EMBER: CREDIT-GUIDED DISTILLATION WITH ONLINE STUDENT EXPLORATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Correct teacher traces do not specify which reasoning steps a changing student should imitate. We introduce Evolving Marginal-Benefit Estimation for Reasoning (EMBER), a framework that adapts supervision from fixed, text-only teacher traces to the current student's capabilities. EMBER estimates a step's credit by comparing terminal returns at adjacent teacher prefixes under the same student continuation policy. Positive credit weights imitation, while the sampled continuations also support reinforcement learning. Preserving credit magnitude allows teacher supervision to weaken as its estimated benefit declines. Our analysis characterizes credit reliability under estimation error and policy drift, and shows that question-only mastery alone does not make teacher credit vanish: value drops along teacher prefixes must also disappear. Experiments on instruction following and mathematical reasoning examine this coupling of guidance and exploration. On AMC 2022–2023, Qwen3.5-4B trained with EMBER exceeds SFT by 6.7 percentage points in avg@32, with training dynamics consistent with a diminishing teacher contribution. These findings support student-relative task return as a criterion for adapting teacher supervision during online learning.

1 INTRODUCTION

A correct teacher solution does not tell us which of its steps a particular student should imitate. A step may unlock a successful continuation, repeat something the student already knows, or divert it from a route it can execute more reliably. We therefore view a teacher step as an *action proposal*: its usefulness is the change in task success when the student receives that step and then resumes its own reasoning. This perspective makes teaching depend on both the proposed action and the student's evolving abilities.

Competence-aware selection and return-aware imitation motivate this view. PACED weights problems by student competence, while TIP uses token uncertainty and teacher–student disagreement (Xu et al., 2026a;b). Such signals can guide where to measure, but do not determine the return of a particular intervention. AggreVaTeD connects imitation to cost-to-go (Sun et al., 2017), and FutureBridge-OPD evaluates teacher bridges through subsequent student behavior (Chen et al., 2026a). We study a complementary setting: individual steps from fixed, black-box teacher traces, assessed by terminal task return under *student continuation on both branches*. Only teacher text is required; teacher likelihoods and new responses at student-visited states are unavailable.

Evolving Marginal-Benefit Estimation for Reasoning (EMBER) estimates values at adjacent teacher-prefix boundaries using a refreshed student policy and a fixed terminal-sampling budget. Their difference measures the benefit of inserting the intervening teacher step. Positive estimated credit weights imitation of that step, while the same sampled student suffixes supply terminal-reward GRPO updates. Teacher weights retain their original magnitude, allowing supervision to weaken as its estimated benefit decreases. A learned prefix-value estimator extends sparse rollout measurements across the teacher traces.

Three distinctions make this mechanism precise. First, intervention benefit and improvement from a neural parameter update are different claims; we characterize the former without assuming the latter. Second, estimation error and policy drift can reverse credit signs, so we derive a conservative acceptance rule under explicit error bounds. Third, increasing question-only accuracy does not suffice to eliminate teacher credit. An exact decomposition shows that handoff additionally requires

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.

compatibility with intermediate teacher prefixes. We also derive a cost-aware allocation of shared boundary measurements.

Reported instruction-following and mathematical reasoning comparisons assess the method across student models. A separate Qwen3.5-4B trajectory study reaches 62.3% question-only avg@32 on AMC 2022–2023, 6.7 percentage points above SFT, alongside declining positive teacher credit and teacher-gradient magnitude. A mechanism study on approximately 1% of OpenR1-Math-220k further supports step-aligned credit: the aligned reference outperforms shuffled, uniform-token, and question-level weighting. Together, these studies motivate selective, evolving supervision while identifying where stronger calibration and controlled comparisons remain necessary.

2 Related Work

Distillation and hybrid training. Sequence-level distillation learns from teacher text (Kim & Rush, 2016), while GKD supervises student-generated sequences with teacher distributions (Agarwal et al., 2024). PACED and TIP select supervision using competence, uncertainty, or disagreement (Xu et al., 2026a;b). R^3 and UFT use solution prefixes for exploration, with UFT combining supervised and reinforcement fine-tuning (Xi et al., 2024; Liu et al., 2025b); dGRPO adds distributional teacher guidance to outcome optimization (Ramos et al., 2026). GAD and SODA provide text-based instruction distillation comparators (Ye et al., 2025; Chen et al., 2026b). Against this background, EMBER studies how to assign evolving task-return credit to individual teacher steps.

Intervention and teaching eligibility. FutureBridge-OPD compares continuations after teacher bridges and student actions, using teacher likelihoods to measure subsequent teacher agreement (Chen et al., 2026a). SKALD gates privileged self-distillation by verified advantage (Jiang et al., 2026); VG-OPD uses criterion-level gains to license expert supervision (Xu & Zhang, 2026). These establish selective, outcome-sensitive teaching as prior work. Our estimand instead compares terminal returns before and after one fixed text step, averaging the student's own continuation policy on both branches under a common token budget. This choice supports teacher access through text alone.

Credit estimation and optimization. CRAFT reuses GRPO siblings for counterfactual distillation credit through teacher–student likelihood differences (Meng & Chen, 2026); GammaOPD studies temporal credit in reverse-KL distillation (Liu et al., 2026). VinePPO estimates step values with Monte Carlo rollouts (Kazemnejad et al., 2025). EMBER combines adjacent task-return contrasts with online value estimation and explicit error, drift, and handoff analysis. AggreVaTeD supplies the cost-to-go imitation connection (Sun et al., 2017), while AWR motivates advantage-weighted likelihood (Peng et al., 2019). Our values use learner continuation, and the resulting teacher loss remains distinct from a policy gradient. PPO and GRPO provide the separate suffix optimization surrogate (Schulman et al., 2017; Shao et al., 2024). The theoretical claims concern interventions and estimation under stated assumptions, with transfer to shared neural parameters evaluated empirically.

