaled at $\tau = 1.1$, which only fixes which distribution t is).

Proposition 1. (i) $\ell(p_T) = 0$. (ii) *Let i, j be vocabulary items with $q = t_i$ and $p = t_j$ such that $p > q > \lambda/\ln 2$ and $p \ln \frac{p}{p-q} \geq \lambda$. Then the distribution p'_S obtained from p_T by moving mass q from j to i (doubling i, starving j) satisfies $\ell(p'_S) = -q\ln 2 + \lambda < 0$.* (iii) *In the two-item case $\mathcal{V} = \{1, 2\}$ with $t = (q, 1-q)$, $q < \frac{1}{2}$ and $q\ln(1/q) > \lambda$, the global minimum of ℓ is attained at the one-hot distribution on the minority item, with value $q\ln q + \lambda < 0$ (any $q \gtrsim 1.2\%$ suffices at the default clip setting of $\lambda = 0.05$).*

Proof. (i) Every summand vanishes at $s = t$. (ii) The term for i becomes $q\ln\frac{q}{2q} = -q\ln 2$; the term for j becomes $\min\{p\ln\frac{p}{p-q},\, \lambda\} = \lambda$ by assumption; all other terms are unchanged, and $-q\ln 2 + \lambda < 0$ since $q > \lambda/\ln 2$. (iii) Parametrize $s = (s, 1-s)$. If the majority term is clipped ($s \geq 1-(1-q)e^{-\lambda/(1-q)}$), the minority term is unclipped there (its clip would require $s \leq qe^{-\lambda/q}$, while $s \geq 1-(1-q)e^{-\lambda/(1-q)} \geq q$), so $\ell(s) = q\ln(q/s) + \lambda$, which is strictly decreasing in s and equals $q\ln q + \lambda$ at $s = 1$. Otherwise the majority term is unclipped. If the minority term is unclipped as well, ℓ is the ordinary KL and $\ell \geq 0$. If the minority term is clipped ($s \leq qe^{-\lambda/q}$), the majority term is nonnegative there, so $\ell \geq \lambda > 0$. In every case $\ell(s) \geq q\ln q + \lambda$, with equality only at $s = 1$, so the one-hot distribution on the minority item is the global minimum. □

Remarks. First, the direction of the bias: the clip caps the *penalty* for under-weighting the teacher's confident tokens at λ per vocabulary item, while the *reward* for over-weighting its mid- and low-probability tokens accumulates freely. Part (ii) says that pairing any $\gtrsim 7.2\%$ token ($\lambda/\ln 2$ at λ=0.05) with a higher-probability token whose starved contribution already saturates the clip ($p\ln\frac{p}{p-q} \geq \lambda$) yields a profitable over-weighting direction. With the full vocabulary, each starved item still costs λ, so the optimum does not go one-hot, but the attractor has moved strictly off the teacher. Second, the magnitude: the observed per-token training loss reaches -0.018 at step 100 in both constant-LR and annealed runs and continues to -0.048 under constant LR while plateauing at $\sim$$-0.026$ once annealed (Figure 4b), i.e., the late decrease the optimizer is exploiting is precisely the negative mass the clip fails to bound. Third, the clip sweep of Section 5 delimits what the proposition alone explains: by part (iii) the reachable floor rises as λ loosens (the minimum $q\ln q + \lambda$ grows with λ), so the observed acceleration under looser clips ($41.0 \to 35.8$ at step 100 from λ=0.05 to 0.10) is not a deeper negative region; it is the optimization dynamics that the clip rate-limits, consistent with the unclipped run degrading fastest while its loss never goes negative (Appendix I, Q17). Tightening to 0.03 leaves the behavior within the reported budget essentially unchanged.

J.1 LOSS TRAJECTORIES ACROSS SEEDS AND VARIANTS

Guide: this subsection extends the loss-drift signature of Section 5 to all three seeds and the full variant family (Figure 13), and fixes the convention under which terminal loss levels are quoted.

Figure 4b showed the seed-42 pair. Panel (a) of Figure 13 repeats the diagnostic for all three seeds. In all six runs the clipped loss crosses zero between steps 18 and 20; the constant and annealed trajectories track each other until the schedules separate; and in every seed the constant-LR run ends deeper than its annealed counterpart (five-step terminal means -0.047, -0.029, -0.029 versus -0.026, -0.020, -0.020 for seeds 42/1/2). Under the proposed schedule the loss converges in the ordinary sense as well. Late in training it flattens (-0.021 at step 150 drifting to -0.026 at 300, against -0.025 to -0.047 for the constant runs), and gradient norms stay bounded throughout

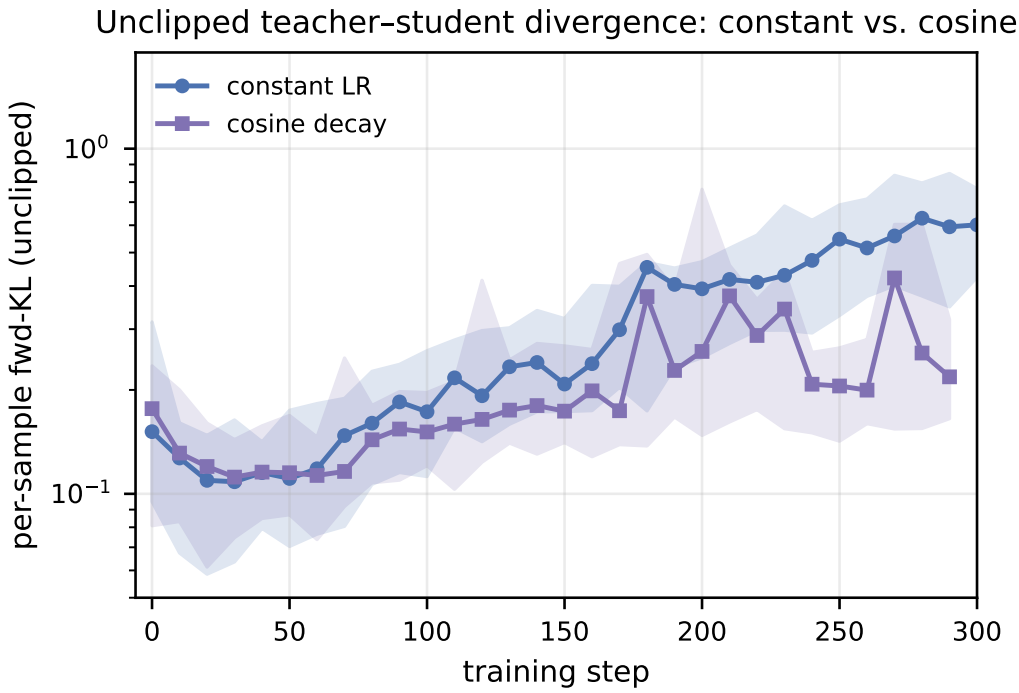


Figure 12: **Unclipped teacher–student divergence: constant vs. cosine LR.** Batch means (lines) and p10–p90 across samples (shading) of the detached, unclipped forward KL logged every ten steps by the DAW statistics lane (upper route of Figure 2); both runs are DAW runs, whose long-horizon outcome is statistically indistinguishable from the baseline's (Table 10); we report these curves as supporting evidence for the baseline's drift dynamics. The mean is token-weighted and need not lie inside the band; log scale. Both schedules stop approaching the teacher within $\sim$30 steps; the constant-LR divergence then climbs monotonically to $5.6\times$ its minimum by step 300 (the final batch, mean 1.20, leaves the band), while under cosine it settles near twice its minimum over the full annealed horizon without running away. Direct evidence for the drift of Section 5: the student leaves the teacher faster the more learning rate is spent.

(maximum 0.155 across the three annealed runs, settling at 0.026–0.058 late in training), far from the 2.7×10^{14} explosion of the unhardened hybrid (Figure 8). One caveat keeps the reading honest: the terminal depth is a within-run drift meter, not a cross-run score. The same constant schedule ends at -0.047 under seed 42 but -0.029 under seeds 1 and 2. Only the early zero crossing and the within-seed ordering are seed-invariant, and those are what the mechanism claims rest on.

Panel (b) holds the seed (and hence the data order) fixed and asks how deep each variant drives the same objective; panel (a) has already warned that absolute depths are run-specific, so it is the ordering that carries information. Three factors leave separate fingerprints. *Dose*: the runs that deliver half the cumulative dose (B=150) all end shallow, whatever their shape, with cosine at -0.026, the constant half-rate at -0.022 (the same dose as the cosine, collected without ever annealing), and DAW+cosine at -0.020. At the loss level too, dose rather than schedule shape is the controlling variable. *Objective-side caps*: two full-dose runs stop shallow for reasons that have nothing to do with the schedule. The symmetric clip bounds the negative region by construction and plateaus at -0.029, yet the collapse proceeds unchanged (Section 5). The fp32 probe stops near -0.030 (not shown), so part of the late drift rides on bf16 logit precision. *Anchor*: the full-dose runs with unbounded objectives end equally deep, to within the run-to-run variability established above (frozen teacher -0.047, EMA-0.99 -0.049). The refresh run resets the meter outright: at the step-200 teacher merge the objective jumps from -0.017 back to $+0.002$ and re-descends only to -0.016 by step 300 while the benchmark score falls from 40.3 to 32.6. The loss level, in other words, tracks how far the current objective can still be pushed, not the score; the shallowest terminal loss in the panel belongs to a run that ends 2.5 points below the constant baseline.

