When Should the Teacher Move? Temporal Coupling and Stability in Self On-Policy Distillation

Anonymous authors
Paper under double-blind review

Abstract

Self on-policy distillation trains a student policy against a teacher derived from its own parameter history, yet the teacher's update schedule—which governs the *temporal coupling* between teacher and student—has not been systematically studied as a stability variable. Through a controlled schedule sweep on Qwen3-8B, we establish that *isolation periods*—complete teacher freezing between updates—are the key structural property enabling stable learning, not teacher age, and we introduce a diagnostic framework of temporal KL structure, refresh shock, and length-tail risk to characterize training dynamics. We further identify *state-oblivious collapse*: the best short-horizon fixed schedule ($M{=}50$) catastrophically fails under long-horizon training because a clock-driven refresh can copy a transiently drifting student into the teacher in a single, irreversible step—a failure mode invisible under short-horizon evaluation and mechanistically distinct from EMA's chronic contamination. To address this, we propose *Consolidation-Gated Teacher Refresh* (CGTR), which preserves isolation periods while gating each refresh on joint evidence of reward improvement and length-tail safety, ensuring every teacher movement responds to genuine student consolidation rather than a clock signal. With a single shared parameter set and no per-dataset retuning, CGTR achieves **zero collapse** and the best final score on all four tasks (Chemistry, Biology, Physics, ToolUse), self-regulating its refresh frequency to each task's learning dynamics.

1 Introduction

Large language model post-training increasingly relies on on-policy learning (Schulman et al., 2017; Ouyang et al., 2022; Shao et al., 2024; Guo et al., 2025), where the model improves by sampling from its own policy rather than imitating static demonstrations. On-policy distillation extends this principle by providing the student with dense token-level supervision from a teacher during on-policy rollouts (Agarwal et al., 2024), reducing the distribution mismatch inherent in classical off-policy distillation. A natural extension is to derive this teacher from the model itself rather than an external source, as recent self-distillation methods do by constructing teacher signals from the same model under richer contexts—privileged reasoning, environment feedback, or compression-oriented instructions (Zhao et al., 2026; Hübotter et al., 2026; Li et al., 2026; Yang et al., 2026; Sang et al., 2026).

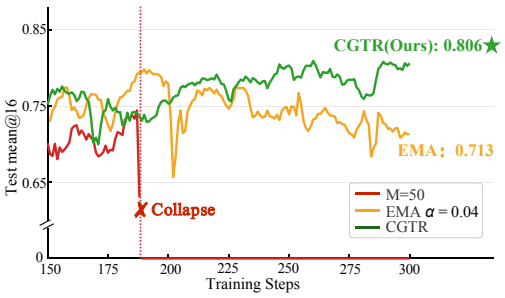


Figure 1: Long-horizon results (300 steps, Chemistry, steps 150–300 shown). $M{=}50$ collapses catastrophically at step 189 (test mean@16 $\rightarrow$ 0); EMA $\alpha{=}0.04$ plateaus at 0.713; **CGTR (ours) achieves the best final score of 0.806** with zero collapse.

Unlike a fixed external teacher, a self-teacher co-evolves with the student it trains, introducing a fundamental tension that prior work has not systematically studied. On the one hand, the teacher must follow the student over time to support bootstrapping and remain relevant. On the other hand,

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.

if the teacher parameters track the student too closely, it loses *temporal independence*, weakening the corrective distillation signal because the teacher no longer retains a historical parameter state capable of counteracting student drift. The teacher must therefore serve as a robust anchor capable of rescuing the student from suboptimal distributions—not a passive copy that ratifies whatever the student does. We call the degree to which the teacher tracks the student *temporal coupling*, and ask: *how does temporal coupling shape training stability and long-horizon learning?*

Each canonical mechanism for controlling temporal coupling embodies a distinct failure mode, but prior work treats these mechanisms as implementation choices rather than stability-critical variables. Tight coupling (e.g., updating the teacher at every step) removes the temporal separation between teacher and student parameters, causing *reference collapse*: the corrective signal from a historical teacher vanishes, leaving training unconstrained against self-reinforcing student drift. EMA (Polyak & Juditsky, 1992; Tarvainen & Valpola, 2017; Grill et al., 2020) mitigates this by diluting each update, but continuous soft tracking means that student drift accumulates in the teacher without bound over long training horizons, leading to *chronic teacher contamination*. Sparse hard refresh (Mnih et al., 2015) avoids both failure modes in the short run: by periodically copying the student into the teacher and then freezing it for many steps, creating *isolation periods* during which student fluctuations cannot contaminate the teacher. These isolation periods help preserve temporal independence by preventing the teacher from immediately tracking student fluctuations. Our 100-step schedule sweep attributes their stability benefit to this effect rather than teacher age alone (Section 3.3), with long-horizon experiments providing further evidence (Figure 1).

Yet isolation periods alone are not sufficient for long-horizon stability when the refresh schedule is fixed. A fixed interval is oblivious to student state: the teacher is updated at every M-th step regardless of whether the student is consolidating or drifting. If a hard refresh fires at the wrong moment (e.g., during a transient entropy spike or a reward plateau), the teacher absorbs corrupted behavior in a single, irreversible copy. Once copied, the contaminated teacher continuously reinforces the corrupted state through subsequent distillation, and there is no partial recovery path thereafter.

This analysis motivates a schedule that preserves the structural isolation advantage of hard refresh while conditioning each refresh on observable evidence that the student has genuinely consolidated. We propose *Consolidation-Gated Teacher Refresh* (CGTR), which maintains a minimum isolation period and gates each refresh on the joint satisfaction of a reward-improvement condition and a length-tail safety condition. When all conditions are met, the refresh is corrective; when any fails, the teacher remains frozen regardless of elapsed time. This design eliminates the need to pre-specify a refresh interval while ensuring that every teacher movement is strictly a response to genuine progress rather than a response to a clock signal alone.

Our contributions are as follows:

- We identify *isolation period* (complete teacher freezing between updates) as the key structural property enabling stable self on-policy distillation, supported by a systematic schedule sweep and a diagnostic framework of temporal KL, refresh shock, and length-tail risk.
- We characterize *state-oblivious collapse*, a long-horizon failure mode of fixed-interval hard refresh in which a single ill-timed update causes irreversible capability loss invisible at short training horizons.
- We propose *Consolidation-Gated Teacher Refresh* (CGTR), which gates each refresh on reward improvement and length-tail safety. CGTR achieves zero collapse and the best score on all four tasks (Chemistry, Biology, Physics, ToolUse) with a single parameter set.

2 Related Work

On-Policy Distillation and Self-Distillation. Knowledge distillation trains a student to match a teacher's predictive distribution (Hinton et al., 2015), which sequence-level distillation adapts for autoregressive generation by having the student imitate teacher-generated sequences (Kim & Rush, 2016). In reinforcement learning, policy distillation transfers behavior between teacher and student policies (Rusu et al., 2015), and on-policy distillation extends this idea by training the student on its own sampled trajectories with dense teacher supervision (Agarwal et al., 2024). Iterative self-training showed that a model can bootstrap its own training signal by generating and filtering self-

produced reasoning (Zelikman et al., 2022). Recent self-distillation methods further remove the need for a stronger external teacher by deriving teacher signals from the same model under richer contexts, such as privileged reasoning information, environment feedback, reward-verifiable directions, or compression-oriented instructions (Zhao et al., 2026; Hübotter et al., 2026; Yang et al., 2026; Sang et al., 2026). These methods demonstrate that self-teachers can provide dense supervision for post-training without a separate teacher model. Our work is orthogonal: rather than studying how to construct the self-teacher signal, we study how the self-teacher should move over training time.