3 Method: Task-Return Credit from Text-Only Proposals

EMBER alternates between estimating the benefit of teacher steps and updating the student with weighted imitation and shared suffix rollouts (Figure 1). We describe the binary-verifier setting here; the instruction-following extension is given in Section 3.5, with additional implementation details in Appendix B.

3.1 Student-Relative Intervention Credit

Let $\mathcal{D} = \{(x, z, y^T)\}$ contain questions, verified answers, and teacher traces $y^T = (a_1^T, \ldots, a_n^T)$ split at semantic boundaries. Define $p_i = (x, a_1^T, \ldots, a_i^T)$ and $p_0 = x$, including the chat template and remaining token budget. For a frozen student decoding policy π_t and terminal verifier $R_z \in \{0, 1\}$, define

$$V_t(p) = \mathbb{E}_{y \sim \pi_t(\cdot|p)} R_z(p \oplus y), \qquad \Delta_{t,i} = V_t(p_i) - V_t(p_{i-1}). \tag{1}$$

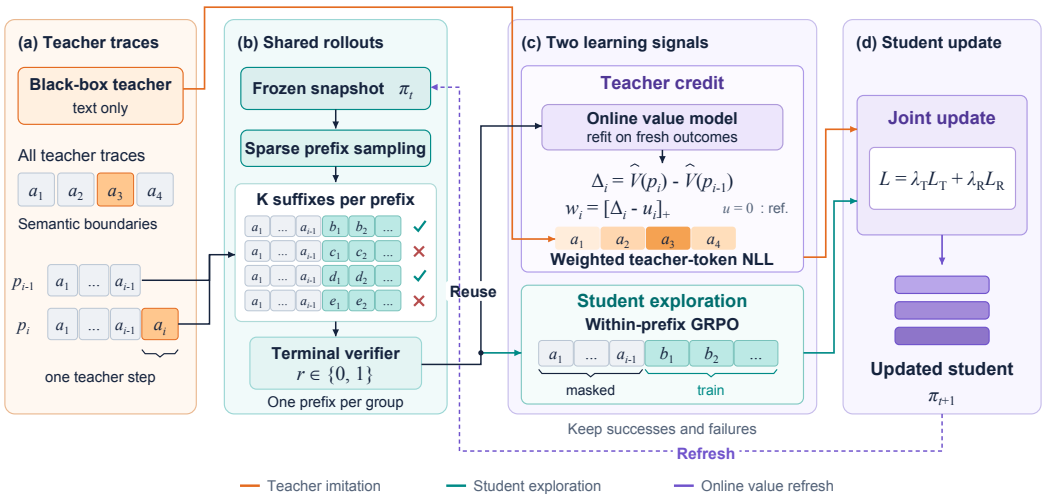


Figure 1: **EMBER overview.** Fixed teacher text defines adjacent prefixes. Shared student suffixes refresh the value estimator and train suffix-only GRPO. Estimated adjacent credit weights teacher-token supervision; joint updates are followed by fresh acquisition. The diagram's Δ_i is an estimated gap and its four suffixes are illustrative. RL masks supplied prefixes. Binary outcomes depict the mathematical setting; instruction following uses frozen rewards (Section 3.5).

Inserting teacher step a_i^T is a deterministic macro-action with zero intermediate reward, so $Q^{\pi_t}(p_{i-1}, a_i^T) = V_t(p_i)$ and $\Delta_{t,i}$ is its student-relative advantage. Both branches use the same future student. Forced tokens consume the common response budget; terminal states use their actual verified reward. Thus the contrast measures the benefit of this step under a fixed verifier, decoding rule, and horizon.

3.2 Online estimation from shared rollouts

A block of B teacher examples receives an expected $G = bB$ prefix groups, with fractional counts carried forward. We use $b = 0.10$ and $K = 16$ continuations per group, each receiving one terminal reward label; Q is trained on all 16 labels. Acquisition mixes frontier prefixes, random adjacent-boundary pairs, and unhinted roots; shared endpoints are sampled once. For each selected prefix p_g, a group $\{y_{g,j}\}_{j=1}^K \sim \pi_t(\cdot \mid p_g)$ supplies the same 16 outcomes for value estimation and suffix reinforcement learning. Constant-reward groups remain in the value-training set even when their RL advantages are zero.

The reference mixture assigns 40% of groups to the early frontier, 40% to randomly selected adjacent-boundary pairs, and 20% to roots. Frontier selection uses the preceding predictor to identify the earliest eligible prefix with predicted success in $[0.05, 0.30]$; random selection provides the cold start and fallback. Answer-revealing prefixes are excluded from frontier RL. Root groups connect acquisition to question-only inference, while random pairs retain coverage beyond the predicted frontier. The mixture specifies a reference design rather than an optimal allocation.

The prefix state-value predictor is $\widehat{V}_{\phi_t}(p) = \sigma(f_{\phi_t}(h_0(p)))$, with frozen causal features h_0 and lightweight online heads fitted by binary cross-entropy on fresh outcomes. Fixed question folds provide cross-fitting: each question is predicted by heads fitted outside its fold. This prevents direct same-question label memorization but does not certify prediction error. Frozen features can be cached in one causal pass over a teacher response; neither future teacher tokens nor the reference answer enters an earlier boundary's features. Heads are warm-started across rounds, but historical outcomes are not relabeled as current-policy measurements. Refreshing Q therefore updates the student-relative target while preserving the feature representation.