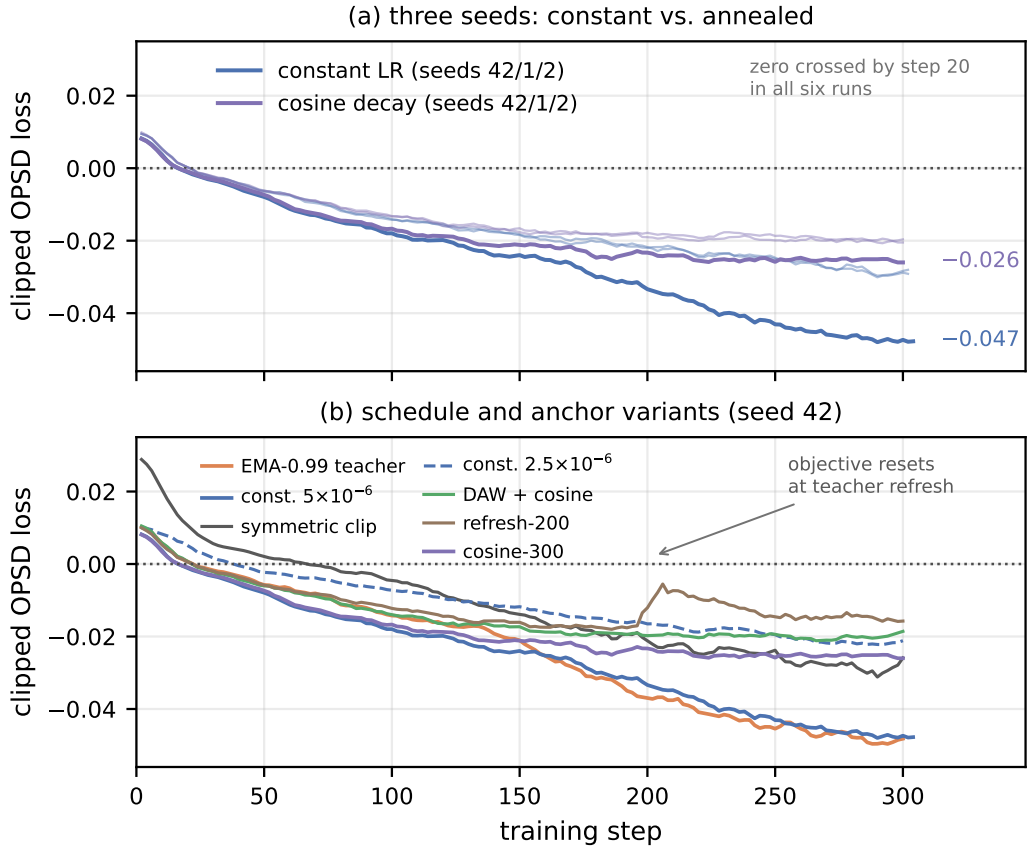


Figure 13: **Training-loss trajectories (1.7B): the annealed objective converges, the constant one keeps drifting.** (a) Clipped OPSD loss, constant LR vs. cosine decay, three seeds each (seed 42 heavier, seeds 1/2 faint). All six runs cross zero by step 20; in every seed the constant run ends deeper than the annealed one (terminal five-step means -0.047 vs. -0.026 for seed 42, -0.029 vs. -0.020 for seeds 1/2). (b) Seed-42 variants: the unbounded full-dose runs sit deepest (frozen teacher -0.047, EMA-0.99 -0.049), the half-dose (B=150) runs cluster at -0.020 to -0.026 regardless of shape, the symmetric clip caps the full-dose descent at -0.029, and the refresh run's objective resets at the step-200 teacher merge ($-0.017 \rightarrow +0.002$, annotated). Loss is logged every two steps; display curves are 5-point edge-padded means; quoted terminal levels are last-five-value means.

A note on conventions: terminal levels in this subsection are means of the last five logged values; the single-point annotations in the main text (-0.048 in Figure 4b, -0.028 in Section 5) differ from these by at most two thousandths and refer to the same plateaus.

K Full Results

Guide: this section collects every per-benchmark number behind the paper's figures and tables, so that any aggregate we report can be recomputed from primary results.

Tables 12, 14, and 15 report all per-benchmark numbers used in this paper, and Table 13 reports the ablation and hyperparameter probes of Appendix B per benchmark.

Table 12: Full 1.7B results (Avg@12 per benchmark; mean over the three). MATH500 uses val_n=4.

	AIME24	AIME25	HMMT25	Mean
Base	54.7	36.4	23.1	38.1

(continued)

	AIME24	AIME25	HMMT25	Mean
baseline step-25	52.8	40.0	25.0	39.3
baseline step-50	53.3	38.1	26.4	39.3
baseline step-75	55.6	43.9	28.1	42.5
baseline step-100	54.2	40.0	28.9	41.0
DAW step-25	52.8	40.0	25.8	39.5
DAW step-50	54.2	40.8	27.2	40.7
DAW step-75	52.8	41.7	28.3	40.9
DAW step-100	55.6	38.3	30.6	41.5
hybrid v2 step-100	55.3	40.3	29.2	41.6
baseline step-200/300	51.7/47.5	35.0/33.9	23.9/23.9	36.9/35.1
DAW step-200/300	52.5/43.6	36.4/36.7	24.7/23.3	37.9/34.5
hybrid step-200/300	50.6/45.0	36.7/33.6	24.2/21.7	37.1/33.4
LR decay step-100/200/300	54.7/52.2/53.3	39.4/38.1/40.3	26.4/25.6/24.2	40.2/38.6/39.3
linear decay step-100/200/300	54.2/52.5/52.2	39.8/37.5/40.0	28.9/29.7/25.3	41.0/39.9/39.2
decay-500 step-100	54.7	39.6	29.1	41.1
decay-500 step-200/300/500	53.3/50.6/45.6	40.7/36.9/34.7	28.1/24.2/24.7	40.7/37.2/35.0
teacher refresh step-100/200/300	53.8/46.1/8.9	40.7/35.6/5.0	28.6/20.8/3.3	41.0/34.2/5.7
refresh+decay combo step-100/200/300	51.9/48.3/21.1	38.5/36.7/13.9	27.4/25.6/8.3	39.3/36.9/14.4
teacher refresh every 200, step-100/200/300	53.8/55.0/43.3	41.5/39.7/34.2	30.4/26.1/20.3	41.9/40.3/32.6
EMA teacher (0.99) step-100/200/300	55.6/52.5/43.3	42.5/38.3/33.3	29.7/26.1/20.8	42.6/39.0/32.5
constant 2.5×10^{-6} step-100/200/300	53.7/55.8/51.1	40.3/39.5/39.2	28.6/28.1/26.1	40.9/41.1/38.8
cosine decay within 100-step budget	56.1	38.3	24.7	39.7
EMA + cosine decay step-100/200/300	54.2/52.6/53.6	40.8/38.2/38.9	27.8/25.6/25.3	40.9/38.8/39.3
fp32 KL computation step-100/200/300	53.6/52.8/51.4	40.8/35.8/31.7	28.5/24.9/23.1	41.0/37.8/35.4
symmetric clip step-300 (seed 42)	50.3	31.9	25.0	35.7
clip removed ($\lambda=\infty$) step-200/300	44.7/46.9	32.8/33.1	20.3/21.1	32.6/33.7
matched thinking-mode prompts step-100/200/300	42.8/39.5/35.0	34.2/31.1/25.6	22.5/22.5/19.7	33.1/31.0/26.8
Multi-seed step-300 (seeds 42/1/2; mean±s.d.):				
constant LR	35.1 / 34.4 / 34.3 → 34.6±0.4			
cosine decay	39.3 / 38.2 / 40.7 → 39.4±1.0			
DAW+cosine decay	41.0 / 38.9 / 39.0 → 39.6±1.0			
symmetric clip	35.7 / 34.4 / 35.0 → 35.1±0.5			
Hyperparameter analyses (step 100, seed 42):				
EMA decay 0.99 / 0.999 / 0.9995	42.6 / 39.4 / 39.5			
clip λ=0.03 / 0.05 (default) / 0.10 / none (∞)	41.1 / 41.0 / 35.8 / 33.2			
DAW ε=0.1 (default) / 0.3	41.5 / 41.1			
hybrid α=0.02 / 0.05 / 0.1 (v1)	41.6 / 41.2 / 36.3			
non-thinking, official clip 10^{-6} / sane 0.05	6.4 / 6.9			

(continued)

	AIME24	AIME25	HMMT25	Mean
MATH500 step 300: base / const / DAW / decay / refresh / combo	90.3 / 89.0 / 89.4 / 91.2 / 46.9 / 66.5			

Table 13: Ablation and hyperparameter probes, per benchmark (step 100, Avg@12, seed 42; Qwen3-1.7B except the 4B row; 4B baseline: 73.3/65.8/41.9, mean 60.4). Source: compact_results.json.