Stability and Failure Modes in Self-Distillation. Recent work has begun to analyze why self-distillation succeeds or fails. Li et al. (2026) show that teacher–student compatibility and thinking-pattern alignment critically affect distillation outcomes. Other works identify information leakage, unstable long-term training, and the need for reward-grounded or periodically synchronized teachers as key concerns (Yang et al., 2026; Sang et al., 2026). These studies treat teacher movement as a fixed implementation choice, and different mechanisms—hard refresh, fixed teachers, and EMA (Mnih et al., 2015; Polyak & Juditsky, 1992; Tarvainen & Valpola, 2017)—are adopted without analyzing whether the optimal interval is task-dependent. We address this gap by treating temporal coupling as a first-class stability variable, identifying what structural property enables long-horizon consolidation, and showing that the optimal hard refresh interval varies with task characteristics—motivating CGTR as an adaptive schedule that generalizes across settings.

3 Analysis of Teacher Update Schedules

3.1 Preliminary: Setup and Update Rules

Let π_θ denote the *student policy* parameterized by θ, and let π_ϕ denote the *self-teacher policy* parameterized by ϕ, which is derived from a historical version of θ. For a student rollout y and feedback context f, we optimize:

$$\mathcal{L}_t(\theta) = \sum_{j=1}^{|y|} \mathrm{JSD}(\pi_\theta(\cdot \mid x, y_{<j}) \,\|\, \mathrm{stopgrad}\, \pi_\phi(\cdot \mid x, f, y_{<j}))\,, \tag{1}$$

where f is the feedback context derived from rollout outcomes. We approximate JSD with top-k logits and importance-sampling correction (Appendix A).

Two canonical update rules define the temporal coupling between ϕ and θ:

Periodic hard refresh. The teacher parameters are updated to match the student every M steps:

$$\phi_t = \begin{cases} \theta_t & \text{if } t \bmod M = 0, \\ \phi_{t-1} & \text{otherwise.} \end{cases} \tag{2}$$

Between consecutive refreshes, π_ϕ is completely frozen, forming an *isolation period* during which student fluctuations cannot contaminate the teacher.

Exponential moving average (EMA). The teacher is updated continuously at every step:

$$\phi_t = (1-\alpha)\,\phi_{t-1} + \alpha\,\theta_t, \quad 0 < \alpha \leq 1. \tag{3}$$

EMA has no isolation period: the teacher absorbs a fraction α of student updates at every step, bounding per-step contamination but allowing drift to accumulate over long horizons.

To control for teacher age, we derive a lag-aligned EMA baseline (α=0.04; Appendix E). Table 1 summarizes the schedule sweep results (setup in Section 5.1; terminology in Appendix F).

3.2 Fast Synchronization Causes Reference Collapse

At M=1, teacher and student parameters are virtually identical, eliminating temporal independence and the corrective signal of a historical teacher. Without this anchor, sequence lengths grow unchecked, exceeding 8,000 at step 10 and reaching 16,908 by step 14, forcing early termination with no test score. At M=2, the teacher lags by only one step, providing negligible temporal independence. Around the reference-collapse marker at step 87, mean response length rises sharply

Table 1: Schedule sweep results on Qwen3-8B / SciKnowEval Chemistry (100 steps). **Test m@16**: test mean@16 accuracy over 16 samples (primary); **B@4**: bootstrap-estimated best-of-4 test accuracy; **Ent.**: token entropy (training mean / std); **Rew.**: mean rollout reward; **KL**: mean student–teacher KL; **Collapse**: denotes the mechanism discussed in Section 3.2(— = none observed). ‡M1 terminated at step 16; no test score available. Bold marks the highest test scores only.

Schedule	Test m@16	B@4	Ent. mean	Ent. std	Rew. mean	KL mean	Collapse
$M = 1^{\ddagger}$	N/A	N/A	$\sim$1.05	—	—	—	step 10
$M = 2$	0.225	0.276	0.900	1.063	0.504	0.031	step 87
$M = 3$	0.673	0.740	1.715	1.167	0.607	0.085	—
$M = 4$	0.656	0.718	1.536	1.131	0.595	0.053	—
$M = 5$	0.653	0.723	2.001	1.091	0.629	0.064	—
$M = 8$	0.596	0.629	1.413	0.713	0.597	0.004	—
$M = 10$	0.669	0.737	0.545	0.278	0.614	0.025	—
$M = 20$	0.629	0.689	0.595	0.119	0.598	0.004	—
$M = 50$	**0.755**	**0.827**	0.499	0.048	0.629	0.005	—
Fixed	0.697	0.751	0.437	0.097	0.637	0.008	—
EMA $\alpha = 0.1$	0.645	0.689	1.417	0.694	0.613	0.005	—
EMA $\alpha = 0.04$	0.713	0.782	0.727	0.142	0.665	0.022	—

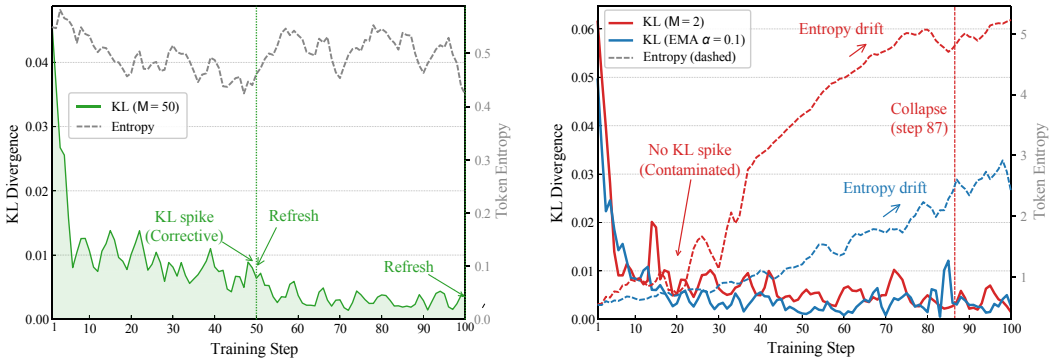


Figure 2: Illustration of two refresh dynamics. **(Left)** A *corrective refresh* (e.g., M=50) produces a transient KL spike as the teacher absorbs genuine improvement; the student then re-adapts under the new target. **(Right)** A *contaminating* pattern (low M or EMA with high α) produces no correction spike and is followed by gradual entropy growth as student drift propagates through the teacher.

toward the 8,192-token cap (Figure 5); the final test score is 0.225. The diagnostic KL mean at M=2 is only 0.031, consistent with tight coupling and a weak historical parameter constraint rather than a strong independent anchor. This *loss of temporal independence* underlies both failures—a pattern we term *reference collapse*. Accuracy-based collapse is assessed separately (Appendix A.6).

Diagnostic: KL temporal structure. What matters is not the KL level but its *temporal structure*: a corrective refresh produces a transient KL spike as the student re-adapts to the new teacher; a contaminating pattern shows no spike but subsequent entropy growth. Figure 2 contrasts these two patterns, providing a diagnostic that distinguishes healthy from contaminating training dynamics.