3.3 MAGNITUDE-PRESERVING TEACHER SUPERVISION

Let $\widehat{\Delta}_{t,i} = \widehat{V}_{\phi_t}(p_i) - \widehat{V}_{\phi_t}(p_{i-1})$. Teacher tokens receive nonnegative credit

$$w_{t,i} = [\widehat{\Delta}_{t,i} - u_{t,i}]_+, \qquad L_T(\theta) = \frac{1}{N_T} \sum_{(x,y^T)\in\mathcal{B}_t} \sum_i \mathrm{sg}(w_{t,i}) \sum_{v\in\mathcal{T}_i} -\log \pi_\theta(y_v^T \mid x, y_{<v}^T), \quad (2)$$

Here N_T counts nonpadding teacher tokens in the reference reduction; Appendix C gives the experimental denominator. $\mathcal{T}_i$ denotes step i's tokens. The reference uses $u = 0$; the conservative variant subtracts an uncertainty-and-drift allowance (Lemma 2). Its sign certificate requires valid error coverage. All teacher examples receive this pass, without dividing by the sum or mean of weights. Preserving credit magnitude allows supervision to weaken as the estimated benefit decreases, instead of amplifying residual noise near handoff.

Separating selection from supervision strength. Credit determines both which teacher steps receive supervision and how strongly the teacher branch acts. For a fixed batch and student parameters, scaling every weight by $c \geq 0$ gives $L_T(cw) = cL_T(w)$ and scales its gradient by the same factor. Dividing weights by their current mean would cancel this effect whenever that mean is positive: uniformly shrinking credit would leave the normalized teacher loss unchanged. The unnormalized rule instead retains an interpretable response to declining estimated benefit while keeping λ_T fixed. This scaling argument concerns a controlled change in the weights, not successive training checkpoints, where token gradients and the relative placement of credit also change. It therefore motivates measuring both credit magnitude and the teacher gradient. Likewise, mass-matched ablations test where credit is assigned without assuming that equal coefficient mass produces identical updates.

3.4 SUFFIX REINFORCEMENT AND JOINT UPDATES

Within each same-prefix group, define

$$A_{g,j} = \frac{r_{g,j} - \bar{r}_g}{s_g + \epsilon_A}, \quad \bar{r}_g = \frac{1}{K}\sum_j r_{g,j}, \quad s_g^2 = \frac{1}{K}\sum_j (r_{g,j} - \bar{r}_g)^2,$$
$$\rho_{g,j,v} = \frac{\pi_\theta(y_{g,j,v} \mid p_g, y_{g,j,<v})}{\pi_t(y_{g,j,v} \mid p_g, y_{g,j,<v})}. \quad (3)$$

Successful and failed suffixes are retained, and different prefixes never share a reward baseline. The joint objective is

$$L_R = -\frac{1}{N_R} \sum_{g,j,v} \min\{\rho_{g,j,v}A_{g,j}, \mathrm{clip}(\rho_{g,j,v}, 1-\epsilon_c, 1+\epsilon_c)A_{g,j}\},$$
$$L_{\text{student}} = \lambda_T L_T + \lambda_R L_R. \quad (4)$$

Here N_R counts generated nonpadding tokens, including zero-advantage groups; set $L_R = 0$ when $N_R = 0$. This GRPO-style surrogate uses token-level reduction and masks all supplied question and prefix tokens. Natural EOS is included; truncation creates no artificial EOS target. No additional NLL anchor or reference-policy KL is introduced. Each round samples from a frozen policy, refits Q, freezes teacher weights, and applies joint updates before refreshing the rollout policy. The suffixes are on-policy conditional on their starting prefixes; only root groups begin from the question-only deployment distribution.

3.5 LMSYS EXTENSION WITH A FROZEN REWARD MODEL

For the LMSYS instruction-following protocol, we use **Skywork-Reward-V2-Qwen3-1.7B** (Liu et al., 2025a) as a pretrained, frozen terminal reward model. Its parameters are never updated with the student. Let $c_i = (a_1^T, \ldots, a_i^T)$ denote the response portion of p_i. The scorer receives the conversation context and the *complete response*, comprising the supplied teacher prefix and the student-generated suffix:

$$r_{i,j} = R_\psi(x, c_i \oplus y_{i,j}), \qquad V_t^R(p_i) = \mathbb{E}_{y\sim\pi_t(\cdot|p_i)} R_\psi(x, c_i \oplus y), \qquad \Delta_{t,i}^R = V_t^R(p_i) - V_t^R(p_{i-1}). \quad (5)$$

We use the model's scalar reward logits, not correctness probabilities. An unconstrained value head is fitted by squared-error regression on fresh rewards, replacing the sigmoid/Bernoulli value objective. Although the reward model is frozen, the Q model is refreshed because the student continuation policy changes. The same rewards supervise Q and form within-prefix GRPO advantages; RL gradients still apply only to generated suffix tokens. Teacher weights use the same positive-part adjacent credit rule. Group centering is used for GRPO only, never for Q targets or gap estimation.

For this extension, sample nonterminal semantic boundaries uniformly, with a separate root allocation under the same group budget; do not apply binary success-rate frontier gates to reward logits. The terminal value of a complete teacher response is its actual reward, not one. The binary and bounded-return guarantees in Section 4 are not asserted for unbounded raw logits. The instruction-following results are reported in Table 1; their GPT-4o evaluation scores are distinct from these training rewards.

4 When is teacher credit reliable, and when does it vanish?

The following results concern bounded terminal return, fixed decoding and horizon, and explicit policy or estimation assumptions. Proofs and counterexamples are in Appendix A. They specialize standard policy-mixture, concentration, and allocation arguments; they are not a convergence theorem for Equation 4.

Intervention value is not an optimizer guarantee.