	AIME24	AIME25	HMMT25	Mean
Frozen-teacher baseline	54.2	40.0	28.9	41.0
DAW band-pass (ε=0.1)	55.6	38.3	30.6	41.5
easy-gate	54.7	41.1	27.5	41.1
hard-emphasis	53.1	40.8	27.2	40.4
hard-gate	51.4	39.2	26.1	38.9
ε=0.3	54.4	39.7	29.2	41.1
RLVR-hybrid v2 (α=0.02)	55.3	40.3	29.2	41.6
dose probe α=0.05	54.2	41.1	28.3	41.2
v1 unhardened (α=0.1)	47.8	38.6	22.5	36.3
EMA teacher, decay 0.99	55.6	42.5	29.7	42.6
EMA teacher, decay 0.999	52.2	38.1	28.1	39.4
EMA teacher, decay 0.9995	54.7	38.6	25.3	39.5
Clip λ=0.03 (tighter)	51.9	40.8	30.6	41.1
Clip λ=0.10 (looser)	48.3	35.6	23.6	35.8
Clip removed (λ=∞)	46.7	32.5	20.6	33.2
4B clip λ=0.10	71.4	63.1	42.5	59.0

Table 14: Full 4B results (Avg@12 per benchmark; mean over the three). MATH500 uses val_n=4.

	AIME24	AIME25	HMMT25	Mean
Base	74.2	65.6	42.5	60.7
baseline step-100/200	73.3/68.6	65.8/62.1	41.9/38.4	60.4/56.4
DAW step-100/200	73.3/70.2	65.3/60.7	44.7/38.9	61.1/56.6
baseline step-300	62.5	59.2	33.3	51.7
DAW step-300	66.4	55.8	33.1	51.8
LR decay step-100/200/300 (seed 42)	72.8/70.2/67.2	66.5/63.7/63.1	42.7/39.4/40.6	60.7/57.8/56.9
LR decay step-300, 3-seed mean	69.5	61.9	41.6	57.7±0.5
teacher refresh 100/200/300	73.2/51.7/4.4	66.1/42.5/1.7	41.8/29.4/5.6	60.4/41.2/3.9
refresh+decay combo 100/200/300	70.5/53.9/1.9	64.4/50.0/0.8	41.7/28.6/1.4	58.9/44.2/1.4
hybrid v2 step-100	76.7	66.9	44.4	62.7
clip λ=0.10 step-100	71.4	63.1	42.5	59.0
MATH500 step 300: base/const/DAW/decay	95.5/93.0/92.5/94.2			

Table 15: Full 8B results (Avg@12 per benchmark; mean over the three). MATH500 uses val_n=4. Bold marks the LR-decay row, the only variant whose step-300 mean stays at base level (64.1 vs. 64.2).

	AIME24	AIME25	HMMT25	Mean
Base	77.8	69.4	45.3	64.2
baseline step-100	78.3	71.4	46.4	65.4
baseline step-200	72.5	64.2	44.2	60.3
baseline step-300	72.2	63.6	40.6	58.8
DAW step-100	79.2	69.4	47.2	65.3
DAW step-200	73.9	65.8	44.4	61.4
DAW step-300	72.2	62.5	39.2	58.0
LR decay step-100/200/300	75.9/75.9/77.2	71.3/69.1/68.3	45.4/46.2/46.7	**64.2/63.7/64.1**
MATH500 step 300: base/const/decay		95.5/93.3/95.9		

L Robustness of the Verdicts

Guide: this section maps each verdict of the paper to its primary evidence and to its level of statistical protection, so that a reader can see at a glance which claims rest on multiple seeds and bootstrap confidence intervals and which are single-seed demonstrations.

Table 16 assigns every load-bearing claim a protection level. We use three levels. **P1** (strongest): the ordering holds seed by seed over three training seeds *and* a paired bootstrap over the 90 underlying problems excludes zero. **P2**: single training seed, but the paired bootstrap excludes zero (and, where applicable, the comparison replicates at another scale). **P3**: single seed, evidence is a curve, spectrum, or mechanism consistency rather than an interval; displayed for orientation, and we do not build further claims on these differences alone.

Table 16: Verdict-to-evidence-to-protection map. "Seeds" counts independent training seeds behind the comparison; bootstrap intervals are 95% percentile CIs over 10,000 problem resamples (Table 10). P1/P2/P3 as defined in the text of this section.

Verdict	Evidence	Seeds	Statistical protection
Constant-LR training ends below base at 300 steps, all scales	Fig. 3b, Tabs. 1, 9	3 (1.7B)	P2: vs. base, -9.1 $[-12.7, -5.5]$ (4B), -5.4 $[-8.6, -2.3]$ (8B), -3.0 $[-6.0, +0.0]$ (1.7B, borderline); holds under Maj/Pass
Peak-to-300 decline is decisive, not noise	Fig. 3a	1	P2: -7.4 $[-10.7, -4.4]$; monotone over six checkpoints
Teacher refresh is catastrophic	Tab. 1, App. H, A	1, 4B replica	P2: -29.4 $[-36.4, -22.6]$ (1.7B), -47.8 $[-55.5, -39.9]$ (4B)
Gentler anchor motion only delays the fall	Tab. 1	1	P2/P3: EMA-0.99 -2.6 $[-5.3, -0.2]$; refresh-200 -2.5 $[-5.5, +0.4]$ (not significant; direction consistent across the mobility spectrum)
Cosine decay removes most of the collapse, all scales	Figs. 4a, 11, 3b	3 (1.7B)	P1 at 1.7B: $+4.2$ $[+0.9, +7.6]$, seed ranges non-overlapping; P2 at 4B/8B: $+5.3$ $[+2.7, +8.1]$, $+5.3$ $[+2.4, +8.3]$, constant side single-seed (4B cosine over three seeds); holds under Maj/Pass

(continued)

Verdict	Evidence	Seeds	Statistical protection
The controlling variable is cumulative dose B, not schedule shape	Fig. 6, Tab. 3	1 per probe	P3: three schedule families within 0.5 pts at $B{=}150$; half-rate constant vs. cosine -0.5 $[-2.8, +1.9]$ (null)
DAW: small within-budget gains, no long-horizon rescue	Fig. 10, Tab. 2	3	P2 null: vs. constant LR at 300, -0.6 $[-3.2, +2.0]$ (1.7B), -0.8 $[-3.3, +1.7]$ (8B)
Hardened hybrid: stable, not better	Fig. 2, Tab. 1	3	P2 null: vs. constant LR at 300, -1.7 $[-4.4, +0.9]$; largest seed spread of our variants
Flagged implementation defects do not drive the collapse	§3, App. F	1	P2 nulls: fp32 $+0.3$ $[-2.1, +2.8]$; symmetric clip $+0.7$ $[-1.8, +3.1]$; matched prompts decline by the same amount
Loosening or removing the clip accelerates the drift	§5, Fig. 9	1 (+4B)	P2: $41.0 \to 35.8$ at $\lambda{=}0.10$; removed ($\lambda{=}\infty$) 33.2 at step 100, -7.8 $[-11.1, -4.5]$ vs. default, -4.8 $[-7.6, -2.1]$ vs. base; 4B direction replicated (59.0 vs. 60.4)

Two boundaries on what we claim. First, differences our intervals cannot order are reported as ties and never used as evidence. DAW and the hardened hybrid are indistinguishable from the baseline at step 300, and we conclude from this that neither rescues the horizon, not that either is better or worse within the budget on step-300 numbers. Second, single-checkpoint gaps below roughly two points are within the ± 1.5 seed band of these 30-problem sets and are treated as noise throughout (Section 7, Figure 10). Our claims rest on curve shapes, intervals, and cross-scale replication rather than on point differences of that size.

M Compute and Reproducibility

Guide: this section reports the total compute behind the paper (estimated from training logs, queue timestamps, and same-type anchors for external-machine runs), the measured anchors, and the environment and release plan.

Hardware and accounting method. The server-side experiments ran on 8×A800-80GB nodes, with training runs and evaluations scheduled in parallel lanes across the node's 8 GPUs. A minority of later runs (19 training replications and 44 checkpoint evaluations, marked as external in our records) were executed on a separate machine whose hardware and logs were not retained; they are priced by analogy with same-type runs (same scale, same lane configuration, same number of steps) at the anchors below and are A800-equivalent estimates. GPU-hours throughout are computed as (GPUs occupied per run) × (wall-clock hours per run), summed over runs. The estimate below is derived from measured anchors and extrapolates 4B/8B step times linearly in parameter count. A 300-step 1.7B training takes 2,617–2,632 s of `train_runtime` in the main-lane configuration (≈44 min) and 7,758–7,964 s in the queue era, where training lanes share the node with concurrent evaluations. One 30-problem, 12-sample benchmark evaluation takes 84–116 min (≈4.6 h per three-benchmark checkpoint). All numbers are estimates pending exact accounting from server timestamps; the script and per-run breakdown, with server runs anchor-priced and external runs analog-priced, are retained with our records.

