PRACTICE, CORRECT, AND COMMIT: SELECTIVE TEACHER CORRECTION FOR REINFORCEMENT LEARNING WITH VERIFIABLE REWARDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Adopting a teacher-corrected checkpoint changes the model that generates subsequent training data in reinforcement learning with verifiable rewards (RLVR), making checkpoint adoption a consequential training decision. We separate producing a corrected candidate from deciding whether to use it for further learning. UQ-Commit implements this separation by retaining the checkpoint after each RL block, constructing a teacher-corrected candidate, and comparing the two before training continues. It adopts correction when the lower bootstrap quantile of the paired validation pass@32 difference exceeds a fixed margin. In an eight-block study with a 4B student and three training seeds, UQ-Commit reaches 53.06 Macro6, the mean pass@32 percentage across six mathematical-reasoning benchmarks. The reported mean gain is 2.25 percentage points over always-accept correction under the same teacher, correction recipe, and GRPO schedule, and 2.18 points over GRPO alone. Separate four-block 4B and 8B comparisons provide additional positive mean gains over GRPO. These results establish selective checkpoint adoption as a useful component of the evaluated teacher-correction pipeline.

1 INTRODUCTION

In teacher-guided reinforcement learning, the model chosen after correction becomes the source of the next training experience. Reinforcement learning with verifiable rewards (RLVR), including Group Relative Policy Optimization (GRPO), updates reasoning models using their own sampled responses (Shao et al., 2024; Yu et al., 2025). Teacher guidance and alternating reinforcement-learning and supervised fine-tuning procedures provide complementary supervision (Agarwal et al., 2024; Byun et al., 2024; Ma et al., 2026). At the boundary between correction and further practice, the choice of checkpoint therefore influences future rollouts, observed errors, and parameter updates.

Our central idea is to separate producing teacher correction from adopting its resulting checkpoint. A staged correction update produces a candidate for continued training, while the checkpoint before correction remains available. Retaining both makes it possible to evaluate the candidate before it generates the next training segment. UQ-Commit makes adoption an explicit decision within the learning procedure, with the current block's completed RL update preserved under either choice.

UQ-Commit implements this decision through practice, correct, and commit (Figure 1). Practice runs a fixed GRPO segment and collects failed responses. Correct fine-tunes a copy of the resulting checkpoint on teacher-generated correction traces. Commit compares the two checkpoints on paired validation prompts and adopts the corrected candidate when the lower bootstrap quantile of the pass@32 difference exceeds a fixed margin. The selected checkpoint initializes the next block. Here, uncertainty refers to sampling variability in the checkpoint comparison.

The primary experiment compares selective adoption with always accepting correction under a common teacher, correction recipe, and GRPO schedule. Across three training seeds in an eight-block 4B study, UQ-Commit reaches 53.06 Macro6, versus 50.80 for always-accept correction: a reported paired mean gain of 2.25 percentage points. Against GRPO alone (50.88 Macro6), the gain is 2.18 points and positive in all three seeds. UQ-Commit has the highest mean on five of six benchmarks. Separate four-block 4B and 8B comparisons provide supporting evidence.

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.

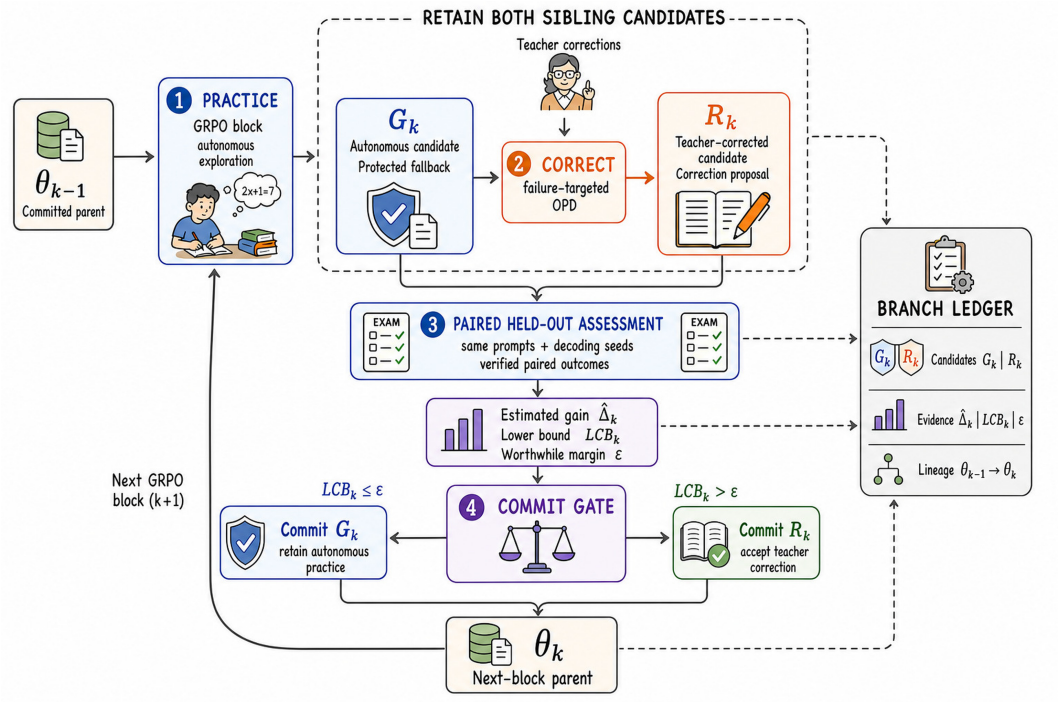

Figure 1: Each block first produces the GRPO checkpoint G_k. Teacher-correction SFT starts from G_k and produces R_k, while G_k remains available. The two candidates are assessed on the same validation prompts. The selected checkpoint initializes the next block. Retaining G_k preserves the current block's GRPO updates.