Proposition 1 (A one-intervention improvement operator)**.** *At a fixed prefix s, let $A_T = Q^\pi(s, a_T) - V^\pi(s)$. Randomly force a_T with probability $\alpha \in [0, 1]$, otherwise follow π immediately, and use π after the intervention. The return is $V^\pi(s) + \alpha A_T$. Consequently, $\alpha = \eta[A_T]_+$, $0 \le \eta \le 1$, gives gain $\eta[A_T]_+^2$. In an isolated finite macro-action softmax with full support, natural-gradient ascent on $w \log \pi(a_T \mid s)$ induces the same first-order policy direction $w(\delta_{a_T} - \pi)$.*

The first statement is an executable intervention, not a weight update. Ordinary Euclidean log-likelihood ascent can decrease return even with $A_T > 0$: for $\pi = (0.1, 0.8, 0.1)$, $Q = (0.9, 1, 0)$ and teacher action one, $A_T = 0.01$ but the directional return derivative along $\nabla \log \pi(a_T)$ is -0.0606. Shared neural parameters add further coupling. We therefore test transfer from intervention value to actual updates rather than asserting it.

Conservative credit under estimation error and policy drift. Write V_t for the acquisition policy and V_+ for the policy at use time.

Lemma 2 (Error-aware acceptance and credit loss)**.** *On an event where, at both endpoints $j \in \{i-1, i\}$, $|\widehat{V}_j - V_t(p_j)| \le e_j$ and $|V_t(p_j) - V_+(p_j)| \le d_j$, set $u_i = e_{i-1} + e_i + d_{i-1} + d_i$ and $w_i = [\widehat{\Delta}_i - u_i]_+$. Then*

$$0 \le w_i \le [\Delta_{+,i}]_+, \qquad 0 \le [\Delta_{+,i}]_+ - w_i \le 2u_i. \tag{6}$$

In particular $w_i > 0$ implies $\Delta_{+,i} > 0$, and the single-intervention operator with $\alpha = \eta w_i$ has gain at least ηw_i^2.

For m directly measured boundaries with fixed counts K_j and iid Bernoulli continuations, simultaneous bounds are $e_j = \sqrt{\log(2m/\delta)/(2K_j)}$ with probability at least $1 - \delta$. If the maximum next-token total-variation change is ε over every reachable continuation history of length at most H_j, one may take $d_j = \min(1, H_j\varepsilon)$. These bounds are conservative and often vacuous at $K = 16$. Empirical average KL, head disagreement, or cross-fitting alone does not establish their hypotheses. For learned Q at unmeasured states, coverage is an assumption to test; $u = 0$ with prediction noise has no sign certificate.

Endogenous handoff has a prefix-compatibility condition.

Lemma 3 (Teacher-credit envelope and handoff criterion)**.** *For one fixed, complete teacher trace with verified terminal value $V_n = 1$, define $P = \sum_i [\Delta_i]_+$, $D = \sum_i [-\Delta_i]_+$, and $W = \sum_i |\Delta_i|$. Then*

$$P = 1 - V_0 + D = \tfrac{1}{2}(1 - V_0 + W). \tag{7}$$

Along a sequence of student policies, $P \to 0$ if and only if $V_0 \to 1$ and $D \to 0$. Moreover $[\Delta_i]_+ \le 1 - V_{i-1}$. If the per-token teacher score gradients have norm at most G, the exact-credit loss satisfies

$$\|\nabla L_T\| \le \frac{G}{N_T} \sum_{(x,y^T),i} |\mathcal{T}_i|[\Delta_i]_+ \le G \max_{(x,y^T),i} (1 - V_{i-1}). \tag{8}$$

The same gradient upper bound holds on the coverage event for conservative weights. Uniformly accurate plug-in gaps add at most $G\zeta$ if their error is bounded by ζ. Thus prefix-wise competence can make the teacher contribution disappear with fixed λ_T, but increasing root accuracy alone need not: values $(1, 0, 1)$ give $P = D = 1$. Loss rescaling to unit mean would defeat this magnitude-based handoff. Vanishing teacher weights from excessive uncertainty is abstention, not evidence of mastery; remaining GRPO gradients can also be zero.

Measure boundaries according to the contrasts they support.

Lemma 4 (Cost-optimal shared-boundary allocation)**.** *For a fixed trace and policy, let independent boundary estimates have variance σ_j^2/K_j, with $\sigma_j^2 = V_j(1 - V_j)$, and target weighted gap MSE $\mathcal{R} = \sum_{i=1}^n \omega_i \mathbb{E}[(\widehat{\Delta}_i - \Delta_i)^2]$. For fixed $\omega_i \ge 0$, put $\omega_0 = \omega_{n+1} = 0$ and $a_j = (\omega_j + \omega_{j+1})\sigma_j^2$. For $c_j > 0$ and $C > 0$, under $\sum_j c_j K_j = C$, the real-count optimum on indices $a_j > 0$ is*

$$K_j^* = \frac{C\sqrt{a_j/c_j}}{\sum_\ell \sqrt{a_\ell c_\ell}}, \qquad \mathcal{R}^* = \frac{(\sum_j \sqrt{a_j c_j})^2}{C}. \tag{9}$$

This is a cost-aware Neyman allocation on the boundary graph, not a new general allocation theorem. An interior boundary informs two adjacent contrasts; generation cost and Bernoulli variance both matter. It motivates a pilot-estimated allocation within the calibration budget, with integer counts and held-fixed main samples. It does not optimize GRPO's learning value or learned-Q generalization. The fixed-K scheme remains the reference until that proposed allocation is evaluated.

5 Experiments

We examine task performance, the evolution of teacher credit, and the effect of credit assignment. The benchmark comparisons and training studies use different students and settings; conclusions are restricted to their respective comparisons. Appendix C provides the experimental settings and definitions.