Table 17: Compute accounting (estimates; see text for anchors, external-run pricing, and assumptions). One benchmark evaluation = 30 problems × 12 samples, thinking mode, up to 38,912 tokens.

Category	Count	Est. A800 GPU-hours
Training, 1.7B (100–500 steps)	66 runs	∼172
Training, 4B	16 runs	∼75
Training, 8B	6 runs	∼64
Evaluation, competition benchmarks	435 (145 ckpts × 3)	∼1,380
Evaluation, MATH500 (val_n=4)	16	∼160
Total	88 runs, 451 evaluations	∼1,850
of which external machine	19 runs, 132 evaluations	∼480

Evaluation, not training, accounts for just over 80% of the total. The 12-sample thinking-mode protocol with a 38,912-token budget is the cost driver of this line of work, and any reproduction can cut it proportionally by reducing samples per problem. The totals include one discarded buggy run (∼3 GPU-hours, Appendix F.2) and exclude development smoke runs (each under ten minutes) and exploratory runs not reported here. Wall-clock, the study spans about three weeks on shared nodes.

Environment and reproducibility. Python 3.10 with a PyTorch/DeepSpeed ZeRO-2/vLLM-colocate stack, FlashAttention-2, bf16 training, PEFT LoRA, and math-verify scoring; exact package pins are listed in Table 4 (Appendix D). All configuration switches used in this paper are gated and default-off (Section 3); seeds are listed in Appendix D; every number in the paper traces to the per-problem evaluation outputs aggregated in our compact results ledger, and all confidence intervals come from one canonical bootstrap script (10,000 problem resamples, fixed random-number-generator (RNG) seed) so that Table 10 is exactly regenerable. We intend to release the extended training code, evaluation pipeline, per-problem outputs, and the results ledger upon acceptance; the anonymous artifact set available during review maps to these categories in Appendix D.

N Run Registry

Guide: this section registers every training/evaluation run behind the paper, with its configuration delta and evaluated checkpoints, so that each table row and figure point can be traced to a specific run.

Table 18 lists all runs. Unless noted, every run uses the official OPSD configuration of Section 3 (seed 42); the registry marks seed-1/2 replications explicitly. The 1.7B `teacherrefresh` key is an early archive superseded by `teacherrefreshv2` (Appendix F.2); it enters no table or figure in the paper and is listed for completeness.

Table 18: Run registry: every training/evaluation run behind this paper. "Steps" lists evaluated checkpoints (three competition benchmarks each unless noted). MATH500 (val_n=4) was additionally evaluated for the runs marked in Section K. Source: `compact_results.json`.

Run key	Configuration	Steps
`bandpass` (1.7B)	DAW band-pass, ε=0.1	25, 50, 75, 100
`bandpass_long` (1.7B)	DAW band-pass, extended horizon	200, 300
`bandpass_s1` (1.7B)	DAW band-pass, seed 1	100
`bandpass_s2` (1.7B)	DAW band-pass, seed 2	100
`base` (1.7B)	no training	base only
`baseline` (1.7B)	official config, seed 42 (early ckpts)	25, 50, 75, 100
`baseline_150` (1.7B)	constant full LR stopped at step 150 (reverse dose unit)	150

(continued)

Run key	Configuration	Steps
`baseline_300_s1` (1.7B)	official config to step 300, seed 1	300
`baseline_300_s2` (1.7B)	official config to step 300, seed 2	300
`baseline_long` (1.7B)	official config, extended horizon	200, 300
`baseline_s1` (1.7B)	official config, seed 1 (step-100 multi-seed)	100
`baseline_s2` (1.7B)	official config, seed 2 (step-100 multi-seed)	100
`clip03` (1.7B)	clip λ=0.03	100
`clip10` (1.7B)	clip λ=0.10	100
`const2p5e6` (1.7B)	constant 2.5×10^{-6} (dose-matched)	100, 200, 300
`dawdecay` (1.7B)	DAW + cosine decay	300
`dawdecay_300_s1` (1.7B)	DAW + cosine decay, seed 1	300
`dawdecay_300_s2` (1.7B)	DAW + cosine decay, seed 2	300
`daweps03` (1.7B)	DAW band-pass, ε=0.3	100
`decay` (1.7B)	cosine decay inside the 100-step budget	100
`decay500` (1.7B)	cosine decay over a 500-step horizon	100, 200, 300, 500
`easygate` (1.7B)	DAW easy-gate (down-weight easiest decile)	100
`ema0099` (1.7B)	EMA teacher, decay 0.99	100, 200, 300
`ema0999` (1.7B)	EMA teacher, decay 0.999	300
`ema09995` (1.7B)	EMA teacher, decay 0.9995	100, 300
`emadecay` (1.7B)	EMA-0.99 teacher + cosine decay	100, 200, 300
`emateacher` (1.7B)	EMA teacher, decay 0.999	100
`fp32` (1.7B)	KL recomputed in fp32	100, 200, 300
`hardemph` (1.7B)	DAW hard-emphasis (up-weight hardest)	100
`hardgate` (1.7B)	DAW hard-gate (down-weight hardest decile)	100
`hybrid` (1.7B)	RLVR-hybrid v1, unhardened, α=0.1	100
`hybrida005` (1.7B)	RLVR-hybrid dose probe, α=0.05	100
`hybridv2` (1.7B)	RLVR-hybrid v2, hardened, α=0.02	100
`hybridv2_long` (1.7B)	RLVR-hybrid v2, extended horizon	200, 300
`hybridv2_s1` (1.7B)	RLVR-hybrid v2, seed 1	100
`hybridv2_s2` (1.7B)	RLVR-hybrid v2, seed 2	100
`jsd05` (1.7B)	JSD objective (β=0.5, fp32)	100, 300
`lineardecay` (1.7B)	linear LR decay to zero	100, 200, 300
`lrdecay` (1.7B)	cosine LR decay to zero over 300 steps	100, 200, 300
`lrdecay_300_s1` (1.7B)	cosine LR decay, seed 1	300
`lrdecay_300_s2` (1.7B)	cosine LR decay, seed 2	300
`matched300` (1.7B)	thinking-mode matched prompts, 300-step horizon	100, 200, 300
`noclip` (1.7B)	clip removed (λ=∞)	100, 200, 300
`nonthink_clip005` (1.7B)	non-thinking, sane clip 0.05	100
`nonthink_clip1e6` (1.7B)	non-thinking, official clip 10^{-6}	100
`refresh200` (1.7B)	teacher refresh every 200 steps	100, 200, 300

(continued)

Run key	Configuration	Steps
`refreshdecay` (1.7B)	teacher refresh every 100 steps + cosine decay	100, 200, 300
`rkl1` (1.7B)	reverse KL objective (β=1)	100, 200, 300
`symclip` (1.7B)	symmetric clip, constant LR	300
`symclip_300_s1` (1.7B)	symmetric clip, seed 1	300
`symclip_300_s2` (1.7B)	symmetric clip, seed 2	300
`teacherrefresh` (1.7B)	teacher refresh every 100 steps (v1 operator archive)	200, 300
`teacherrefreshv2` (1.7B)	teacher refresh every 100 steps	100, 200, 300
`4b_bandpass` (4B)	4B DAW band-pass	100
`4b_bandpass_long` (4B)	4B DAW, extended horizon	200, 300
`4b_base` (4B)	4B base model (no training)	base only
`4b_baseline` (4B)	4B official config	100
`4b_baseline_long` (4B)	4B extended horizon	200, 300
`4b_clip10` (4B)	4B clip λ=0.10	100
`4b_hybridv2` (4B)	4B RLVR-hybrid v2	100
`4b_lrdecay_long` (4B)	4B cosine LR decay to zero	100, 200, 300
`4b_lrdecay_s1` (4B)	4B cosine LR decay, seed 1	300
`4b_lrdecay_s2` (4B)	4B cosine LR decay, seed 2	300
`4b_refreshdecay` (4B)	4B refresh + decay combo	100, 200, 300
`4b_teacherrefresh` (4B)	4B teacher refresh every 100 steps	100, 200, 300
`8b_bandpass` (8B)	8B DAW band-pass	100
`8b_bandpass_long` (8B)	8B DAW, extended horizon	200, 300
`8b_base` (8B)	8B base model (no training)	base only
`8b_baseline` (8B)	8B official config	100
`8b_baseline_long` (8B)	8B extended horizon	200, 300
`8b_lrdecay` (8B)	8B cosine LR decay to zero	100, 200, 300

O PER-PROBLEM ANALYSIS

Guide: this section moves from difficulty buckets (Appendix H) to individual problems, using the retained per-problem records (90 problems × 12 samples).