Our contributions are:

- **Decision formulation.** We formulate checkpoint adoption as a distinct decision in staged teacher correction, made before the selected model generates the next RL training segment.
- **Selection policy.** We instantiate this decision with paired bootstrap assessment of validation improvement, retaining the completed GRPO checkpoint as an available starting point for continued training.
- **Empirical finding.** Under the primary common schedule and correction recipe, the complete selective policy improves final Macro6 over always-accept correction and GRPO; additional schedule and scale comparisons support its value within the evaluated pipeline.

2 Related Work

RLVR and training dynamics. GRPO, DAPO, and DeepSeek-R1 establish the RLVR setting studied here (Shao et al., 2024; Yu et al., 2025; Guo et al., 2025). ProRL investigates extended RL training (Liu et al., 2025a), while pass@k analyses examine the relationship between RL-trained and base-model reasoning coverage (Yue et al., 2025). Studies of entropy dynamics (Cui et al., 2025), optimization bias (Liu et al., 2025c), and training behavior across open base models (Zeng et al., 2025) further characterize how sampling and optimization interact. This literature motivates examining the checkpoint that will generate each new segment of training data.

Teacher guidance and staged correction. Generalized Knowledge Distillation (GKD) uses teacher feedback on student-generated sequences (Agarwal et al., 2024). UFT incorporates solution-prefix guidance (Liu et al., 2025b), LUFFY combines expert trajectories with on-policy experience (Yan et al., 2025), and CHORD dynamically weights supervised expert updates alongside reinforcement learning (Zhang et al., 2026). ARES alternates reinforcement learning with SFT on teacher-corrected multimodal reasoning (Byun et al., 2024); ReLIFT interleaves reinforcement learning with

online fine-tuning on difficult questions (Ma et al., 2026). The latter two are direct antecedents of the practice-and-correct structure. UQ-Commit adds an adoption decision where correction exposes a separate candidate checkpoint.

Distillation objectives. On-Policy Self-Distillation (OPSD) uses a single model in teacher and student roles, with privileged solution information in the teacher context (Zhao et al., 2026). Self-Supervised On-Policy Distillation (SSOPD) derives supervision from successful completions within the student's rollout groups (Tan & Hong, 2026). Distilled RL incorporates teacher signals into the RL objective (Wang et al., 2026), TrOPD restricts distillation to reliable regions and handles outliers (Xing et al., 2026), and On-Policy Delta Distillation transfers a teacher-minus-base delta signal (Heo et al., 2026). These approaches design the supervision within an update. Our implementation uses external teacher traces and SFT; access to reference answers is shared by the two correction-based arms. UQ-Commit selects between the resulting checkpoint and the retained pre-correction checkpoint before further training.

Checkpoint selection and policy evaluation. Uncertainty-Guided Checkpoint Selection (UGCS) ranks observed checkpoints using uncertainty-related training signals (Nguyen et al., 2025). High-confidence off-policy evaluation and policy improvement base candidate-policy decisions on lower bounds for expected return from logged experience (Thomas et al., 2015a;b). UQ-Commit directly evaluates both checkpoints on paired validation prompts before the next training segment is generated. Its bootstrap threshold serves as an empirical adoption criterion; Section 3.3 specifies its interpretation under adaptive validation reuse. The selected checkpoint determines subsequent training data, whereas selecting along an already observed trajectory leaves that trajectory unchanged.

3 METHOD

3.1 TWO CANDIDATES AT A TRAINING BOUNDARY

Let θ_{k-1} denote the checkpoint used to start training block k. A block consists of a fixed number of GRPO steps, which generate rollout data D_k and produce the pre-correction checkpoint

$$G_k = \mathcal{U}_{\mathrm{GRPO}}(\theta_{k-1}; D_k). \tag{1}$$

From the failed rollout responses, we construct a correction batch F_k. The teacher T provides correction traces, and a supervised update starting from G_k produces

$$R_k = \mathcal{U}_{\mathrm{corr}}(G_k; F_k, T). \tag{2}$$

Both candidates share the training history up to G_k. A binary decision c_k specifies whether the additional correction is adopted. The next training checkpoint is R_k when $c_k = 1$ and G_k when $c_k = 0$. This notation distinguishes the decision from the parameter state it selects. Throughout the paper, **commit** means adopting this selected checkpoint for subsequent training; the **training parent** is simply the checkpoint that initializes the next block.

The procedure requires a correction operator that exposes a separate R_k while preserving G_k. It can therefore be applied directly to staged correction updates such as the trace-SFT operator studied here. Expert supervision incorporated inseparably into every RL update does not automatically provide this pair.