3.3 Sparse Hard Refresh Achieves Best Short-Horizon Performance

Among all schedule-sweep settings, M=50 achieves the best test score (0.755 at step 100), with entropy mean 0.499 and KL mean 0.005. These results confirm that isolation periods—the complete freezing of the teacher between updates—are the key mechanism for stability and performance.

The advantage of M=50 over EMA α=0.04 is not attributable to teacher age: lag-aligned EMA (α=0.04; Appendix E) matches M=50 in average teacher age yet produces markedly worse entropy stability (mean 0.727 vs. 0.499). The difference is structural: each isolation window com-

Table 2: Two design requirements for stable teacher update, validated in Sections 3.2–3.4. Teacher isolation indicates whether the teacher remains fully frozen for a sufficient interval; state-aware refresh indicates whether an update depends on the observed student state.

Schedule	Teacher isolation	State-aware refresh	Observed failure
Tight refresh ($M \leq 2$)	Insufficient	No	Reference collapse
EMA	None	No	Steady-state contamination
Fixed refresh (M large)	Sufficient	No	State-oblivious collapse
CGTR (ours)	Sufficient	Yes	—

pletely freezes the teacher, whereas EMA incorporates a fraction of student updates at every step and *never resets the teacher-student gap*—drift accumulates unboundedly regardless of the per-step mixing rate—a failure mode we term *steady-state contamination*. This establishes a key principle: *isolation period, not teacher age, enables stable self-distillation*. Furthermore, entropy alone is an unreliable consolidation signal: it can be high during productive exploration or suppressed by over-regularization. Reward improvement directly verifies that the student has made genuine progress.

3.4 The Fragility of Fixed Isolation Periods

The schedule-sweep results reveal a second structural observation: while sparse hard refresh creates the best short-horizon stability, a fixed update interval is inherently *oblivious to student state*. The teacher is refreshed at every M-th step regardless of whether the student is currently consolidating or drifting. If a refresh fires during a transient entropy spike or a reward plateau, the teacher absorbs corrupted behavior in a single, irreversible copy—unlike EMA, which limits contamination per step by α, a hard copy applies the full student state instantaneously with no partial recovery path. The longer the isolation period, the larger the potential drift that can accumulate before the next refresh, amplifying this irreversibility risk in proportion to the very property that made isolation beneficial—a failure mode we term *state-oblivious collapse*.

Compounding this, the optimal interval M^* is dataset-dependent: different tasks have different learning dynamics, reward densities, and trajectory length characteristics, so the collapse boundary shifts across datasets. A single pre-specified M is therefore unlikely to be simultaneously safe across all tasks. This motivates an adaptive refresh rule that preserves isolation periods while conditioning each update on whether the student has genuinely consolidated, which we develop in Section 4.

Diagnostic: length-tail expansion as an early warning. Length-tail expansion can appear before abrupt changes in entropy or sequence length become visible in mean values. Figure 5 illustrates how tail expansion in $M=2$ precedes the abrupt changes at step 87 by roughly 12 steps. Monitoring response length tail ratio growth provides a fast, stateless early-warning signal, separate from the accuracy-based collapse criterion.

3.5 Two Design Requirements

The schedule sweep above reveals two design requirements that any viable teacher update rule must satisfy simultaneously.

(A) Sufficient teacher isolation. The teacher must remain frozen long enough to avoid immediately absorbing student fluctuations. This isolation helps preserve temporal independence: it prevents student drift from immediately entering the teacher, preserving a corrective historical reference. Tight hard refresh ($M \leq 2$) provides too little isolation, while EMA provides none.

(B) State-aware refresh. Each teacher update must be triggered by evidence that the student has consolidated, rather than by elapsed training steps alone. A refresh that fires during transient drift copies the corrupted state into the teacher in a single, irreversible step. Clock-based schedules violate this requirement by construction.

Table 2 maps each canonical schedule to these requirements; no clock-based schedule satisfies both, motivating a rule that combines sufficient teacher isolation with state-aware refreshes.

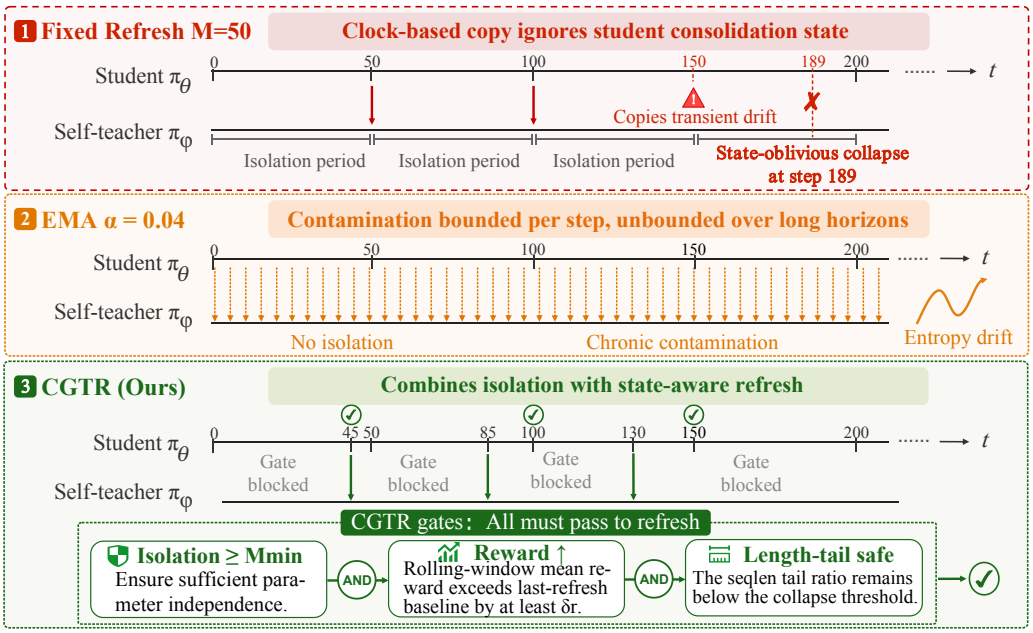


Figure 3: Comparison of teacher-update strategies. Fixed refresh ignores student state, EMA continuously tracks student updates, and CGTR preserves isolation while refreshing only when the isolation, reward, and length-tail gates jointly pass.

4 Consolidation-Gated Teacher Refresh

Consolidation-Gated Teacher Refresh (CGTR) is designed to satisfy both design requirements from Section 3.5 simultaneously: a hard isolation floor provides sufficient teacher isolation, while conditional gating on observable consolidation signals makes each refresh state-aware. Each condition can be evaluated from a single rolling window with no additional model inference or learned component (Figure 3). The full update rule is:

$$\phi_t = \begin{cases} \theta_t & \text{if } \begin{array}{l}\text{IsolationGate} \\ \wedge\ \text{RewardGate} \\ \wedge\ \text{LengthGate}\end{array}, \\ \phi_{t-1} & \text{otherwise.} \end{cases} \tag{4}$$

where t_{last} is the step of the most recent teacher refresh. The three conditions are defined as follows.