Figure 14(a) histograms the per-problem accuracy change at step 300 relative to base. Under constant LR, 28 of 90 problems degrade and 19 improve (43 unchanged), mean -3.0pp; under cosine decay the balance inverts to 17 degraded and 28 improved, mean $+1.2$pp. The aggregate gap is a broad shift of the whole distribution, not a few outlier problems. Figure 14(b) scatters step-300 accuracy against base accuracy per problem: the constant-LR points fall systematically below the diagonal in the mid band while hugging it at both edges. Quantitatively, on the 45 problems the base solves partially ($0 <$ base $< 100\%$), constant LR loses 6.3pp on average while decay gains 1.7pp; on the 45 edge problems (base 0% or 100%) both schedules are flat ($+0.4$/$+0.7$pp). The collapse is thus a consolidation failure on partially-learned problems, and decay's rescue is concentrated in exactly the same band.

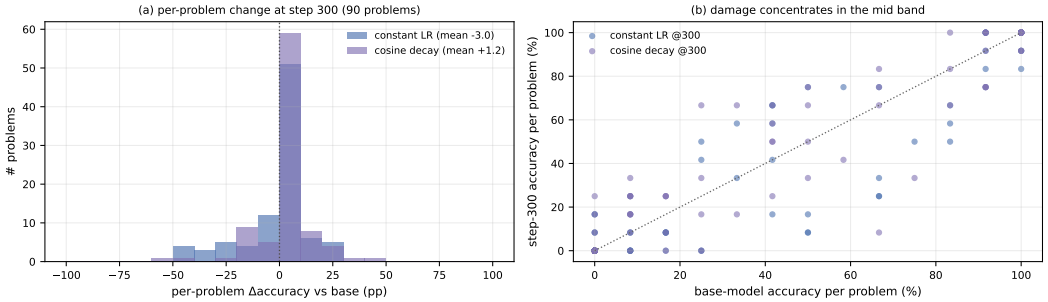


Figure 14: **Per-problem view of the collapse (1.7B, 90 problems).** (a) Histogram of per-problem accuracy change at step 300 vs. base: decay shifts the whole distribution right of the constant-LR run. (b) Step-300 vs. base accuracy per problem: constant LR sinks below the diagonal in the mid band; both schedules are flat on problems the base never or always solves.

P Per-Benchmark Trajectories

Guide: this section disaggregates the macro-average curves of the main text into the three competition benchmarks individually, at all three scales.

Figure 15 shows that the collapse shape (a peak inside the published budget, then a monotone decline to a terminal point near or below base) holds on each benchmark individually rather than being an artifact of averaging. Two per-benchmark notes. AIME25 is where the original paper's own 50–100-step decline was visible ($43.9 \rightarrow 41.1$); our extended curve confirms it was the onset (panel a2). HMMT25, the hardest set, shows the shallowest absolute decline at 1.7B simply because its base is low (23.1); the proportional loss is comparable. At 4B the refresh failure craters on all three benchmarks alike (panel b1), and at 8B the decay run ends at or above each benchmark's base line individually (panel b2), not just on average.

Q Response-Length Statistics

Guide: this section quantifies generation length across all variants and scales, the behavioral signature behind the failure gallery of Appendix A.

Figure 16 and Table 19 report mean response length. The variants quoted in the main text stay inside the 51–66k character band (mean over benchmarks) at every scale and step. This includes the fully collapsed constant-LR run, whose step-300 length (53.8k at 1.7B) is essentially the base model's (52.1k). The dose-controlled collapse carries no length signature at all, which is precisely why it is invisible from the training dashboard. The later probe runs widen the envelope only modestly (45.7k for the unclipped run, 68.4k for EMA-0.99). The refresh family is the opposite extreme: lengths grow monotonically with each merge, to 102k at 1.7B (refresh-100, step 300), 140k for refresh+decay, and 162k/214k at 4B. Length is therefore a selective alarm: it fires for the anchor-moving failures and stays silent for the optimizer-side drift. Blank cells in Table 19 reflect the same two causes described in Appendix R: checkpoints never retained for some steps, and auxiliary metrics (here, response length) not aggregated into the results ledger for runs that were evaluated. The measured lengths suffice for the claims they support.

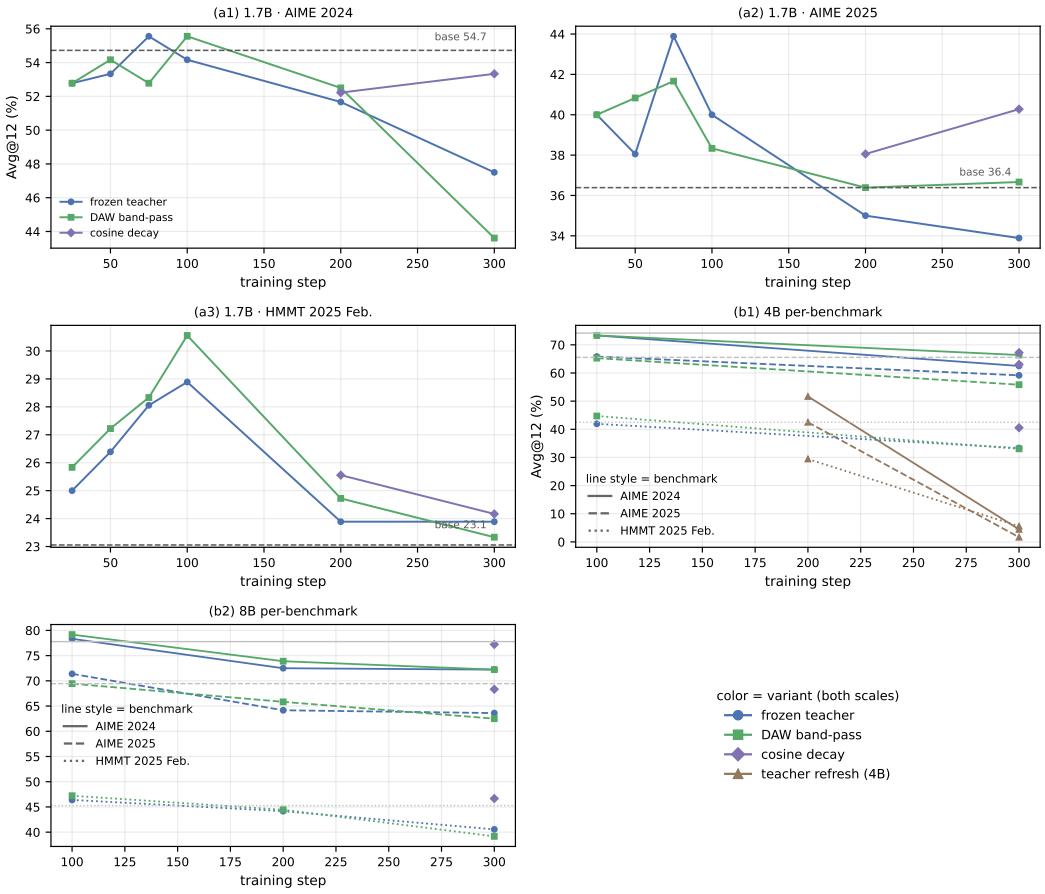


Figure 15: **Per-benchmark trajectories.** (a1–a3) 1.7B on AIME24, AIME25, HMMT25 individually: the peak-then-decline shape is per-benchmark, not an averaging artifact; dashed lines mark each benchmark's base score. (b1–b2) 4B and 8B per benchmark (line style), variants in color: the refresh collapse and the decay rescue replicate on every benchmark at both scales.

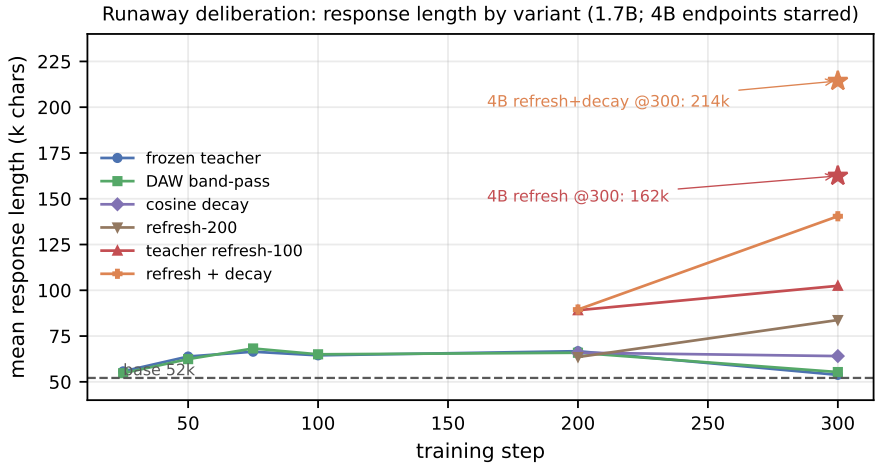


Figure 16: **Response length by variant.** Mean generation length (thousand characters, three-benchmark mean) vs. training step. Non-refresh variants overlap the base model; the refresh family inflates two- to four-fold (4B endpoints starred).

Table 19: Mean response length in thousand characters (12 samples per problem, thinking mode). Non-refresh variants stay within the 51–70k band at every scale; the refresh family inflates two- to four-fold. Source: `compact_results.json`.