3.2 PRACTICE AND TARGETED CORRECTION

Practice uses GRPO with answer-verification rewards. The resulting failed responses identify concrete problems and reasoning attempts on which to request teacher feedback. In the primary protocol, failed rollout rows are uniformly shuffled and at most 125 are selected per block. This is a budget over failed responses; it does not imply 125 distinct problem prompts.

For each selected response, the teacher receives the problem, the failed student solution, concept tags, and the reference answer when available. It samples one correction trace y_i^T. Starting from G_k, we perform one full-parameter SFT epoch. For a nonempty correction batch, the trace-level

objective is

$$\mathcal{L}_{\text{corr}}(\theta) = -\frac{1}{|F_k|} \sum_{i \in F_k} \frac{1}{|y_i^T|} \sum_{t=1}^{|y_i^T|} \log \pi_\theta(y_{i,t}^T \mid x_i, y_{i,<t}^T). \tag{3}$$

Only correction-trace tokens contribute to the loss; prompt tokens are masked. This operator is supervised learning from sampled teacher traces. **Always-accept correction** applies the same operator and continues from R_k at every block.

3.3 Paired validation and adoption

UQ-Commit compares the candidates on a validation pool that is separate from gradient-training data and the six final reporting benchmarks. At block k, it uses the active fold $H_{f(k)}$, where $f(k) = 1 + ((k-1) \bmod F)$. Both candidates use the same prompts, decoding-seed identifiers, sampling settings, and verifier.

For candidate M, let $v_{i,s,j}(M)$ be the binary correctness of sample j for prompt i and decoding seed s. The empirical any-of-32 outcome is

$$a_{i,s}(M) = \mathbf{1}\left\{ \max_{1 \le j \le 32} v_{i,s,j}(M) = 1 \right\}. \tag{4}$$

For each prompt, we average the paired difference over the S decoding seeds, then average over the m prompts in the active fold:

$$d_i = \frac{1}{S} \sum_{s=1}^{S} \big[a_{i,s}(R_k) - a_{i,s}(G_k) \big], \qquad \widehat{\Delta}_k = \frac{1}{m} \sum_{i=1}^{m} d_i. \tag{5}$$

A bootstrap replicate samples m prompt indices with replacement and recomputes the mean of their paired differences. All decoding outcomes belonging to a prompt remain together. From B such replicates, we take the lower α quantile

$$L_k = Q_\alpha\left(\widehat{\Delta}_k^{*(1)}, \ldots, \widehat{\Delta}_k^{*(B)} \right), \qquad c_k = \mathbf{1}\{L_k > \epsilon\}. \tag{6}$$

The next checkpoint is selected by

$$\theta_k = \begin{cases} R_k, & c_k = 1, \\ G_k, & c_k = 0. \end{cases} \tag{7}$$

In the evaluated configuration, $m = 200$, $S = 3$, $B = 10{,}000$, $\alpha = 0.05$, and $\epsilon = 0.005$. The margin is 0.5 percentage points on the pass@32 scale. Equality retains G_k. The lower bootstrap quantile incorporates prompt-sampling variability into the adoption rule; the positive margin specifies the improvement required to switch checkpoints. These are fixed design choices in the reported experiments.

Immediate validation pass@32 is an observable local criterion available before either candidate has generated the next training segment. Its role is to guide the checkpoint choice; final utility is evaluated separately by the reporting benchmarks after the training policy has run. In particular, the criterion does not estimate the unobserved downstream trajectory of a rejected checkpoint.

The five-fold rotation reuses folds 1-3 in blocks 6-8 of the primary study. At those later uses, the candidates can already depend on earlier decisions involving the same fold. We therefore interpret L_k as an empirical selection threshold in this adaptive procedure. The nominal 95% lower-bound interpretation does not supply calibrated coverage for reused folds or a simultaneous guarantee over the training trajectory. All six reporting benchmarks remain outside the adoption rule.

Table 1: Independent experimental cohorts and the role of each comparison. Seeds 42, 123, and 456 are used within each endpoint cohort. Results are compared within cohorts and are not pooled. The 8B comparison uses matched GRPO endpoints.

Cohort	Configuration	Purpose and results
Primary	Qwen3-4B-Base / Qwen3-8B. 8 blocks, 64 GRPO steps per block.	Compare adoption policies under one correction recipe. Table 2; Figure 3.
Four-block schedule	Qwen3-4B-Base / Qwen3-8B. 4 blocks, 64 GRPO steps per block.	Examine a second schedule and failure-allocation policy. Table 4; Appendix B.
Scale study	Qwen3-8B-Base / Qwen3-14B. Separate matched endpoints.	Descriptive comparison at a larger model scale. Table 4; Appendix B.

Algorithm 1. Practice, correct, and commit.

1. Start block k from the previously selected checkpoint θ_{k-1}.
2. Run the GRPO block; save G_k and collect failed responses.
3. Select the correction batch and generate teacher traces; fine-tune a copy of G_k to obtain R_k.
4. Evaluate G_k and R_k on the active validation fold and compute the paired prompt differences.
5. Bootstrap prompts, compute L_k, and adopt R_k if $L_k > \epsilon$; otherwise adopt G_k.
6. Record the candidate identifiers, assessment evidence, and selected checkpoint, then continue training from that checkpoint.