Isolation gate. The teacher is held frozen for at least $M_{\min}$ steps after the last refresh:

$$t - t_{\text{last}} \geq M_{\min}. \tag{5}$$

This directly enforces teacher isolation: during the isolation window, the teacher cannot absorb student fluctuations. By preventing immediate tracking, the isolation period helps preserve *temporal independence* (Section 3.5), keeping the distillation target anchored to an earlier parameter state that can counter current student deviations. The condition is a hard prerequisite; the two state-aware gates below are evaluated only once it is met.

Reward gate. The mean training reward over the most recent w steps must exceed the reward recorded at the last teacher refresh by at least δ_r:

$$\bar{r}_{t-w:t} > r_{\text{ref}} + \delta_r, \tag{6}$$

where r_{ref} is the rolling-average reward captured at the last refresh (updated after each successful refresh), and $\delta_r \geq 0$ is the improvement threshold. $\bar{r}_{t-w:t}$ is computed online from training rollouts and requires no additional model inference. This implements a *ratchet*: r_{ref} rises with each refresh, so every subsequent teacher update requires strictly higher performance than the previous one.

Length-tail gate. The maximum-to-mean sequence-length ratio over the most recent w steps must remain below a safety threshold:

$$\frac{\max(\texttt{seqlen}_{t-w:t})}{\text{mean}(\texttt{seqlen}_{t-w:t})} \leq L_{\max}. \tag{7}$$

This makes refreshes sensitive to *behavioral instability*: a rising relative length tail is an early warning of unstable training (Section 3.4) that reward improvement alone cannot detect in time.

Compared to a fixed-M schedule, CGTR introduces three interpretable hyperparameters ($M_{\min}$, δ_r, $L_{\max}$), each with a direct physical interpretation: $M_{\min}$ bounds the isolation period, δ_r sets the performance consolidation threshold, and $L_{\max}$ limits admissible relative length dispersion. These are set once from schedule-sweep diagnostics, in contrast to the opaque integer M of a fixed schedule.

5 RESULTS

We evaluate CGTR and fixed-schedule baselines in a 300-step long-horizon comparison (Chemistry), run per-gate ablations to verify each condition's contribution, and test three methods on Biology, Physics, and ToolUse to assess cross-task generalization. Table 3 summarizes the Chemistry long-horizon results; Table 4 reports the gate ablation; Table 5 reports the cross-task comparison.

Table 3: Long-horizon comparison results (300 steps, no auxiliary recovery). Test fin/peak = test mean@16 at end / best step. Ent mn/fin = entropy mean / final. Coll. follows Appendix A.6 (— = none observed); #Ref = teacher refreshes.

Sched.	Test fin	Test peak	Ent mn	Ent fin	Coll.	#Ref
M=50	0.000	0.744	0.539	0.525	189	6
EMA-04	0.713	0.798	1.539	3.015	—	—
CGTR	**0.806**	**0.809**	0.654	0.780	—	**3**

5.1 EXPERIMENTAL SETUP

We evaluate base Qwen3-8B (Yang et al., 2025) without instruction tuning on the level-L3 Chemistry, Biology, and Physics subsets of SciKnowEval (Feng et al., 2025) and on ToolAlpaca ToolUse (Tang et al., 2023), using exact-match binary rewards. Chemistry uses a 1,890/210 train/test split. Training uses 4×A100 GPUs, FSDP, global batch size 16 (n=4 rollouts per question), AdamW lr 10^{-5}, distillation top-k=20, and JSD β=0.5. The primary metric is test mean@16; we also report bootstrap-estimated B@4. The 100-step sweep compares $M \in \{1, 2, 3, 4, 5, 8, 10, 20, 50\}$, a fixed teacher, and EMA $\alpha \in \{0.1, 0.04\}$; Chemistry long-horizon and cross-task runs use 300-step training on held-out data. Appendix A provides full details.

5.2 FIXED-INTERVAL REFRESH COLLAPSES UNDER LONG-HORIZON TRAINING

Table 3 and Figure 4 report the long-horizon comparison; M=50 collapses catastrophically at step 189: mean@16 drops from its peak of 0.744 (step 187) to 0.000 and does not recover for the remainder of training. The mechanism is consistent with the structural prediction of Section 3.4: the hard refresh at step 150 fires while the student is in an elevated entropy state (≈ 0.98), copying a transiently unstable snapshot into the teacher. By step 189, entropy collapses toward zero and the model degenerates into repetitive outputs. Subsequent clock-driven refreshes copy the degraded student state into the teacher, leaving no recovery path.

EMA α=0.04 avoids catastrophic failure by capping per-step contamination, but the trade-off is chronic drift: entropy grows monotonically from ≈ 0.7 to 3.0 by step 300, and performance plateaus at 0.713 after step 190 with no further gain.

5.3 CGTR ACHIEVES BEST LONG-HORIZON PERFORMANCE WITHOUT PER-DATASET TUNING

CGTR achieves the highest final score (0.806) and peak score (0.809) in the Chemistry long-horizon run (Table 3), with zero collapse over 300 steps. Its entropy remains at a lower level on average than

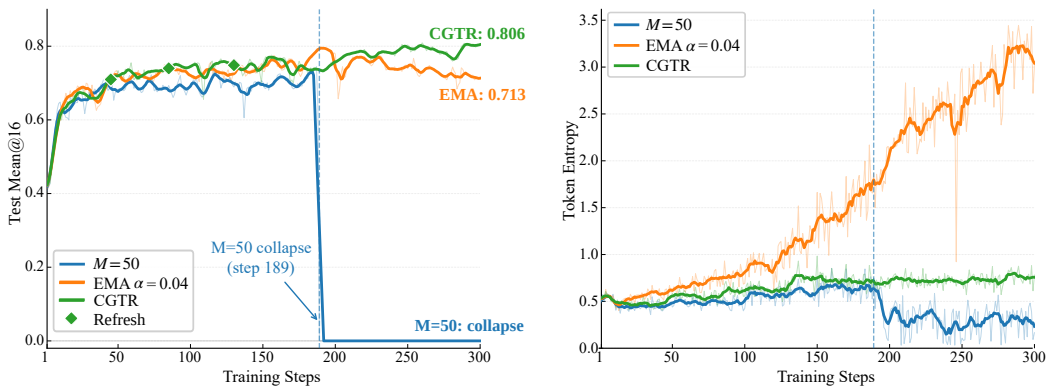


Figure 4: Long-horizon comparison (300 steps, Chemistry, no auxiliary recovery). **(Left)** Test mean@16: M=50 (blue) collapses at step 189; EMA α=0.04 (orange) plateaus at 0.713; CGTR (green) achieves 0.806 with zero collapse and 3 adaptive refreshes. **(Right)** Token entropy: EMA entropy drifts upward throughout training, while CGTR entropy remains stably low; M=50 exhibits a sharp entropy decrease after collapse.

Table 4: Gate ablation on Chemistry (300 steps). Test fin/peak = test mean@16 at end / best step. #Ref = teacher refreshes; Coll. follows Appendix A.6 (— = none observed).

Variant	Test fin	Test peak	#Ref	Coll.
CGTR (full)	0.806	0.809	3	—
w/o Isolation gate	0.731	0.768	12	—
w/o Reward gate	0.000	0.781	6	218
w/o Length-tail gate	0.771	0.792	4	—

EMA's (0.654 vs. 1.539). The reward ratchet and length-tail guard ensure every teacher update captures consolidation rather than transient drift, avoiding failure modes that trap both fixed schedules.