5.1 Instruction following and mathematical reasoning

The EMBER runs in Tables 1–2 use the full OpenR1-Math-220k dataset on six A100 80GB PCIe GPUs, with the optimization settings of Table 4 (Appendix C.3).

Table 1 follows the instruction-following evaluation of Chen et al. (2026b). Its GPT-4o Score is $\frac{100}{N}\sum_i S_i/(S_i+R_i)$, where S_i and R_i are judge ratings for student and GPT-4o reference responses. A score of 50 indicates parity; it is neither accuracy nor a win rate. EMBER achieves the highest reported LMSYS score for three of four student families: 51.7 for Qwen2.5-7B-Instruct, 49.3 for Llama-3.2-3B-Instruct, and 52.3 for Llama-3.1-8B-Instruct. Its Qwen2.5-3B-Instruct score of 48.7 trails SODA's 49.2. Transfer benefits also vary: EMBER exceeds SODA on Dolly for three students but falls below it on SelfInst for both 3B students.

For Qwen2.5-Math-7B, Table 2 uses the benchmark layout of Huang et al. (2025). EMBER obtains a reported average of 52.3 versus Prefix-RFT's 51.8, with higher scores on five benchmarks and a slightly lower Olympiad score (55.5 versus 55.7). These tables retain literature baselines and source metric definitions; they provide cross-study context, without establishing matched replications or statistically significant improvements.

5.2 Training dynamics and the handoff

We train Qwen3.5-4B for 69 updates on 2,200 OpenR1 questions (Open R1 Team, 2025), using A800-40G-PCIE hardware. Ten checkpoints measure question-only avg@32 on the 83-question

Table 1: Main results on instruction following: GPT-4o Score (higher is better; 50 indicates parity with the reference). Baseline scores are taken from Chen et al. (2026b), Table 1, which attributes Base, SeqKD, and GAD to Ye et al. (2025); EMBER rows report our results. Bold indicates the highest score for each student and benchmark, including ties. Literature comparisons do not by themselves establish matched training access or compute.

Model	Method	LMSYS	Dolly	SelfInst	Vicuna
GPT-5-Chat	Teacher	51.7	49.8	49.7	49.9
Qwen2.5-3B-Instruct	Base	45.8	45.1	45.6	47.3
	SeqKD	47.5	44.8	45.7	48.0
	GAD	48.9	46.7	47.7	49.4
	SODA	**49.2**	46.1	**48.2**	**49.8**
	EMBER (ours)	48.7	**46.9**	45.9	49.0
Qwen2.5-7B-Instruct	Base	48.7	47.6	48.3	49.1
	SeqKD	49.2	47.2	48.3	49.5
	GAD	50.8	48.5	50.1	**51.4**
	SODA	51.5	49.6	50.8	**51.4**
	EMBER (ours)	**51.7**	**49.7**	**51.3**	**51.4**
Llama-3.2-3B-Instruct	Base	44.0	45.8	47.0	46.9
	SeqKD	47.6	47.0	47.1	48.1
	GAD	48.1	48.5	49.1	48.9
	SODA	49.1	49.5	**50.5**	49.9
	EMBER (ours)	**49.3**	**49.7**	49.7	**50.3**
Llama-3.1-8B-Instruct	Base	46.9	46.6	48.4	47.9
	SeqKD	49.7	47.7	48.7	48.7
	GAD	50.3	48.8	49.5	50.2
	SODA	51.8	**49.9**	51.6	**51.9**
	EMBER (ours)	**52.3**	49.7	**52.3**	51.7

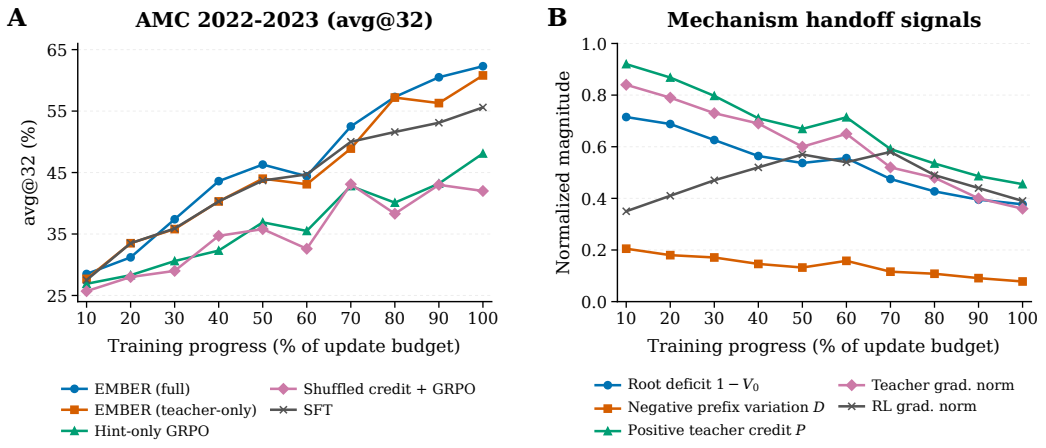


Figure 2: **AMC 2022–2023 training dynamics.** Qwen3.5-4B, 2,200 OpenR1 questions, 69 updates. **A:** Question-only avg@32 (83 questions). **B:** Full EMBER's credit and normalized gradient signals. Markers denote the ten recorded checkpoints; progress is the update budget.

AMC 2022–2023 set. Figure 2 compares full EMBER with teacher-only EMBER, Hint-only GRPO, Shuffled credit + GRPO, and SFT; Table 3 gives the endpoints.