The reported endpoint is θ_K, the checkpoint selected by the policy at the final block. Rejected correction does not undo the current GRPO block. A decision ledger links the selected checkpoint at each boundary to the checkpoint that actually initializes the next block; Appendix A specifies the fields needed for this check.

4 Experiments

4.1 Experimental question and cohorts

The primary experiment compares three complete training policies under the same eight-block GRPO schedule: GRPO alone, always-accept correction, and UQ-Commit. The two correction-based policies use the same teacher, reference-answer access, uniform failure-selection rule, per-block correction cap, and trace-SFT configuration. Their realized training data can diverge after their first different checkpoint choice, because subsequent rollouts are generated by different models. Thus, the endpoint contrast evaluates selective adoption as a training policy under a common recipe.

The additional trajectory-monitoring and paired-continuation experiments are diagnostic cohorts. They are identified separately in Appendix C so that their intermediate scores are not confused with the endpoints in the main comparison.

4.2 Data, evaluation, and implementation

The training mixture draws 6,000 examples from DAPO (Yu et al., 2025) and 2,000 from DeepMath-103K (He et al., 2026), retaining 7,969 after processing. The validation pool contains 1,000 prompts, and the six reporting sets contain 1,546 prompts in total. A normalized exact-match audit reported no prompt overlap between training and validation, validation and reporting, or training and reporting. This audit concerns normalized prompt identity. Data preparation uses split seeds 42, 43, and 44; these are distinct from the three training-run seeds.

The reporting benchmarks are MATH-500, AIME 2024, AIME 2025, AMC 2023, Minerva Math, and OlympiadBench (Hendrycks et al., 2021; Lewkowycz et al., 2022; He et al., 2024). Each prompt

Table 2: Primary schedule-matched 4B results. Entries are empirical pass@32 percentages, mean ± sample standard deviation over three training seeds. Higher is better. Bold marks the highest mean in each column. MATH denotes MATH-500; A24/A25 denote AIME 2024/2025; AMC denotes AMC 2023; Min. denotes Minerva Math; Oly. denotes OlympiadBench. Always-accept applies trace-level correction SFT and adopts the corrected checkpoint at every block.

Method	MATH	A24	A25	AMC	Min.	Oly.	Macro6
GRPO	69.40 ±0.69	27.78 ±1.92	**34.44 ±5.09**	90.83 ±1.44	36.40 ±0.97	46.42 ±0.58	50.88 ±1.18
Always-accept	70.07 ±1.51	23.33 ±3.33	32.22 ±1.92	95.00 ±2.50	37.03 ±0.40	47.16 ±0.70	50.80 ±1.24
UQ-Commit	**71.40 ±1.25**	**32.22 ±1.92**	33.33 ±3.34	**95.83 ±1.44**	**37.63 ±0.91**	**47.91 ±0.69**	**53.06 ±1.03**

is counted as solved when at least one of 32 sampled responses passes the verifier. **Macro6** is the unweighted mean of the six benchmark pass@32 percentages. It measures verified solution coverage at a fixed sampling budget, with equal weight per benchmark. It does not measure the accuracy of a single sampled response. All six benchmark scores accompany the primary aggregate.

The unit of experimental replication is a training seed. We report the mean and sample standard deviation across seeds 42, 123, and 456. Within-cohort differences pair runs by training seed. Decode seeds remain nested evaluation draws, rather than additional training replications. Paired gains are retained from the original unrounded computations and can differ by 0.01 from differences between rounded table entries. With three runs per policy, seed-wise effects are descriptive rather than significance claims.

The 4B practice configuration uses 64 steps per block, response groups of 8, a batch size of 256, an actor learning rate of 10^{-6}, two PPO epochs, a KL coefficient of 0.001, and rollout temperature/top-p of 1.0/1.0. The 8B configuration uses batch size 128, tensor parallelism 2, and GPU-memory utilization 0.50. Correction uses one full-parameter SFT epoch with AdamW, a cosine learning-rate schedule, and learning rate 5×10^{-6}. Further reported settings appear in Appendix A.

The validation pool is permuted with seed 44 into five folds of 200 prompts. Assessment uses decoding seeds 0, 1, and 2, temperature 0.6, top-p 0.95, a maximum of 4,096 generated tokens per sequence, and 32 sequences per prompt and seed. The verifier extracts and normalizes boxed answers and checks exact-match correctness. The bootstrap seed is 20260716.

4.3 Selective correction improves the primary endpoint

UQ-Commit produces the highest Macro6 mean in the primary comparison. The paired gains over GRPO are +0.22, +2.86, and +3.45 points for seeds 42, 123, and 456, with a reported mean of +2.18. Against always-accept correction, the mean gain is +2.25 points. Figure 3 displays these primary contrasts in its two SM rows, alongside the separate schedule and scale cohorts.

The benchmark breakdown shows where the primary gain appears. UQ-Commit has the highest reported mean on MATH-500, AIME 2024, AMC 2023, Minerva Math, and OlympiadBench. Relative to GRPO, the displayed means improve by 4.44 points on AIME 2024 and 5.00 points on AMC 2023. These two competition benchmarks contribute prominently to the aggregate improvement. The evidence supports the Macro6 comparison under this pass@32 protocol and identifies the tasks on which selective correction has the clearest observed benefit.