Frozen-teacher controls. To test whether CGTR's gain comes simply from freezing the teacher, we compare it with a teacher fixed at initialization and with the M=50 teacher snapshot at step 100 frozen through step 300. Both controls avoid collapse but finish below CGTR (0.725 and 0.765 vs. 0.806; Appendix H). Thus, the gain is not reproduced by either simple freezing strategy in this run; this comparison does not cover every possible freeze time.

Adaptive refresh behavior. CGTR triggers three teacher refreshes at steps 45, 85, and 130. After the third refresh at step 130 (reward $=$ 0.844), the threshold rises to 0.894. The student does not reach this level in the remaining 170 steps, so the teacher stays frozen from step 130 to 300. During this self-determined isolation window, test accuracy improves from 0.752 to 0.806 (+0.054)—demonstrating that a long static teacher is beneficial once the reward ratchet has ensured it encodes a genuinely consolidated state.

5.4 Ablation: Each Gate Is Necessary

We remove one gate at a time on Chemistry (300 steps) to quantify each condition's independent contribution (Table 4).

Removing the **reward gate** is the most damaging: without the ratchet condition, refreshes fire during transient reward dips, doubling the refresh count (6 vs. 3) and collapsing at step 218 (test fin $=$ 0.000); the length-tail gate guards only against generation explosion—it monitors seqlen tail growth, not reward-level bad refreshes—and therefore cannot intercept an ill-timed copy triggered during a reward dip before length pathology has emerged. Removing the **isolation gate** ($M_{\min}$=0) quadruples the refresh rate (12 vs. 3); without a structural freeze floor the teacher tracks the student too closely, and chronic contamination during reward-plateau windows degrades the final score to 0.731 without triggering hard collapse. Removing the **length-tail gate** allows one premature

Table 5: Cross-task generalization (300 steps). Test fin = final mean@16; Test peak = best step value. Coll. follows Appendix A.6 (— = none observed).

Task	Sched.	Test fin	Test peak	Ent fin	Coll.
Biology	$M{=}50$	0.000	0.560	0.000	130
	EMA-04	0.471	0.569	3.522	—
	CGTR	**0.485**	**0.573**	0.204	—
Physics	$M{=}50$	0.670	0.702	0.450	—
	EMA-04	0.660	0.728	2.271	—
	CGTR	**0.703**	**0.735**	0.278	—
ToolUse	$M{=}50$	0.467	0.612	1.843	—
	EMA-04	0.481	0.641	3.206	—
	CGTR	**0.536**	**0.643**	**0.892**	—

refresh at step 127 (reward above threshold but tail ratio $3.41{>}L_{\max}$), slightly corrupting the teacher and dropping the final score by 0.035; the run avoids collapse, confirming the length-tail gate as a complementary guard against ill-timed updates. Importantly, all ablated variants reach comparable peak scores (0.768–0.792), yet their final outcomes diverge sharply (0.000–0.771). This shows that peak performance alone can obscure schedule fragility, making long-horizon retention essential for evaluation.

Cross-task generalization. We run all three schedules for 300 steps on Biology, Physics, and ToolUse to test whether the failure modes generalize. The cross-task results (Table 5) are consistent across all four settings. **Biology:** $M{=}50$ collapses at step 130 (final $\rightarrow$ 0.000), replicating the Chemistry failure on a different task; CGTR avoids collapse and outperforms EMA (final 0.485 vs. 0.471), triggering twice (steps 127, 192). **Physics:** all three schedules avoid collapse; CGTR still leads on final test (0.703), peak (0.735), and entropy (0.278), with a single gate trigger at step 180 reflecting Physics' characteristically slower reward dynamics. **ToolUse:** no hard collapse; $M{=}50$ nonetheless declines from peak 0.612 to final 0.467 under state-oblivious refreshes, EMA declines similarly (final 0.481, entropy 3.206), while CGTR holds the best final (0.536) and lowest entropy (0.892) with two well-timed refreshes (steps 95, 190).

Multi-seed verification. Across three seeds on each of Chemistry and Biology, $M{=}50$ collapses in all runs (3/3 per task), whereas CGTR has no observed collapse (0/3 per task). CGTR's final mean@16 is 0.805 ± 0.007 on Chemistry and 0.482 ± 0.003 on Biology (mean$\pm$std; Appendix G).

Summary. Across all four tasks, CGTR achieves the best final and peak scores with no observed collapse, using one shared parameter set without per-dataset retuning. All thresholds were fixed from the 100-step Chemistry sweep before the long-horizon and cross-task evaluations; selection and sensitivity are documented in Appendices C and D.

6 Conclusion

In this work, we show that *isolation periods*—not teacher age—are the structural property that enables stable self on-policy distillation, and that fixed-interval schedules carrying this advantage nonetheless fail catastrophically at long horizons through *state-oblivious collapse*: a single ill-timed refresh irreversibly copies a transiently drifting student into the teacher, a failure mode mechanistically distinct from EMA's chronic contamination and invisible under short-horizon evaluation. Our proposed *Consolidation-Gated Teacher Refresh* (CGTR) resolves this tension by preserving isolation periods while gating each refresh on joint evidence of reward improvement and length-tail safety, ensuring every teacher movement is a response to genuine student consolidation rather than a clock signal. CGTR achieves zero collapse and the best final score on all four tasks (Chemistry, Biology, Physics, ToolUse) with a single shared parameter set—demonstrating that the right question for teacher scheduling is not *"how old should the teacher be?"* but *"is the student ready for the teacher to move?"*

AI USE STATEMENT

Generative AI tools were used to improve the readability and presentation of author-written text, including language polishing and LaTeX formatting, and to support literature retrieval and discovery. The authors manually reviewed every AI-assisted edit and verified all identified references against the original publications. These tools were not used to generate research ideas or hypotheses, design or execute the methodology or experiments, implement the proposed method, generate or clean datasets, conduct data analysis, interpret results, formulate or prove mathematical claims, translate text, or draft manuscript sections. The authors take full responsibility for the final content of this work.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024. doi: 10.48550/arXiv.2306.13649. URL https://arxiv.org/abs/2306.13649. arXiv:2306.13649.

Kehua Feng, Xinyi Shen, Weijie Wang, Xiang Zhuang, Yuqi Tang, Qiang Zhang, and Keyan Ding. Sciknoweval: Evaluating multi-level scientific knowledge of large language models, 2025. URL https://arxiv.org/abs/2406.09098.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21271–21284, 2020. URL https://arxiv.org/abs/2006.07733.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

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

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

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327. Association for Computational Linguistics, 2016. doi: 10.18653/v1/D16-1139.

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, 2026. URL https://arxiv.org/abs/2604.13016.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015. doi: 10.1038/nature14236.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022. URL https://arxiv.org/abs/2203.02155.

Boris T. Polyak and Anatoli B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992. doi: 10.1137/0330046.

Andrei A. Rusu, Sergio Gomez Colmenarejo, Çaglar Gülçehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. Policy distillation. *arXiv preprint arXiv:1511.06295*, 2015. URL https://arxiv.org/abs/1511.06295.