Full EMBER reaches 62.3% accuracy, versus 55.6% for SFT and 60.8% for teacher-only EMBER. The 1.5-point teacher-only gap suggests a modest additional contribution from suffix RL in this run. The larger gaps to Hint-only GRPO (48.1%) and Shuffled credit + GRPO (42.0%) are consistent

Table 2: **Main mathematical reasoning results.** All rows use Qwen2.5-Math-7B. Baseline scores are transcribed from Table 1 of Huang et al. (2025); the EMBER row reports our results. Literature baselines are not our reproductions. The highest displayed value in each metric is bold.

	Math reasoning performance (%)					
Model / method	**AIME** 24 / 25	**AMC**	**MATH** 500	**Minerva**	**Olympiad**	**Avg.**
Qwen2.5-Math-7B	11.5 / 4.9	31.3	43.6	7.4	15.6	19.0
Previous RFT results (reported in the source)						
SimpleRL-Zero	27.0 / 6.8	54.9	76.0	25.0	34.7	37.4
Oat-Zero	**33.4** / 11.9	61.2	78.0	34.6	43.4	43.7
Same-dataset/base comparisons within the source study						
RFT	25.1 / 15.3	62.0	84.4	39.3	46.8	45.5
SFT	22.2 / 22.3	52.8	82.6	40.8	43.7	44.1
RL w/ SFT Loss	19.5 / 16.4	49.7	80.4	34.9	39.4	40.1
SFT+RFT	25.8 / 23.1	62.7	87.2	39.7	50.4	48.2
UFT	20.8 / 16.5	58.8	83.8	33.8	51.6	44.2
ReLIFT	28.2 / 20.1	64.9	87.4	33.8	52.5	47.8
LUFFY	29.4 / 23.1	65.6	87.6	37.5	**57.2**	50.1
Prefix-RFT	31.8 / 26.4	68.2	88.4	40.3	55.7	51.8
EMBER (ours)	32.2 / **27.6**	**68.6**	**88.7**	**41.0**	55.5	**52.3**

All scores, including Avg., are retained as reported. The source study's matched baselines use approximately 46k length-filtered OpenR1-Math-220k examples and DeepSeek-R1 teacher traces; these details apply to that study and do not establish a matched EMBER protocol. This comparison is separate from the Qwen3.5-4B training trajectories.

Table 3: **Final checkpoint of the Qwen3.5-4B trajectory study.** AMC 2022–2023 avg@32 is read at the end of the 69-update budget. Differences are full EMBER minus the listed configuration.

Configuration	AMC avg@32 (%)	Gap to full (pp)
EMBER (full)	**62.3**	0.0
EMBER (teacher-only)	60.8	1.5
Hint-only GRPO	48.1	14.2
Shuffled credit + GRPO	42.0	20.3
SFT	55.6	6.7

with a role for selective teacher supervision. These endpoint contrasts do not imply dominance throughout training: teacher-only leads at 20% progress, and full EMBER temporarily declines at 60%. The first checkpoint is already within training, so its 28.5% accuracy is not an untrained-model baseline.

Mechanism interpretation. From the first to final checkpoint, root deficit $1 - V_0$ falls from 0.715 to 0.377, negative prefix variation D from 0.205 to 0.078, and positive teacher credit P from 0.920 to 0.455. Both terms thus contribute to declining credit. However, the supplied root deficit equals one minus AMC accuracy, and $P = (1 - V_0) + D$ holds at every checkpoint. This is an algebraic consistency check of Lemma 3, rather than independent empirical validation.

The teacher-gradient series decreases from 0.840 to 0.360, while the RL-gradient series rises to 0.580 before ending at 0.390. These trends are consistent with diminishing teacher influence alongside continued RL activity. We interpret these curves as within-branch trends. Positive credit remains substantial at the endpoint, indicating a partial handoff.

Table 4: **Does teacher credit need to match the step?** Mechanism study on 2,200 training questions, approximately 1% of OpenR1-Math-220k. All four arms retain online Q and suffix GRPO under a common, Q-independent prefix-sampling schedule while varying teacher-weight assignment.

Teacher-credit assignment	**Test success** (%) ↑	**Compute** GPU-h
Step-aligned credit (reference)	**55.67**	1.79
Within-question shuffled credit	55.25	1.79
Uniform token weight	55.15	0.33
Question-level weight	55.28	1.79

Separate Qwen 3B study on 2,200 training questions; question-only mean success, not pass@16. Shuffling and question-level averaging preserve each question's token-weighted credit mass; uniform weighting preserves the minibatch total (Appendix C.2). These transformations do not match gradient norms. Reported GPU-hours do not establish a matched efficiency comparison.

5.3 Step-aligned credit on a small training subset

Table 4 examines teacher-credit assignment in a focused Qwen 3B mechanism study using 2,200 training questions, approximately 1% of OpenR1-Math-220k. All four arms retain online Q estimation and suffix GRPO under a common, Q-independent prefix-sampling schedule. This design varies the assignment of teacher weights while holding the acquisition rule fixed. Step-aligned credit reaches 55.67% success, exceeding within-question shuffled credit (55.25%), uniform-token weighting (55.15%), and question-level weighting (55.28%) by 0.42, 0.52, and 0.39 percentage points, respectively.

What the controls test. Each control removes a different aspect of alignment. Within-question shuffling moves credit between teacher steps while preserving each question's token-weighted credit mass. Question-level averaging preserves that mass but assigns the same weight to every supervised token within the question, eliminating distinctions between steps. Uniform-token weighting removes both step-level and question-level variation while preserving the minibatch's total token-weighted credit mass. The transformations apply to each arm's current weights; they do not impose equal gradient norms. Appendix C.2 gives their exact definitions.

Evidence for step-level allocation. The aligned reference achieves a higher reported success rate than each control. This pattern supports the benefit of allocating teacher supervision according to the contribution of individual steps, even within this approximately 1% training subset. Preserving aggregate teacher weight alone does not recover the aligned reference's performance. The study thus provides targeted empirical support for step-aligned credit under the fixed acquisition design. Its central finding is the improvement obtained by retaining the correspondence between a teacher step and its estimated contribution.