The comparison with always-accept correction connects this benefit to the adoption policy within the evaluated pipeline. Both correction arms have access to the same form of teacher supervision and use the same SFT procedure. The final mean difference therefore goes beyond the presence of teacher traces alone. It evaluates the combined effects of validation-based selection, its realized correction frequency, and the ensuing training trajectory; the experiment does not separate those components from one another.

Table 3: Correction-adoption counts in the primary cohort. Percentages are calculated from the reported counts. These totals describe the realized policy; they do not reconstruct which individual blocks adopted correction.

Training seed	Adopt R_k	Retain G_k	Adoption rate
42	5 / 8	3 / 8	62.5%
123	5 / 8	3 / 8	62.5%
456	6 / 8	2 / 8	75.0%
Total	16 / 24	8 / 24	66.7%

Table 4: Additional endpoint cohorts. Values are Macro6 percentages, mean ± sample standard deviation over three training seeds. Each cohort is a separate comparison. Low-Mastery uses difficulty-based failure selection and always adopts correction; it changes both failure allocation and adoption relative to UQ-Commit. Full six-benchmark results are retained in Appendix B. A dash denotes a policy not evaluated in that comparison.

Cohort	GRPO	Low-Mastery correction	UQ-Commit	Gain over GRPO
Four-block 4B	49.20 ±1.20	51.11 ±1.95	**51.62 ±1.98**	+2.42
8B scale study	53.39 ±1.10	—	**56.25 ±2.28**	+2.86

4.4 The policy uses both available checkpoints

Both outcomes occur in every primary run. Across 24 block decisions, UQ-Commit adopts 16 corrected candidates and retains 8 GRPO checkpoints. The method therefore implements a selective training path in practice. These counts establish the frequency of each action; the endpoint comparison in Table 2 establishes the performance of the resulting complete policy.

4.5 Additional schedule and scale comparisons

On the four-block 4B schedule, UQ-Commit reaches 51.62 Macro6, with a reported mean gain of 2.42 points over GRPO and a positive difference in each evaluated seed. The paired differences from Low-Mastery correction are +0.49, +0.58, and +0.44 points. This comparison describes the two complete policies, since their failure-selection and checkpoint-adoption rules both differ.

At 8B, UQ-Commit reaches 56.25 Macro6, compared with 53.39 for GRPO. The paired differences are +1.11, +0.74, and +6.75 points. This provides a positive comparison in each observed run at the larger student/teacher scale. The individual effects are reported alongside the aggregate so that the descriptive nature of this three-seed scale study remains explicit.

5 Assessment Cost and Scope

Assessment is an additional operation that evaluates both candidates before a checkpoint is chosen. One complete block assessment generates $2 \times 200 \times 3 \times 32 = 38{,}400$ sequences, or 307,200 sequences over eight blocks. These are autoregressive generations, each capped at 4,096 tokens, rather than individual forward passes. The maximum sequence length is a decoding cap, not a measured average token count. Resampling the saved prompt outcomes does not require further model generations.

Correction generation, SFT, and checkpoint assessment are performed before the adoption decision, including when G_k is retained. The primary comparison matches the GRPO schedule and correction configuration. It therefore measures the model-quality benefit of the additional selection procedure

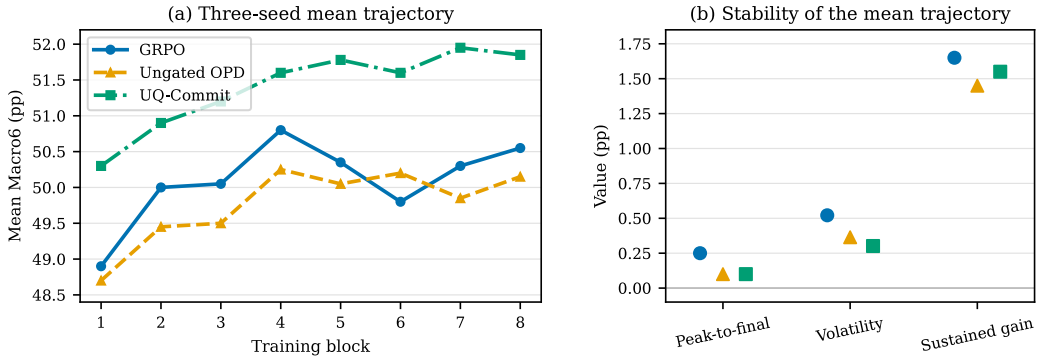

Figure 2: Separate trajectory-monitoring cohort with three seeds and eight blocks, distinct from the endpoint cohorts in the main text. The legend label "Ungated OPD" denotes always-accept trace-SFT correction. Panel (a) shows the mean Macro6 trajectory; panel (b) summarizes that mean trajectory. These are descriptive summaries of the averaged curves, rather than seed-level variability estimates.