Scale	Variant	step	AIME24	AIME25	HMMT25	Mean
1.7B	Base model (no training)	–	44.8	53.3	58.2	52.1
1.7B	Frozen teacher (baseline)	25	54.2	56.5	56.5	55.7
1.7B	Frozen teacher (baseline)	50	60.5	65.5	65.5	63.8
1.7B	Frozen teacher (baseline)	75	59.7	68.2	71.4	66.4
1.7B	Frozen teacher (baseline)	100	58.8	65.8	68.8	64.5
1.7B	Frozen teacher (baseline)	200	61.0	65.7	73.5	66.7
1.7B	Frozen teacher (baseline)	300	50.3	52.1	59.0	53.8
1.7B	DAW band-pass	100	61.7	65.0	68.4	65.0
1.7B	DAW band-pass	200	61.8	71.0	64.8	65.9
1.7B	DAW band-pass	300	54.5	51.8	59.7	55.4
1.7B	RLVR-hybrid v2 (hardened)	200	62.9	65.9	69.8	66.2
1.7B	RLVR-hybrid v2 (hardened)	300	50.5	55.7	59.4	55.2
1.7B	Cosine LR decay to zero	200	64.5	65.4	67.9	65.9
1.7B	Cosine LR decay to zero	300	59.3	63.6	69.3	64.1
1.7B	Linear LR decay to zero	300	63.2	67.3	66.7	65.7
1.7B	Constant 2.5×10^{-6} (dose-matched)	300	57.8	62.0	64.3	61.4
1.7B	Cosine decay over 500 steps	300	54.1	58.6	60.6	57.8
1.7B	Cosine decay over 500 steps	500	57.5	56.7	58.7	57.6
1.7B	DAW + cosine decay	300	59.5	63.2	65.5	62.7
1.7B	EMA + cosine decay	300	73.3	80.3	78.2	77.3
1.7B	EMA teacher (0.99)	300	66.6	68.9	69.8	68.4
1.7B	Symmetric clip (constant LR)	300	54.4	56.3	59.9	56.9
1.7B	Clip removed ($\lambda{=}\infty$)	300	47.5	44.4	45.1	45.7
1.7B	fp32 KL computation	300	55.4	60.1	63.5	59.6
1.7B	Matched thinking-mode prompts	300	48.9	50.8	51.3	50.3
1.7B	Teacher refresh every 100 steps	200	87.8	87.6	92.0	89.1
1.7B	Teacher refresh every 100 steps	300	93.2	117.8	96.3	102.4
1.7B	Teacher refresh every 200 steps	200	58.6	64.8	67.1	63.5
1.7B	Teacher refresh every 200 steps	300	80.1	83.7	87.4	83.8
1.7B	Refresh + LR decay (combo)	200	85.7	90.8	91.9	89.5
1.7B	Refresh + LR decay (combo)	300	135.1	140.9	145.2	140.4
4B	Base model (no training)	–	44.2	52.0	57.4	51.2
4B	Frozen teacher (baseline)	300	51.8	49.3	64.2	55.1
4B	Cosine LR decay to zero	300	48.7	56.0	59.3	54.6
4B	Teacher refresh every 100 steps	200	97.9	90.8	98.4	95.7
4B	Teacher refresh every 100 steps	300	167.1	163.9	156.5	162.5
4B	Refresh + LR decay (combo)	200	86.3	108.4	100.8	98.5
4B	Refresh + LR decay (combo)	300	212.7	217.0	213.2	214.3
8B	Base model (no training)	–	43.0	53.0	62.1	52.7
8B	Frozen teacher (baseline)	300	50.8	60.3	64.0	58.4
8B	Cosine LR decay to zero	300	49.9	62.7	66.7	59.8

R Full Scoring-Rule Results

Guide: this section reports, for every run and checkpoint in the paper, all three scoring rules our 12-sample protocol supports (mean accuracy, Avg@12; majority vote, Maj@12; and any-correct, Pass@12), per benchmark, so that any aggregate quoted in the text can be recomputed under any rule. Two kinds of entries are blank. Where a step is absent entirely, its checkpoint was never retained. Training wrote large numbers of checkpoints to the server, and evaluated checkpoints were not cleaned up in time; the resulting disk pressure meant those intermediate checkpoints were not kept for evaluation. Where a row reports Avg@12 but leaves Maj@12 and Pass@12 blank,

the evaluation was run but only the primary metric was aggregated into the results ledger; the auxiliary metrics remain in the raw evaluation outputs. This applies to all four tables in this section (Tables 20–23). The available data suffice for every claim and figure in the paper.

Tables 20–23 list every run and checkpoint. Two readings matter beyond completeness. First, the headline ordering is not a scoring-rule artifact: at step 300 the decay-over-constant gap survives majority vote and any-correct scoring on every row (e.g., 1.7B Maj $42.2 \to 50.0$, Pass $61.1 \to 65.6$), and the refresh collapse is so deep that it shows under every rule at once (1.7B Pass 23.3). Second, the three rules do not move together under drift: at 1.7B majority vote falls with mean accuracy from the step-75 peak ($51.1 \to 42.2$ Maj against $42.5 \to 35.1$ Avg), while at 8B the mean-accuracy drop (-5.4) barely touches majority vote ($70.0 \to 70.0$) or any-correct ($77.8 \to 78.9$). The drift first erodes per-sample reliability and only later flips the modal answer, consistent with the mid-band localization of Appendix H.

Table 20: Full scoring-rule results, 1.7B core families. Each run occupies three consecutive rows (AIME24, AIME25, HMMT25), reporting Avg@12, Maj@12, and Pass@12 in %. –: not evaluated. Source: `compact_results.json`.

Variant	step	Benchmark	Avg@12	Maj@12	Pass@12
Base model (no training)	–	AIME24	54.7	73.3	80.0
		AIME25	36.4	43.3	70.0
		HMMT25	23.1	30.0	46.7
Frozen teacher (baseline)	25	AIME24	52.8	66.7	73.3
		AIME25	40.0	53.3	63.3
		HMMT25	25.0	26.7	50.0
Frozen teacher (baseline)	50	AIME24	53.3	63.3	76.7
		AIME25	38.1	43.3	70.0
		HMMT25	26.4	30.0	50.0
Frozen teacher (baseline)	75	AIME24	55.6	66.7	76.7
		AIME25	43.9	53.3	73.3
		HMMT25	28.1	33.3	56.7
Frozen teacher (baseline)	100	AIME24	54.2	63.3	76.7
		AIME25	40.0	53.3	70.0
		HMMT25	28.9	40.0	56.7
Frozen teacher (baseline)	200	AIME24	51.7	66.7	80.0
		AIME25	35.0	43.3	60.0
		HMMT25	23.9	26.7	53.3
Frozen teacher (baseline)	300	AIME24	47.5	60.0	76.7
		AIME25	33.9	43.3	60.0
		HMMT25	23.9	23.3	46.7
DAW band-pass	25	AIME24	52.8	66.7	76.7
		AIME25	40.0	46.7	70.0
		HMMT25	25.8	30.0	46.7
DAW band-pass	50	AIME24	54.2	60.0	76.7
		AIME25	40.8	53.3	66.7
		HMMT25	27.2	30.0	56.7
DAW band-pass	75	AIME24	52.8	73.3	76.7
		AIME25	41.7	53.3	73.3
		HMMT25	28.3	33.3	53.3
DAW band-pass	100	AIME24	55.6	73.3	76.7
		AIME25	38.3	50.0	70.0
		HMMT25	30.6	36.7	56.7
DAW band-pass	200	AIME24	52.5	66.7	80.0
		AIME25	36.4	53.3	73.3
		HMMT25	24.7	36.7	40.0
DAW band-pass	300	AIME24	43.6	53.3	76.7

(continued)