Hejian Sang, Yuanda Xu, Zhengze Zhou, Ran He, Zhipeng Wang, and Jiachen Sun. Crisp: Compressed reasoning via iterative self-policy distillation, 2026. URL https://arxiv.org/abs/2603.05433.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

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

Qiaoyu Tang, Ziliang Deng, Hongyu Lin, Xianpei Han, Qiao Liang, Boxi Cao, and Le Sun. Toolalpaca: Generalized tool learning for language models with 3000 simulated cases, 2023. URL https://arxiv.org/abs/2306.05301.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang,

Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled rlvr, 2026. URL https://arxiv.org/abs/2604.03128.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 15476–15488. Curran Associates, Inc., 2022. doi: 10.52202/068431-1126. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/639a9a172c044fbb64175b5fad42e9a5-Paper-Conference.pdf.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models, 2026. URL https://arxiv.org/abs/2601.18734.

A Experimental Setup and Evaluation Protocol

A.1 Data and Prompt Format

Datasets. We evaluate on two benchmark families spanning four tasks:

- **Science Q&A** (Chemistry, Physics, Biology): Undergraduate-level scientific multiple-choice reasoning using the reasoning subsets (level L3) from SciKnowEval (Feng et al., 2025). Each question provides four options; correctness is judged by exact option match against the gold label.
- **Tool use**: Mapping a tool-API specification and user request to the correct tool call, using ToolAlpaca (Tang et al., 2023). The reward is a binary signal indicating whether the predicted tool name and arguments exactly match the reference call.

Dataset splits. SciKnowEval (Feng et al., 2025) is split 9:1 into train and test sets: Chemistry 1,890/210, Biology 450/50, Physics 720/80. ToolAlpaca (Tang et al., 2023) contains 3,938 tool-use instances covering 426 real-world APIs; we use a 9:1 split (3,544 train / 394 test). No question appears in both splits; the test set is held out throughout training.

Prompt format. **Science Q&A.** Each question is presented as a two-turn conversation. The system turn contains a fixed instruction:

> *Given a question and four options, please select the right answer. Respond in the following format:* `<reasoning> ... </reasoning><answer>... </answer>`*. For the answer, only output the letter (A, B, C, or D).*

The user turn contains the question body with options appended. **ToolUse.** The prompt contains the tool-API specification and user request, and the response is the predicted tool call. Across tasks, the maximum prompt length is 2,048 tokens and the maximum generated response length is 8,192 tokens.

Self-teacher input. Following SDPO (Hübotter et al., 2026), the student receives the original prompt x. For each rollout y, the teacher re-evaluates y token by token under x augmented with a successful response from the same rollout group when available. Otherwise, available environment feedback is used; without either source, the sample provides no self-distillation target. A sample's own success is not reused.

A.2 Reward Function

Science Q&A. The reward is a binary correctness signal:

$$r = \mathbf{1}\big[\text{extract}(\hat{y}) = y^*\big],$$

where $\hat{y}$ is the model response and y^* is the ground-truth option letter. Extraction parses the last `<answer>...</answer>` tag in the response; if parsing fails the reward is 0. **ToolUse.** The reward is 1 only when both the predicted tool name and argument string exactly match the reference call; otherwise it is 0. No external verifier or neural reward model is used. Ground-truth answers and reference calls are not inserted into student prompts; teacher-only context follows the feedback construction above.

A.3 Rollout and Decoding Parameters

Training rollouts. Temperature 1.0, top-p=1.0, top-k=-1 (disabled), n=4 responses per question (GRPO group size (Shao et al., 2024)).

Test evaluation. Following SDPO (Hübotter et al., 2026), we sample n=16 responses per question using temperature 0.6, top-p=0.95, top-k=-1 (disabled), and `do_sample=True`. Mean@16 is the mean correctness across the 16 sampled responses, averaged over test questions. B@4 follows the SDPO estimator: for each question, we draw 1,000 size-4 samples with replacement from

its 16 validation responses, take the maximum reward in each sample, average over the bootstrap samples, and then average over test questions. Schedule sweep: one evaluation at step 100. Long-horizon comparison (Chemistry): every step (`test_freq`$= 1$). Cross-task runs (Biology, Physics, ToolUse): every 10 steps.

A.4 Optimizer and Training Details

AdamW (Loshchilov & Hutter, 2019): β_1=0.9, β_2=0.999, weight decay 0.01, gradient clipping ℓ_2-norm 1.0, lr 10^{-5}, constant schedule. The long-horizon comparison uses a 10-step linear warm-up; the schedule sweep uses none. PPO (Schulman et al., 2017) mini-batch size 4 with gradient accumulation.

A.5 Distillation Loss

The distillation objective is a token-level top-k approximation of the generalized Jensen–Shannon Divergence (JSD). For each token position, the top-k=20 logit indices of the *student* are selected; teacher log-probabilities are gathered at the same positions. A tail bucket absorbs the residual probability mass $1 - \sum_{i \in \text{top-}k} p_i$, computed via `logsumexp` for numerical stability, yielding a $(k{+}1)$-class categorical for both student and teacher. Let $p_j{=}\pi_\theta(\cdot \mid x, y_{<j})$ and $q_j{=}\pi_\phi(\cdot \mid x, f, y_{<j})$. The loss is the generalized JSD:

$$\mathcal{L}_{\text{distill},j} = (1{-}\beta)\, D_{\text{KL}}(p_j \| M_j) + \beta\, D_{\text{KL}}(q_j \| M_j), \tag{8}$$

where $M_j{=}(1{-}\beta)p_j{+}\beta q_j$ and β=0.5 (symmetric JSD). Per-token loss is clipped by importance-sampling weight $\min(\rho,\, c)$ with c=2.0, where ρ is the token-level IS ratio between the updated policy and the rollout policy, correcting for within-mini-batch distribution shift.

A.6 Collapse Detection and Early Termination

For performance reporting, *collapse* is the first evaluated step with test mean@16 < 0.01, using the evaluation frequency above. Persistence and recovery are separate observations. Length and entropy are diagnostics; *reference collapse* describes failure of the teacher reference, not an accuracy threshold crossing.

The short-horizon sweep has only a final test evaluation. $M{=}1$ is terminated at step 16 without a test score; $M{=}2$ finishes at 0.225. The step 10 and step 87 entries in Table 1 are reported reference-collapse markers, not measured accuracy threshold crossings. Figure 5 retains the step 87 marker for $M{=}2$: its mean response length rises from approximately 2,555 tokens at step 87 to 6,690 at step 88 and remains near the 8,192-token cap thereafter. Earlier accuracy crossings are undetermined. Long-horizon $M{=}50$ first crosses the accuracy threshold at step 189 (test $\rightarrow 0.000$).

B Detailed Schedule-Sweep Results

B.1 Additional Metrics

Table 6 provides additional metrics for the schedule sweep.

C CGTR Hyperparameter Values and Selection Rationale

Table 7 lists all CGTR gate parameters and their values used in every experiment reported in this paper. *No per-dataset tuning was performed*: the same values are used for Chemistry (long-horizon comparison), Biology, Physics, and ToolUse.