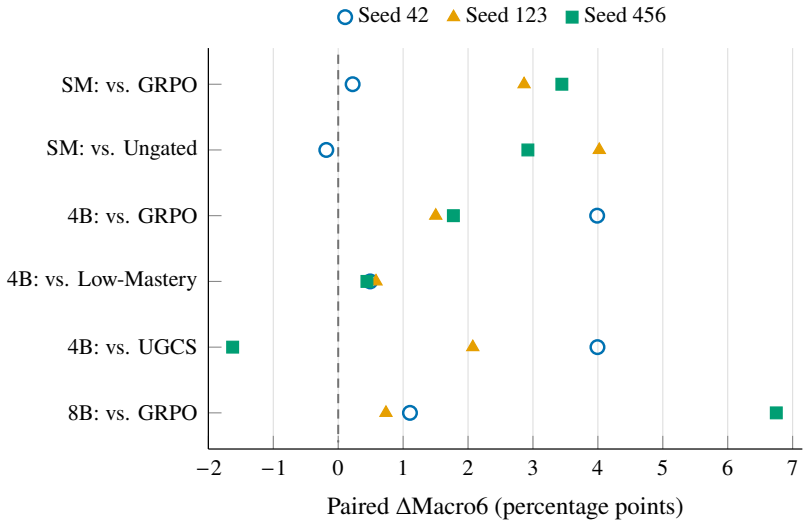

Figure 3: Training-seed paired effects on Macro6. SM denotes the primary schedule-matched eight-block 4B protocol; the other rows belong to the respective four-block 4B and 8B cohorts. Each point is one training seed (42, 123, or 456), and the dashed line marks no gain. Each contrast is computed within its own cohort. UGCS denotes the selection-only comparison, distinct from the training-policy endpoint comparisons in the tables. The label "Ungated" denotes always-accept trace-SFT correction.

under that schedule; it is not a comparison at equal total computational cost. Matching a correction recipe also does not imply equal realized token counts once training trajectories diverge.

Table 5: Protocol-level resource accounting for the primary eight-block 4B study. Counts are derived from the reported configuration. The table reports sequence and update counts, rather than measured token totals, GPU-hours, or wall-clock times.

Operation	Per block	Eight-block run
GRPO practice, all three policies	64 update steps	512 update steps
Teacher correction, correction-based policies	At most 125 generated traces	At most 1,000 generated traces
Correction SFT, correction-based policies	One epoch on at most 125 traces	Eight scheduled SFT epochs, each on its block-specific set
Paired assessment, UQ-Commit	38,400 generated sequences	307,200 generated sequences

6 Conclusion

UQ-Commit separates producing teacher correction from adopting its checkpoint for further reinforcement learning. It retains the completed GRPO checkpoint and uses paired validation to select the starting point for the next training block. Under the primary common schedule and correction recipe, the complete policy reaches 53.06 Macro6, with reported mean gains of 2.25 points over always-accept correction and 2.18 points over GRPO. These results establish the value of selective checkpoint adoption within the evaluated pipeline and identify a concrete decision point for improving teacher-guided training.

AI use statement

Generative AI assistants were used for coding assistance, experiment monitoring, and iterative refinement of figure layouts. The manuscript was written primarily by the authors, with AI assistance in writing limited to language polishing and formatting. The authors designed the research questions and protocols, fixed the scientific decision thresholds, and verified artifact hashes and numerical claims against the original outputs. The authors take full responsibility for the content of the paper.

Reproducibility statement

Section 3 specifies the checkpoint-selection rule and the paired bootstrap procedure. Section 4 describes the data splits, evaluation metric, and training configuration; Appendix A records additional implementation settings and the decision-ledger schema. Appendix B provides the full additional benchmark results, and Appendix C describes the diagnostic cohorts.

REFERENCES

R. Agarwal, N. Vieillard, Y. Zhou, P. Stanczyk, S. Ramos Garea, M. Geist, and O. Bachem. On-Policy Distillation of Language Models: Learning from Self-Generated Mistakes. International Conference on Learning Representations, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/5be69a584901a26c521c2b51e40a4c20-Abstract-Conference.html.

J.-S. Byun, J. Chun, J. Kil, and A. Perrault. ARES: Alternating Reinforcement Learning and Supervised Fine-Tuning for Enhanced Multi-Modal Chain-of-Thought Reasoning Through Diverse AI Feedback. Proceedings of EMNLP, pp. 4410–4430, 2024. URL https://aclanthology.org/2024.emnlp-main.252/.

G. Cui, Y. Zhang, J. Chen, L. Yuan, Z. Wang, Y. Zuo, H. Li, Y. Fan, H. Chen, W. Chen, Z. Liu, H. Peng, L. Bai, W. Ouyang, Y. Cheng, B. Zhou, and N. Ding. The Entropy Mechanism of Reinforcement Learning for Reasoning Language Models. arXiv preprint arXiv:2505.22617, 2025. URL https://arxiv.org/abs/2505.22617.

D. Guo, D. Yang, Z. Zhang, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025. doi: 10.1038/s41586-025-09422-z. URL https://www.nature.com/articles/s41586-025-09422-z.

C. He et al. OlympiadBench: A Challenging Benchmark for Promoting AGI with Olympiad-Level Bilingual Multimodal Scientific Problems. Proceedings of ACL, pp. 3828–3850, 2024. URL https://aclanthology.org/2024.acl-long.211/.

Z. He et al. DeepMath-103K: A Large-Scale, Challenging, Decontaminated, and Verifiable Mathematical Dataset for Advancing Reasoning. International Conference on Learning Representations, 2026. URL https://openreview.net/pdf?id=kHB5Te5IWm.

D. Hendrycks et al. Measuring Mathematical Problem Solving with the MATH Dataset. Advances in Neural Information Processing Systems Datasets and Benchmarks Track, 2021. URL https://arxiv.org/abs/2103.03874.