Variant	step	Benchmark	Avg@12	Maj@12	Pass@12
		AIME25	36.7	43.3	56.7
		HMMT25	23.3	33.3	43.3
RLVR-hybrid v2 (hardened, α=0.02)	100	AIME24	55.3	70.0	80.0
		AIME25	40.3	56.7	70.0
		HMMT25	29.2	36.7	46.7
RLVR-hybrid v2 (hardened)	200	AIME24	50.6	66.7	76.7
		AIME25	36.7	46.7	70.0
		HMMT25	24.2	30.0	53.3
RLVR-hybrid v2 (hardened)	300	AIME24	45.0	60.0	76.7
		AIME25	33.6	36.7	53.3
		HMMT25	21.7	26.7	36.7
Cosine LR decay to zero	100	AIME24	54.7	66.7	76.7
		AIME25	39.4	53.3	73.3
		HMMT25	26.4	30.0	50.0
Cosine LR decay to zero	200	AIME24	52.2	63.3	73.3
		AIME25	38.1	53.3	73.3
		HMMT25	25.6	30.0	53.3
Cosine LR decay to zero	300	AIME24	53.3	63.3	73.3
		AIME25	40.3	60.0	66.7
		HMMT25	24.2	26.7	56.7
Linear LR decay to zero	100	AIME24	54.2	66.2	81.3
		AIME25	39.8	48.7	66.9
		HMMT25	28.9	32.5	53.7
Linear LR decay to zero	200	AIME24	52.5	66.7	76.7
		AIME25	37.5	46.7	66.7
		HMMT25	29.7	33.3	53.3
Linear LR decay to zero	300	AIME24	52.2	66.7	76.7
		AIME25	40.0	43.3	70.0
		HMMT25	25.3	33.3	50.0
Constant 2.5×10^{-6} (dose-matched)	100	AIME24	53.7	67.6	81.5
		AIME25	40.3	49.8	67.9
		HMMT25	28.6	33.1	54.4
Constant 2.5×10^{-6} (dose-matched)	200	AIME24	55.8	68.3	83.6
		AIME25	39.5	51.5	67.2
		HMMT25	28.1	32.6	53.9
Constant 2.5×10^{-6} (dose-matched)	300	AIME24	51.1	63.3	76.7
		AIME25	39.2	53.3	66.7
		HMMT25	26.1	33.3	50.0
Cosine decay over 500 steps	100	AIME24	54.7	70.1	82.5
		AIME25	39.6	49.5	68.4
		HMMT25	29.1	32.7	54.3
Cosine decay over 500 steps	200	AIME24	53.3	66.8	80.6
		AIME25	40.7	52.2	67.3
		HMMT25	28.1	32.5	54.2
Cosine decay over 500 steps	300	AIME24	50.6	63.3	73.3
		AIME25	36.9	56.7	63.3
		HMMT25	24.2	33.3	43.3
Cosine decay over 500 steps	500	AIME24	45.6	60.0	73.3
		AIME25	34.7	46.7	56.7

(continued)

Variant	step	Benchmark	Avg@12	Maj@12	Pass@12
		HMMT25	24.7	30.0	53.3
Cosine decay within 100-step budget	100	AIME24	56.1	63.3	80.0
		AIME25	38.3	53.3	66.7
		HMMT25	24.7	26.7	50.0
DAW + cosine decay	300	AIME24	53.6	66.7	80.0
		AIME25	41.9	50.0	76.7
		HMMT25	27.5	33.3	60.0
EMA + cosine decay	100	AIME24	54.2	67.2	81.4
		AIME25	40.8	52.3	70.5
		HMMT25	27.8	31.6	52.0
EMA + cosine decay	200	AIME24	52.6	64.5	75.5
		AIME25	38.2	46.8	62.4
		HMMT25	25.6	27.6	51.2
EMA + cosine decay	300	AIME24	53.6	66.7	76.7
		AIME25	38.9	50.0	56.7
		HMMT25	25.3	26.7	43.3

Table 21: Full scoring-rule results, 1.7B probe and seed runs (same per-run three-row convention as Table 20). Runs shown as Avg/–/– have only Avg@12 synced to the ledger so far.

Variant	step	Benchmark	Avg@12	Maj@12	Pass@12
Constant LR, 150 steps (reverse unit)	150	AIME24	48.3	66.7	76.7
		AIME25	38.6	53.3	66.7
		HMMT25	23.3	23.3	50.0
Frozen teacher (baseline), seed 1	300	AIME24	47.8	63.3	73.3
		AIME25	31.1	43.3	56.7
		HMMT25	24.2	23.3	53.3
Frozen teacher (baseline), seed 2	300	AIME24	47.8	66.7	70.0
		AIME25	33.1	43.3	53.3
		HMMT25	21.9	26.7	50.0
Cosine LR decay to zero, seed 1	300	AIME24	49.7	56.7	80.0
		AIME25	37.8	46.7	70.0
		HMMT25	26.9	26.7	56.7
Cosine LR decay to zero, seed 2	300	AIME24	54.4	66.7	76.7
		AIME25	40.0	46.7	63.3
		HMMT25	27.5	36.7	53.3
DAW + cosine decay, seed 1	300	AIME24	52.2	66.7	76.7
		AIME25	38.3	53.3	66.7
		HMMT25	26.1	26.7	46.7
DAW + cosine decay, seed 2	300	AIME24	51.4	60.0	76.7
		AIME25	39.2	53.3	73.3
		HMMT25	26.4	33.3	53.3
DAW band-pass	100	AIME24	54.2	70.0	76.7
		AIME25	38.9	53.3	70.0
		HMMT25	28.1	33.3	56.7

(continued)

Variant	step	Benchmark	Avg@12	Maj@12	Pass@12
DAW band-pass	100	AIME24	55.0	70.0	80.0
		AIME25	42.2	60.0	73.3
		HMMT25	28.1	40.0	53.3
RLVR-hybrid v2 (hardened)	100	AIME24	52.8	70.0	76.7
		AIME25	38.9	56.7	70.0
		HMMT25	27.5	33.3	50.0
RLVR-hybrid v2 (hardened)	100	AIME24	50.8	73.3	76.7
		AIME25	37.5	46.7	66.7
		HMMT25	24.7	30.0	46.7
Symmetric clip (constant LR)	300	AIME24	50.3	70.0	76.7
		AIME25	31.9	43.3	50.0
		HMMT25	25.0	26.7	50.0
Symmetric clip (constant LR), seed 1	300	AIME24	45.3	60.0	76.7
		AIME25	36.1	43.3	56.7
		HMMT25	21.9	26.7	46.7
Symmetric clip (constant LR), seed 2	300	AIME24	47.8	60.0	76.7
		AIME25	35.8	43.3	60.0
		HMMT25	21.4	20.0	53.3
Clip λ=0.03 (tighter)	100	AIME24	51.9	66.7	76.7
		AIME25	40.8	50.0	63.3
		HMMT25	30.6	43.3	56.7
Clip λ=0.10 (looser)	100	AIME24	48.3	70.0	76.7
		AIME25	35.6	43.3	66.7
		HMMT25	23.6	30.0	50.0
Clip removed (λ=∞)	100	AIME24	46.7	60.0	76.7
		AIME25	32.5	40.0	56.7
		HMMT25	20.6	23.3	46.7
Clip removed (λ=∞)	200	AIME24	44.7	63.3	76.7
		AIME25	32.8	43.3	63.3
		HMMT25	20.3	20.0	46.7
Clip removed (λ=∞)	300	AIME24	46.9	66.7	76.7
		AIME25	33.1	40.0	56.7
		HMMT25	21.1	23.3	53.3
fp32 KL computation	100	AIME24	53.6	65.6	76.6
		AIME25	40.8	49.6	65.1
		HMMT25	28.5	31.2	52.9
fp32 KL computation	200	AIME24	52.8	65.5	76.8
		AIME25	35.8	44.3	62.5
		HMMT25	24.9	28.3	51.4
fp32 KL computation	300	AIME24	51.4	63.3	80.0
		AIME25	31.7	43.3	56.7
		HMMT25	23.1	26.7	46.7
Matched thinking-mode prompts	100	AIME24	42.8	56.7	76.7
		AIME25	34.2	50.0	70.0
		HMMT25	22.5	30.0	53.3
Matched thinking-mode prompts	200	AIME24	39.5	52.8	64.1
		AIME25	31.1	42.2	56.5
		HMMT25	22.5	26.1	47.6

(continued)

Variant	step	Benchmark	Avg@12	Maj@12	Pass@12
Matched thinking-mode prompts	300	AIME24	35.0	53.3	73.3
		AIME25	25.6	46.7	56.7
		HMMT25	19.7	26.7	46.7
Non-thinking, clip 0.05	100	AIME24	10.8	23.3	43.3
		AIME25	6.9	13.3	20.0
		HMMT25	2.8	6.7	10.0
Non-thinking, clip 10^{-6} (official)	100	AIME24	11.4	13.3	30.0
		AIME25	5.6	10.0	23.3
		HMMT25	2.2	3.3	13.3
EMA teacher (0.99)	100	AIME24	55.6	73.3	80.0
		AIME25	42.5	56.7	63.3
		HMMT25	29.7	33.3	53.3
EMA teacher (0.99)	300	AIME24	43.3	60.0	73.3
		AIME25	33.3	40.0	60.0
		HMMT25	20.8	23.3	46.7
EMA teacher (0.999)	100	AIME24	52.2	66.7	73.3
		AIME25	38.1	53.3	63.3
		HMMT25	28.1	36.7	46.7
EMA teacher (0.999)	300	AIME24	42.2	66.7	76.7
		AIME25	33.3	43.3	53.3
		HMMT25	23.1	20.0	50.0
EMA teacher (0.9995)	100	AIME24	54.7	63.3	73.3
		AIME25	38.6	46.7	66.7
		HMMT25	25.3	33.3	53.3
EMA teacher (0.9995)	300	AIME24	44.7	60.0	73.3
		AIME25	33.3	43.3	53.3
		HMMT25	20.0	26.7	43.3
DAW easy-gate	100	AIME24	54.7	66.7	76.7
		AIME25	41.1	50.0	76.7
		HMMT25	27.5	33.3	50.0
DAW hard-emphasis	100	AIME24	53.1	66.7	76.7
		AIME25	40.8	53.3	66.7
		HMMT25	27.2	30.0	46.7
DAW hard-gate	100	AIME24	51.4	60.0	76.7
		AIME25	39.2	50.0	70.0
		HMMT25	26.1	30.0	53.3
DAW ε=0.3	100	AIME24	54.4	66.7	80.0
		AIME25	39.7	46.7	66.7
		HMMT25	29.2	30.0	56.7
RLVR-hybrid v1 (unhardened, α=0.1)	100	AIME24	47.8	66.7	80.0
		AIME25	38.6	46.7	63.3
		HMMT25	22.5	23.3	53.3
RLVR-hybrid dose probe (α=0.05)	100	AIME24	54.2	66.7	76.7
		AIME25	41.1	60.0	73.3
		HMMT25	28.3	36.7	50.0
JSD (β=0.5)	100	AIME24	45.8	63.3	76.7
		AIME25	36.1	43.3	60.0
		HMMT25	22.8	23.3	46.7
JSD (β=0.5)	300	AIME24	48.3	66.7	73.3