Threshold selection protocol. All gate thresholds were determined from the *100-step schedule sweep* (Section 3.2–3.4), which involves only the short-horizon diagnostics on Chemistry. The long-horizon Chemistry run (300 steps) and all cross-task runs (Biology, Physics, ToolUse) were conducted *after* the parameters were fixed, with no further adjustment. This ordering ensures that the cross-task results constitute a proper held-out evaluation: the gate thresholds had zero exposure to the long-horizon or cross-task data at the time of selection.

Table 6: Additional schedule-sweep metrics (100 steps). Succ = success_group_fraction mean; Ent. fin = last-step entropy. †M1 terminated step 16; ent. mean over steps 1–16 only.

Setting	T m@16	T B@4	Ent. mn	Ent. sd	Ent. fin	Rew. mn	KL mn	Succ
M1†	N/A	N/A	∼1.05	—	0.492	—	—	—
M2	0.2250	0.2760	0.900	1.063	5.226	0.504	0.031	0.746
M3	0.6729	0.7396	1.715	1.167	3.663	0.607	0.085	0.773
M4	0.6562	0.7175	1.536	1.131	4.013	0.595	0.053	0.773
M5	0.6533	0.7228	2.001	1.091	3.548	0.629	0.064	0.793
M8	0.5964	0.6291	1.413	0.713	2.641	0.597	0.004	0.681
M10	0.6693	0.7373	0.545	0.278	0.839	0.614	0.025	0.770
M20	0.6286	0.6894	0.595	0.119	0.697	0.598	0.004	0.706
M50	0.7548	0.8273	0.499	0.048	0.416	0.629	0.005	0.816
Fixed	0.6967	0.7508	0.437	0.097	0.165	0.637	0.008	0.760
EMA α=0.1	0.6449	0.6890	1.417	0.694	2.248	0.613	0.005	0.684
EMA α=0.04	0.7128	0.7815	0.727	0.142	1.864	0.665	0.022	0.792

Table 7: CGTR gate parameters used in all experiments. Rationale for each value is discussed in the text below.

Parameter	Value	Role
$M_{\min}$	30	Minimum isolation period (steps)
w	5	Rolling window size (steps)
Check interval	5	Gate evaluation frequency (steps)
δ_r	0.05	Reward improvement threshold
$L_{\max}$	3.2	Length-tail refresh-guard threshold ($\max/\mu$)

Selection of $M_{\min} = 30$. The minimum isolation period was set to be strictly less than the best fixed-schedule baseline (M=50), so that CGTR retains the freedom to refresh more frequently when conditions warrant, while still exceeding the $M \leq 5$ regime where isolation is too brief to preserve temporal independence (Section 3.2). The value 30 was read directly from the schedule-sweep failure boundary—a structural property of the model and task class, not a tuned quantity—and fixed before any long-horizon run.

Selection of $\delta_r = 0.05$. The reward signal $r_t = \mathbf{1}[\text{correct}]$ has a binary scale $[0, 1]$ with per-window noise $\approx \pm 0.02$ (estimated from the rolling variance over the schedule-sweep runs at step 100). A threshold of 0.05 places the bar at roughly $2.5\times$ the noise level, requiring a gain that is unlikely to be a sampling fluctuation. This choice is grounded in the noise floor of the reward estimator, as estimated from the binary-reward Chemistry sweep with the fixed rollout batch configuration (n=4).

Selection of $L_{\max} = 3.2$. We use $L_{\max} = 3.2$ as an empirical refresh-guard setting, fixed with the other thresholds after the Chemistry diagnostic sweep and before the long-horizon and cross-task experiments. The scale-normalized ratio is also sensitive to isolated maxima: Figure 5 shows early threshold exceedances when only 1–3 of 64 responses reach the token cap while the mean remains low. We therefore do not treat 3.2 as a collapse classifier or claim that all stable steps lie below it. The setting instead reflects the scale of the late M=2 excursion: over steps 75–88, ratio growth is accompanied by sustained growth in mean response length, with the ratio exceeding 4.0 immediately before the reported step-87 marker. Setting $L_{\max} = 3.2$ makes such unusually dispersed tails sufficient to veto a candidate refresh. Because all three gates must pass, an isolated exceedance only delays a refresh; it does not label the run as unstable. The Biology sensitivity analysis (Appendix D) tests $L_{\max} \in \{2.8, 3.2, 4.0\}$, with final scores 0.480, 0.485, and 0.469, respectively, and no observed collapse. These results support the combined rule within the tested range, rather than a standalone stability boundary at 3.2.

D HYPERPARAMETER SENSITIVITY ANALYSIS

To provide the strongest possible evidence that the gate thresholds generalize across tasks, we conduct the sensitivity sweep on **Biology**—a task whose parameters were never involved in any threshold selection. We vary each gate threshold independently (300 steps) while holding the other two at their default values. Table 8 reports test fin and #Ref for each setting.

Table 8: Hyperparameter sensitivity on Biology (300 steps). Bolded rows are the default (selected from Chemistry schedule sweep only). Test fin = final mean@16; #Ref = refreshes; Coll. follows Appendix A.6 (— = none observed).

Param.	Value	Test fin	#Ref	Coll.
$M_{\min}$	20	0.471	3	—
	30	**0.485**	**2**	**—**
	40	0.479	2	—
δ_r	0.02	0.463	4	—
	0.05	**0.485**	**2**	**—**
	0.10	0.474	1	—
$L_{\max}$	2.8	0.480	2	—
	3.2	**0.485**	**2**	**—**
	4.0	0.469	3	—

Across all nine variants test fin ranges from 0.463 to 0.485, a spread of 0.022. Seven of nine variants match or exceed the EMA-04 baseline (0.471); the two exceptions (δ_r=0.02: 0.463; $L_{\max}$=4.0: 0.469) fall within 0.008 of EMA and correspond to configurations that are deliberately suboptimal by design (over-frequent refreshes and a loose length guard, respectively). Crucially, **no variant collapses**, whereas M=50 collapses at step 130 on Biology with the same setup—the safety guarantee is robust across the entire parameter range tested. These results confirm that the default thresholds, derived solely from Chemistry diagnostics, transfer to a held-out task without requiring retuning, and that the performance plateau is broad rather than sharp.

E LAG ALIGNMENT DERIVATION

For a hard refresh with interval M, the teacher parameters are copied from the student at every M-th step and held frozen in between. At any step within a refresh cycle, the teacher parameters were set k steps ago, where k is uniformly distributed over $\{1, 2, \ldots, M\}$. The average age of the teacher parameters is therefore $(M + 1)/2$ steps, and the mean lag relative to the *current* student is $(M - 1)/2$ steps.

For EMA with rate α, the teacher at step t is a weighted average of all past student snapshots with exponentially decaying weights:

$$\phi_t = \sum_{k=0}^{\infty} \alpha(1-\alpha)^k \theta_{t-k}. \tag{9}$$

The mean age of this exponential weighting is $(1 - \alpha)/\alpha$ steps. Setting the two mean ages equal gives the *lag alignment condition*:

$$\frac{1-\alpha}{\alpha} = \frac{M-1}{2} \quad \Longrightarrow \quad \alpha \approx \frac{2}{M+1}. \tag{10}$$

For M=50, this yields $\alpha \approx 2/51 \approx 0.039$; we use $\alpha = 0.04$ as the lag-aligned EMA baseline throughout the paper. If lag alone explained the performance difference between M=50 and EMA, then lag-aligned EMA should match sparse hard refresh across all training horizons. As shown in Section 3.3, it does not—the advantage of hard refresh is structural (isolation periods) rather than attributable to teacher age.