B. Heo, J. Hwang, S. Yun, and D. Han. On-Policy Delta Distillation. arXiv preprint arXiv:2607.15161, 2026. URL https://arxiv.org/abs/2607.15161.

A. Lewkowycz et al. Solving Quantitative Reasoning Problems with Language Models. Advances in Neural Information Processing Systems, 35, pp. 3843–3857, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/18abbeef8cfe9203fdf9053c9c4fe191-Abstract.html.

M. Liu, S. Diao, X. Lu, J. Hu, X. Dong, Y. Choi, J. Kautz, and Y. Dong. ProRL: Prolonged Reinforcement Learning Expands Reasoning Boundaries in Large Language Models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 17998–18031, 2025a. URL https://doi.org/10.52202/085713-0608.

M. Liu, G. Farina, and A. Ozdaglar. UFT: Unifying Supervised and Reinforcement Fine-Tuning. Advances in Neural Information Processing Systems, 38, 2025b. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/92e886487a8354b03d8bf4416eae6d7d-Abstract-Conference.html.

Z. Liu, C. Chen, W. Li, P. Qi, T. Pang, C. Du, W. S. Lee, and M. Lin. Understanding R1-Zero-Like Training: A Critical Perspective. In *Second Conference on Language Modeling*, 2025c. URL https://arxiv.org/abs/2503.20783.

L. Ma, H. Liang, M. Qiang, L. Tang, X. Ma, Z. H. Wong, J. Niu, C. Shen, R. He, Y. Li, B. Cui, and W. Zhang. Learning What Reinforcement Learning Can't: Interleaved Online Fine-Tuning for Hardest Questions. International Conference on Learning Representations, 2026. URL https://openreview.net/pdf/7765781b6d919f5b9c7a7bbd14b7ea7dc4db41ef.pdf.

M. Nguyen, D. Nguyen, D. Do, S. Venkatesh, and H. Le. Uncertainty-Guided Checkpoint Selection for Reinforcement Finetuning of Large Language Models. arXiv preprint arXiv:2511.09864, 2025. URL https://arxiv.org/abs/2511.09864.

Z. Shao et al. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv preprint arXiv:2402.03300, 2024. URL https://arxiv.org/abs/2402.03300.

Z. Tan and Y. Hong. Self-Supervised On-Policy Distillation for Reasoning Language Models. arXiv preprint arXiv:2605.17497, 2026. URL https://arxiv.org/abs/2605.17497.

P. Thomas, G. Theocharous, and M. Ghavamzadeh. High-Confidence Off-Policy Evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015a. doi: 10.1609/aaai.v29i1.9541. URL https://ojs.aaai.org/index.php/AAAI/article/view/9541.

P. Thomas, G. Theocharous, and M. Ghavamzadeh. High Confidence Policy Improvement. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2380–2388. PMLR, 2015b. URL https://proceedings.mlr.press/v37/thomas15.html.

C. Wang, Z. Li, J. Bai, Y. Zhang, H. Deng, G. Lan, and Y. Wang. Distilled Reinforcement Learning for LLM Post-training. arXiv preprint arXiv:2607.17247, 2026. URL https://arxiv.org/abs/2607.17247.

X. Xing, H. Wang, B. Gao, Z. Li, and Y. Tang. Trust Region On-Policy Distillation. arXiv preprint arXiv:2606.01249, 2026. URL https://arxiv.org/abs/2606.01249.

J. Yan, Y. Li, Z. Hu, Z. Wang, G. Cui, X. Qu, Y. Cheng, and Y. Zhang. Learning to Reason under Off-Policy Guidance. Advances in Neural Information Processing Systems, 38, pp. 117157–117186, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/a9d5c33e38442450f9781bd7bad3c81e-Abstract-Conference.html.

Q. Yu et al. DAPO: An Open-Source LLM Reinforcement Learning System at Scale. Advances in Neural Information Processing Systems, 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/a4277440d50f1f15d2cb4c14f7e0c0d2-Abstract-Conference.html.

Y. Yue, Z. Chen, R. Lu, A. Zhao, Z. Wang, Y. Yue, S. Song, and G. Huang. Does Reinforcement Learning Really Incentivize Reasoning Capacity in LLMs Beyond the Base Model? In *Advances in Neural Information Processing Systems*, volume 38, pp. 57654–57689, 2025. URL https://doi.org/10.52202/085713-1933.

W. Zeng, Y. Huang, Q. Liu, W. Liu, K. He, Z. Ma, and J. He. SimpleRL-Zoo: Investigating and Taming Zero Reinforcement Learning for Open Base Models in the Wild. In *Second Conference on Language Modeling*, 2025. URL https://arxiv.org/abs/2503.18892.

W. Zhang, Y. Xie, Y. Sun, Y. Chen, G. Wang, Y. Li, B. Ding, and J. Zhou. On-Policy RL Meets Off-Policy Experts: Harmonizing Supervised Fine-Tuning and Reinforcement Learning via Dynamic Weighting. International Conference on Learning Representations, 2026. URL https://openreview.net/pdf?id=dCm9bBrk5d.

S. Zhao, Z. Xie, M. Liu, J. Huang, G. Pang, F. Chen, and A. Grover. Self-Distilled Reasoner: On-Policy Self-Distillation for Large Language Models. arXiv preprint arXiv:2601.18734, 2026. URL https://arxiv.org/abs/2601.18734.