6 Limitations and Conclusion

EMBER treats teacher steps as proposals whose value changes with the student. Adjacent prefix-return contrasts connect selective imitation and suffix reinforcement learning through shared rollouts. The analysis identifies conditions for reliable credit and shows that vanishing exact teacher credit requires both question-level mastery and compatibility with teacher prefixes. The observed trajectories support a diminishing teacher contribution; a mechanism study on approximately 1% of OpenR1-Math-220k also shows improved success with step-aligned credit over all three weight-assignment controls.

The approach remains limited by sparse value measurements, policy drift, verifier errors, and transfer from supplied prefixes to question-only inference. Reward-model extensions additionally inherit scorer bias. The reported point estimates describe performance in the evaluated settings; they do not establish statistical significance or a compute efficiency advantage.

AI USE STATEMENT

The research idea, methodological design, and core algorithms were developed by the human authors. AI tools were used only to assist with implementation, and all AI-generated code was reviewed and verified by the authors. GPT-6 Astra was additionally used for language polishing and improving the clarity and readability of the manuscript.

REPRODUCIBILITY STATEMENT

Appendix A gives the proofs and counterexamples; Appendix B specifies implementation details; Appendix C records experimental settings and definitions. The source package includes numerical trajectories, plotting code, result tables, and algebraic checks.

REFERENCES

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

Chishui Chen, Yaoyou Fan, Te Sun, Yi Yang, Chenghao Sun, Delin Mao, Hongbo Qiao, Zuowei Zhang, Junxi Wang, Chenxing Sun, Yangen Hu, Lu Pan, Xuyang Liu, and Linfeng Zhang. Look ahead before you distill: Future trajectory validation of teacher guidance for agentic on-policy distillation. *arXiv preprint arXiv:2608.01953, version 2*, 2026a. URL https://arxiv.org/abs/2608.01953v2.

Xiwen Chen, Jingjing Wang, Wenhui Zhu, Peijie Qiu, Xuanzhao Dong, Yueyue Deng, Hejian Sang, Zhipeng Wang, Alborz Geramifard, and Feng Luo. SODA: Semi on-policy black-box distillation for large language models. *arXiv preprint arXiv:2604.03873, version 5*, 2026b. URL https://arxiv.org/abs/2604.03873v5.

Zeyu Huang, Tianhao Cheng, Zihan Qiu, Zili Wang, Yinghui Xu, Edoardo Maria Ponti, and Ivan Titov. Blending supervised and reinforcement fine-tuning with prefix sampling. *arXiv preprint arXiv:2507.01679, version 3*, 2025. URL https://arxiv.org/abs/2507.01679v3.

Yubo Jiang, Fengying Xie, Zhiguo Jiang, and Haopeng Zhang. Distill skills into weights, not prompts: Abstract skills as privileged signals for on-policy self-distillation. *arXiv preprint arXiv:2608.09826*, 2026. URL https://arxiv.org/abs/2608.09826.

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. VinePPO: Refining credit assignment in RL training of LLMs. *arXiv preprint arXiv:2410.01679, version 2*, 2025. URL https://arxiv.org/abs/2410.01679.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of EMNLP*, 2016. URL https://arxiv.org/abs/1606.07947.

Chris Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiacai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, Yang Liu, and Yahui Zhou. Skywork-Reward-V2: Scaling preference data curation via human-ai synergy. *arXiv preprint arXiv:2507.01352*, 2025a. URL https://arxiv.org/abs/2507.01352. Model: https://huggingface.co/Skywork/Skywork-Reward-V2-Qwen3-1.7B.

Mingyang Liu, Gabriele Farina, and Asuman Ozdaglar. UFT: Unifying supervised and reinforcement fine-tuning. In *Advances in Neural Information Processing Systems*, 2025b. URL https://arxiv.org/abs/2505.16984.

Shiqi Liu, Zeyu He, Letian Tao, Guojian Zhan, Jiaxin Gao, Feihong Zhang, Jingliang Duan, Wei Xiong, Kehua Sheng, Bo Zhang, Yang Guan, and Shengbo Eben Li. Beyond token-local imitation: Reward-compatible temporal credit assignment for on-policy distillation. *arXiv preprint arXiv:2609.16937, version 2*, 2026. URL https://arxiv.org/abs/2609.16937v2.

Zibin Meng and Kani Chen. CRAFT: Counterfactual credit assignment from free sibling rollouts for self-distilled agentic reinforcement learning. *arXiv preprint arXiv:2606.29476*, 2026. URL https://arxiv.org/abs/2606.29476.

Open R1 Team. OpenR1-Math-220k. Dataset on Hugging Face, 2025. URL https://huggingface.co/datasets/open-r1/OpenR1-Math-220k. Dataset card accessed September 22, 2026.

Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019. URL https://arxiv.org/abs/1910.00177.

Miguel Moura Ramos, Duarte M. Alves, and André F. T. Martins. A recipe for long-context reasoning in large language models via on-policy optimization and distillation. *arXiv preprint arXiv:2605.12227, version 3*, 2026. URL https://arxiv.org/abs/2605.12227v3.

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. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Wen Sun, Arun Venkatraman, Geoffrey J. Gordon, Byron Boots, and J. Andrew Bagnell. Deeply AggreVaTeD: Differentiable imitation learning for sequential prediction. *arXiv preprint arXiv:1703.01030*, 2017. URL https://arxiv.org/abs/1703.01030.