F TERMINOLOGY

We collect here the key terms used throughout the paper.

Temporal independence. A teacher is *temporally independent* when its historical state retains predictive information not already present in the current student's on-policy distribution. As $M \to 1$ or $\alpha \to 1$, temporal independence vanishes and the distillation objective becomes self-confirming.

Isolation period. In periodic hard refresh, the steps between two consecutive refreshes form an *isolation period* during which the teacher is completely frozen. Any student drift that occurs during this period is not absorbed by the teacher. There is no isolation period in EMA. An isolation period prevents immediate tracking of student fluctuations, thereby helping preserve temporal independence. Very short isolation, however, may be insufficient.

Teacher contamination. When the teacher tracks the student closely (low M or high α), it absorbs the student's transient noise, errors, and distribution shifts. The distillation objective then reinforces rather than corrects these deviations. We call this *teacher contamination*. In EMA, contamination is bounded per step by α but accumulates over long horizons; in hard refresh, a single ill-timed update can contaminate the teacher fully and irreversibly.

Refresh shock. At each hard refresh, the teacher parameters jump from ϕ_{t-1} to θ_t. The magnitude of this jump, $\|\theta_t - \phi_{t-1}\|$, characterizes whether the refresh is *corrective* (the student has improved during the isolation period) or *contaminating* (the student has drifted or collapsed). A corrective refresh is followed by a transient KL spike as the student re-adapts; a contaminating refresh shows no such spike.

Length-tail risk. The distribution of generated sequence lengths is heavy-tailed under reinforcement-style training. Its tail can expand before abrupt changes appear in mean entropy or reward. We track `seqlen_max` and `seqlen_mean` as early-warning diagnostics.

G MULTI-SEED VERIFICATION

We repeat the 300-step comparison of $M{=}50$, EMA $\alpha{=}0.04$, and CGTR on Chemistry and Biology over three seeds in total: the original run (seed 0) and two additional runs (seeds 1–2). The training setup follows Appendix A, with the CGTR hyperparameters in Appendix C held fixed without retuning. Evaluation is performed every step on Chemistry and every 10 steps on Biology.

Table 9 reports the mean and standard deviation across seeds for final and peak test mean@16, together with the number of runs that collapse and their collapse-time statistics. The original results in Tables 3 and 5 are single-run measurements; the values here summarize all three seeds.

Table 9: Multi-seed comparison on Chemistry and Biology (300 steps, three seeds per task and schedule). Test fin/peak = final / peak test mean@16, reported as mean±std across seeds. Collapsed = runs meeting Appendix A.6's accuracy criterion / total runs. Collapse step = mean±std of the first detected collapse step among collapsed runs, with the observed range in parentheses; — indicates no observed collapse within 300 steps.

Task	Sched.	Test fin	Test peak	Collapsed	Collapse step (range)
Chemistry	$M{=}50$	0.000 ± 0.000	0.752 ± 0.009	3/3	208 ± 21 (189–230)
	EMA-04	0.711 ± 0.012	0.795 ± 0.009	0/3	—
	CGTR	$\mathbf{0.805 \pm 0.007}$	$\mathbf{0.809 \pm 0.005}$	0/3	—
Biology	$M{=}50$	0.000 ± 0.000	0.554 ± 0.006	3/3	140 ± 26 (120–170)
	EMA-04	0.471 ± 0.008	0.567 ± 0.006	0/3	—
	CGTR	$\mathbf{0.482 \pm 0.003}$	$\mathbf{0.573 \pm 0.005}$	0/3	—

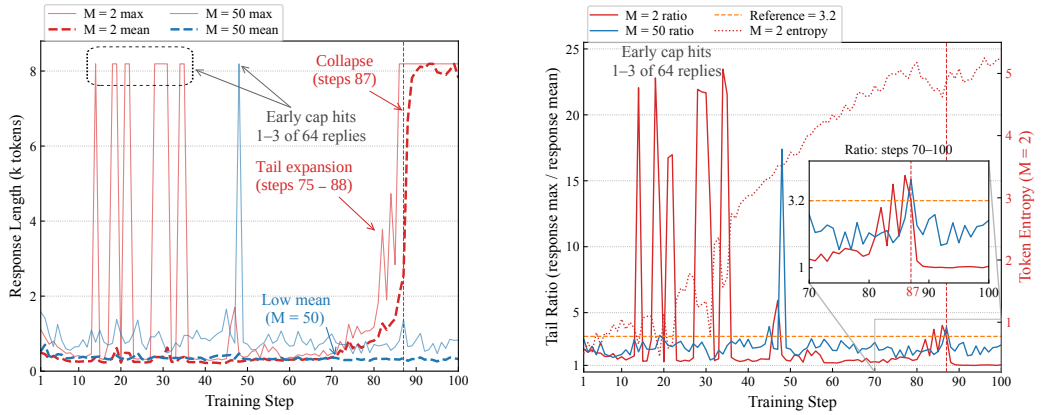


Figure 5: Response-length dynamics on Chemistry over 100 training steps. **(Left)** Per-step maximum (solid) and mean (dashed) response lengths for $M{=}2$ and $M{=}50$. Early cap hits involve only 1–3 of 64 replies, while $M{=}2$ develops sustained late mean-length growth toward the 8,192-token cap. **(Right)** Corresponding maximum-to-mean ratios and $M{=}2$ token entropy (dotted, right axis); the inset magnifies steps 70–100. Vertical dashed lines mark the reported $M{=}2$ reference-collapse step 87, and the orange line denotes $L_{\max}{=}3.2$.

$M{=}50$ collapses in all three seeds on both tasks, although the observed collapse time varies from 189 to 230 on Chemistry and from 120 to 170 on Biology. CGTR and EMA have no observed collapse in these runs, with CGTR achieving the highest mean final score on both tasks. These results support repeatability across the tested seeds; zero observed collapses in three runs is not a general stability guarantee.

H Frozen-Teacher Controls

We add two frozen-teacher controls on Chemistry for 300 steps. **Fixed (init)** holds the pretrained teacher fixed throughout training. **Fixed@100** uses the teacher snapshot active in the $M{=}50$ run at step 100 and freezes it through step 300. The latter is a freeze-after-warmup control without CGTR's state-dependent gating.

Table 10: Frozen-teacher controls on Chemistry (300 steps, single run). Fixed (init) never refreshes the pretrained teacher; Fixed@100 freezes the $M{=}50$ teacher snapshot from step 100 through step 300. Test fin/peak = final / peak test mean@16; Ent fin = final token entropy; Coll. follows Appendix A.6 (— = none observed).

Schedule	Test fin	Test peak	Ent fin	Coll.
Fixed (init)	0.725	0.738	0.158	—
Fixed@100	0.765	0.768	0.420	—
CGTR	**0.806**	**0.809**	0.780	—

Both frozen-teacher controls avoid collapse, but their final scores remain below CGTR: 0.725 for Fixed (init), 0.765 for Fixed@100, and 0.806 for CGTR. The 0.041 difference between Fixed@100 and CGTR shows that the tested freeze-after-warmup control does not reproduce CGTR's final performance. Because this is a single-run comparison at one selected freeze time, it does not establish superiority over every possible frozen snapshot.