A Implementation and Decision Records

Table 6: Additional reported configuration. These settings supplement Section 4.2.

Component	Setting
4B student / teacher	Qwen/Qwen3-4B-Base / Qwen/Qwen3-8B
8B student / teacher	Qwen/Qwen3-8B-Base / Qwen/Qwen3-14B
Correction optimizer	AdamW with the implementation's default betas; cosine schedule
Correction learning rate	5×10^{-6}
Correction training	One full-parameter SFT epoch per block; prompt tokens masked
Precision and memory	bf16; FSDP full sharding; activation checkpointing
SFT per-GPU micro-batch	8 for 4B; 4 for 8B
Validation partition	Seed 44; five folds of 200 prompts
Validation decode seeds	0, 1, 2
Assessment decoding	Temperature 0.6; top-p 0.95; at most 4,096 generated tokens; 32 samples per prompt and seed
Bootstrap	10,000 paired prompt resamples; seed 20260716; fifth percentile
Adoption margin	0.005 in probability units, equivalent to 0.5 percentage points

Decision-ledger schema. For each block, the execution record should contain the incoming checkpoint identifier; identifiers for G_k and R_k; the selected-failure manifest; validation-fold identifier; decode seeds; paired per-prompt outcomes; estimated mean difference; lower bootstrap quantile; margin; decision; and the selected next checkpoint. The incoming identifier of block $k+1$ must match the selected identifier from block k. This schema specifies how to audit the executed policy. The aggregate counts in Table 3 are not a substitute for a complete block-level ledger.

B Complete Additional Endpoint Results

Table 7: Four-block 4B cohort. Empirical pass@32 percentages, mean ± sample standard deviation over three training seeds. Higher is better. Bold marks the highest mean among the displayed policies, including ties at the displayed precision.

Method	MATH	A24	A25	AMC	Min.	Oly.	Macro6
GRPO	70.40 ±1.00	18.89 ±1.92	33.33 ±3.33	87.50 ±0.00	**37.74 ±0.77**	47.32 ±0.79	49.20 ±1.20
Low-Mastery	**71.20 ±0.20**	**24.44 ±5.09**	**36.67 ±3.34**	88.33 ±1.44	37.62 ±2.03	**48.42 ±0.60**	51.11 ±1.95
UQ-Commit	71.00 ±0.40	**24.44 ±5.09**	**36.67 ±3.33**	**93.33 ±5.20**	36.40 ±0.97	47.87 ±1.77	**51.62 ±1.98**

Table 8: 8B endpoint cohort, using the same reporting metric and three training seeds. Higher is better. Bold marks the highest reported mean in each column.

Method	MATH	A24	A25	AMC	Min.	Oly.	Macro6
GRPO	71.53 ±0.90	**32.22 ±1.92**	35.55 ±1.93	91.67 ±1.44	39.70 ±1.33	49.65 ±1.86	53.39 ±1.10
UQ-Commit	**71.87 ±0.50**	31.11 ±1.92	**45.56 ±9.62**	**96.67 ±1.44**	**41.18 ±1.27**	**51.14 ±2.02**	**56.25 ±2.28**

C Diagnostic Training Trajectories

C.1 Separate trajectory-monitoring cohort

Figure 2 displays the separate trajectory-monitoring cohort. UQ-Commit has a reported adjacent-block mean-trajectory change of 0.30 points and a peak-to-final regression of 0.10 points in this cohort. The reported final-minus-initial gains are 1.65 points for GRPO, 1.45 for always-accept correction, and 1.55 for UQ-Commit. These quantities characterize the displayed trajectories; the primary endpoint comparison remains Table 2.

C.2 Paired continuations from a shared boundary

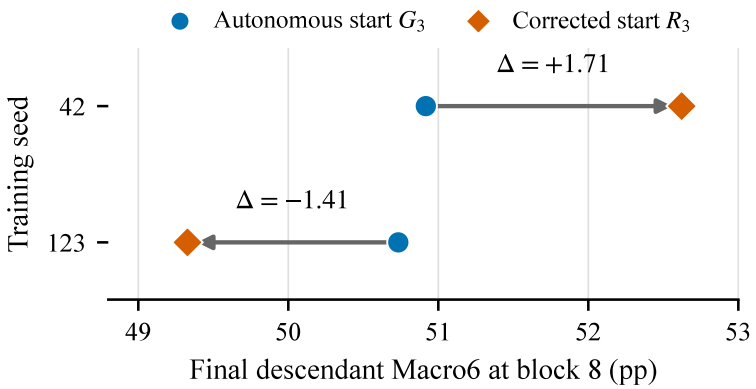

Figure 4: Paired continuations from a shared boundary. The two branches start from G_3 and R_3 and follow a shared downstream schedule to block 8. For training seeds 42 and 123, the corrected-start minus GRPO-start descendant Macro6 differences are +1.71 and -1.41 points, respectively. The experiment diagnoses downstream sensitivity to checkpoint choice.

These continuations show that selecting a checkpoint can change the endpoint reached after further training. Their role is to assess downstream sensitivity to the checkpoint choice, rather than the predictive accuracy of the adoption statistic. Table 2 evaluates the deployed policy.