(continued)

Variant	step	Benchmark	Avg@12	Maj@12	Pass@12
		AIME25	34.2	43.3	70.0
		HMMT25	23.3	23.3	50.0
Reverse KL (β=1)	100	AIME24	44.4	63.3	76.7
		AIME25	35.0	50.0	66.7
		HMMT25	18.3	23.3	46.7
Reverse KL (β=1)	200	AIME24	46.7	56.7	76.7
		AIME25	37.2	43.3	63.3
		HMMT25	26.1	36.7	53.3
Reverse KL (β=1)	300	AIME24	47.2	60.0	76.7
		AIME25	37.8	53.3	60.0
		HMMT25	23.1	23.3	50.0
Teacher refresh every 100 steps (early archive; not used)	200	AIME24	46.7	60.0	76.7
		AIME25	39.4	53.3	66.7
		HMMT25	22.8	20.0	56.7
Teacher refresh every 100 steps (early archive; not used)	300	AIME24	46.7	60.0	76.7
		AIME25	39.4	53.3	66.7
		HMMT25	22.8	20.0	56.7
Teacher refresh every 100 steps	100	AIME24	53.8	67.5	79.6
		AIME25	40.7	51.6	65.8
		HMMT25	28.6	33.4	52.5
Teacher refresh every 100 steps	200	AIME24	46.1	60.0	76.7
		AIME25	35.6	40.0	63.3
		HMMT25	20.8	23.3	40.0
Teacher refresh every 100 steps	300	AIME24	8.9	16.7	33.3
		AIME25	5.0	13.3	16.7
		HMMT25	3.3	10.0	20.0
Teacher refresh every 200 steps	100	AIME24	53.8	67.7	81.4
		AIME25	41.5	51.6	70.8
		HMMT25	30.4	35.9	56.2
Teacher refresh every 200 steps	200	AIME24	55.0	66.7	76.7
		AIME25	39.7	53.3	63.3
		HMMT25	26.1	30.0	50.0
Teacher refresh every 200 steps	300	AIME24	43.3	63.3	70.0
		AIME25	34.2	43.3	60.0
		HMMT25	20.3	23.3	46.7
Refresh + LR decay (combo)	100	AIME24	51.9	64.2	75.5
		AIME25	38.5	47.8	64.6
		HMMT25	27.4	31.2	52.4
Refresh + LR decay (combo)	200	AIME24	48.3	60.0	73.3
		AIME25	36.7	50.0	63.3
		HMMT25	25.6	30.0	53.3
Refresh + LR decay (combo)	300	AIME24	21.1	30.0	50.0
		AIME25	13.9	26.7	36.7
		HMMT25	8.3	13.3	30.0

Table 22: Full scoring-rule results, 4B (same per-run three-row convention as Table 20).

Variant	step	Benchmark	Avg@12	Maj@12	Pass@12
4B base model (no training)	–	AIME24	74.2	83.3	86.7
		AIME25	65.6	70.0	83.3
		HMMT25	42.5	46.7	66.7
4B frozen teacher (baseline)	100	AIME24	73.3	80.0	90.0
		AIME25	65.8	76.7	80.0
		HMMT25	41.9	53.3	63.3
4B frozen teacher (baseline)	200	AIME24	68.6	81.4	84.7
		AIME25	62.1	74.7	85.3
		HMMT25	38.4	45.3	60.9
4B frozen teacher (baseline)	300	AIME24	62.5	76.7	80.0
		AIME25	59.2	70.0	83.3
		HMMT25	33.3	36.7	56.7
4B dAW band-pass	100	AIME24	73.3	80.0	86.7
		AIME25	65.3	76.7	86.7
		HMMT25	44.7	56.7	63.3
4B dAW band-pass	200	AIME24	70.2	82.6	88.3
		AIME25	60.7	73.2	82.9
		HMMT25	38.9	48.5	61.8
4B dAW band-pass	300	AIME24	66.4	80.0	83.3
		AIME25	55.8	70.0	86.7
		HMMT25	33.1	33.3	60.0
4B cosine LR decay to zero	100	AIME24	72.8	84.1	90.4
		AIME25	66.5	77.9	86.9
		HMMT25	42.7	52.7	63.5
4B cosine LR decay to zero	200	AIME24	70.2	82.5	86.1
		AIME25	63.7	75.7	84.6
		HMMT25	39.4	47.3	61.5
4B cosine LR decay to zero	300	AIME24	67.2	80.0	83.3
		AIME25	63.1	73.3	86.7
		HMMT25	40.6	46.7	63.3
4B cosine LR decay to zero	300	AIME24	69.7	80.0	83.3
		AIME25	62.5	73.3	83.3
		HMMT25	41.4	53.3	60.0
4B cosine LR decay to zero	300	AIME24	71.7	80.0	80.0
		AIME25	60.0	73.3	80.0
		HMMT25	42.8	56.7	73.3
4B teacher refresh every 100 steps	100	AIME24	73.2	85.2	90.7
		AIME25	66.1	78.8	86.7
		HMMT25	41.8	51.4	63.1
4B teacher refresh every 100 steps	200	AIME24	51.7	66.7	76.7
		AIME25	42.5	60.0	70.0
		HMMT25	29.4	40.0	53.3
4B teacher refresh every 100 steps	300	AIME24	4.4	16.7	23.3
		AIME25	1.7	16.7	20.0
		HMMT25	5.6	16.7	26.7
4B refresh + LR decay (combo)	100	AIME24	70.5	82.6	86.6
		AIME25	64.4	75.3	86.1
		HMMT25	41.7	51.5	63.9

(continued)

Variant	step	Benchmark	Avg@12	Maj@12	Pass@12
4B refresh + LR decay (combo)	200	AIME24	53.9	73.3	90.0
		AIME25	50.0	73.3	80.0
		HMMT25	28.6	46.7	60.0
4B refresh + LR decay (combo)	300	AIME24	1.9	16.7	23.3
		AIME25	0.8	6.7	6.7
		HMMT25	1.4	6.7	6.7
4B rLVR-hybrid v2 (hardened, α=0.02)	100	AIME24	76.7	80.0	86.7
		AIME25	66.9	76.7	80.0
		HMMT25	44.4	56.7	63.3
4B clip λ=0.10 (looser)	100	AIME24	71.4	80.0	83.3
		AIME25	63.1	76.7	86.7
		HMMT25	42.5	50.0	66.7

Table 23: Full scoring-rule results, 8B (same per-run three-row convention as Table 20).

Variant	step	Benchmark	Avg@12	Maj@12	Pass@12
8B base model (no training)	–	AIME24	77.8	80.0	90.0
		AIME25	69.4	76.7	83.3
		HMMT25	45.3	53.3	60.0
8B frozen teacher (baseline)	100	AIME24	78.3	83.3	90.0
		AIME25	71.4	80.0	83.3
		HMMT25	46.4	56.7	66.7
8B frozen teacher (baseline)	200	AIME24	72.5	83.3	90.0
		AIME25	64.2	76.7	83.3
		HMMT25	44.2	46.7	66.7
8B frozen teacher (baseline)	300	AIME24	72.2	80.0	90.0
		AIME25	63.6	80.0	83.3
		HMMT25	40.6	50.0	63.3
8B dAW band-pass	100	AIME24	79.2	80.0	93.3
		AIME25	69.4	80.0	80.0
		HMMT25	47.2	63.3	70.0
8B dAW band-pass	200	AIME24	73.9	80.0	90.0
		AIME25	65.8	76.7	86.7
		HMMT25	44.4	50.0	70.0
8B dAW band-pass	300	AIME24	72.2	83.3	86.7
		AIME25	62.5	70.0	83.3
		HMMT25	39.2	53.3	63.3
8B cosine LR decay to zero	100	AIME24	75.9	83.5	91.6
		AIME25	71.3	82.7	90.9
		HMMT25	45.4	55.6	67.1
8B cosine LR decay to zero	200	AIME24	75.9	83.8	92.1
		AIME25	69.1	80.5	86.4
		HMMT25	46.2	56.9	67.2
8B cosine LR decay to zero	300	AIME24	77.2	80.0	90.0
		AIME25	68.3	80.0	83.3
		HMMT25	46.7	60.0	63.3