Zhiheng Xi, Wenxiang Chen, Boyang Hong, Senjie Jin, Rui Zheng, Wei He, Yiwen Ding, Shichun Liu, Xin Guo, Junzhe Wang, Honglin Guo, Wei Shen, Xiaoran Fan, Yuhao Zhou, Shihan Dou, Xiao Wang, Xinbo Zhang, Peng Sun, Tao Gui, Qi Zhang, and Xuanjing Huang. Training large language models for reasoning through reverse curriculum reinforcement learning. *arXiv preprint arXiv:2402.05808*, 2024. URL https://arxiv.org/abs/2402.05808.

Xun Xu and Zaixi Zhang. Who teaches which token? verifier-gated multi-expert on-policy distillation for scientific reasoning. *arXiv preprint arXiv:2609.15404, version 2*, 2026. URL https://arxiv.org/abs/2609.15404v2.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, and Zhipeng Wang. PACED: Distillation and on-policy self-distillation at the frontier of student competence. *arXiv preprint arXiv:2603.11178, version 3*, 2026a. URL https://arxiv.org/abs/2603.11178v3.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, Zhipeng Wang, and Alborz Geramifard. TIP: Token importance in on-policy distillation. *arXiv preprint arXiv:2604.14084, version 4*, 2026b. URL https://arxiv.org/abs/2604.14084v4.

Tianzhu Ye, Li Dong, Zewen Chi, Xun Wu, Shaohan Huang, and Furu Wei. Black-box on-policy distillation of large language models. *arXiv preprint arXiv:2511.10643*, 2025. URL https://arxiv.org/abs/2511.10643.

A PROOFS, SCOPE, AND COUNTEREXAMPLES

A.1 OPERATIONAL IDENTIFICATION AND INSUFFICIENT PROXIES

Fix a prefix, a decoding policy, a response-token horizon, and a deterministic terminal verifier. Define $Y(0)$ as the return when the student continues immediately, and $Y(1)$ as the return when the specified teacher text is forced first and the same student completes the response. Then $\mathbb{E}Y(1) - \mathbb{E}Y(0) = Q^\pi(s, a_T) - V^\pi(s)$ by definition. Intervening on the prefix makes both arms directly sampleable; one need not infer an unobserved treatment effect from a confounded observational corpus. Independent random seeds estimate the same mean difference as valid coupled seeds. Coupling changes variance, not the estimand. This identifies the effect of exact tokens under the stated protocol. It does not separate mathematical content from extra hints, answer exposure, formatting, or the tokens consumed by the intervention.

Competence and local disagreement cannot identify the sign. Consider two actions, student probabilities $(1/2, 1/2)$, and the same teacher proposal a_1 in two tasks. The action returns are $(1, 0)$ in task one and $(0, 1)$ in task two. Both have $V^\pi = 1/2$, identical local student entropy, and identical local teacher disagreement if the teacher distribution is held fixed. Yet the teacher advantages are $+1/2$ and $-1/2$. No function only of these shared proxy inputs can identify both signs. This is an insufficiency counterexample, not a claim that these proxies have no predictive value in real data. Future student behavior, rather than the task's immediate reward function, can equivalently create the two different return vectors.

A.2 PROOF OF PROPOSITION 1

Condition on the intervention coin. Expected return is $(1 - \alpha)V^\pi(s) + \alpha Q^\pi(s, a_T)$, giving the first claim. Since $A_T \in [-1, 1]$, $\alpha = \eta[A_T]_+$ is a legal probability for $\eta \in [0, 1]$, and $\alpha A_T = \eta[A_T]_+^2$. This mixes a single forced text option with immediate student continuation; it does not insert teacher actions at every subsequent state.

For the natural-gradient statement, let p be a strictly positive categorical macro-action distribution with independent logits z. Its Fisher matrix is $F = \mathrm{diag}(p) - pp^\top$, and $\nabla_z \log p_T = e_T - p$. A solution to $Fv = w(e_T - p)$ is $v = we_T/p_T$; the minimum-norm solution differs only by a constant-logit shift. Thus the induced probability tangent is $\dot{p} = Fv = w(e_T - p)$, and for fixed continuation returns q, $\dot{J} = q^\top \dot{p} = w(q_T - p^\top q) = wA_T$. This is a first-order identity, not an exact finite natural-gradient step or an assertion about an autoregressive network's Fisher approximation. Semantic steps must define a legitimate macro-action partition for this categorical statement; the direct text-intervention statement does not require that partition.

A Euclidean likelihood counterexample. Let $p = (0.1, 0.8, 0.1)$, $q = (0.9, 1, 0)$, and $T = 1$. Then $J = p^\top q = 0.89$ and $A_T = 0.01 > 0$. For a Euclidean logit update $z \leftarrow z + \eta \nabla_z \log p_T$,

$$\left.\frac{dJ}{d\eta}\right|_{\eta=0} = \sum_j p_j(q_j - J)(\mathbf{1}\{j = T\} - p_j) = p_T(q_T - J) - \sum_j p_j^2(q_j - J) = -0.0606.$$

Multiplication by a frozen positive advantage changes the magnitude, not the negative sign. Binary terminal rewards realize each q_j as its Bernoulli mean. This counterexample already holds without parameter sharing across states. In a language model, changing the continuation policy and other prefixes creates additional gaps between intervention and optimization.

A.3 PROOF OF LEMMA 2

For each endpoint, $|\widehat{V}_j - V_+(p_j)| \le e_j + d_j$ by the triangle inequality. Consequently $|\widehat{\Delta}_i - \Delta_{+,i}| \le u_i$. It follows that $\widehat{\Delta}_i - u_i \le \Delta_{+,i}$ and $\widehat{\Delta}_i - u_i \ge \Delta_{+,i} - 2u_i$. Monotonicity and one-Lipschitzness of positive-part thresholding yield

$$[\Delta_{+,i} - 2u_i]_+ \le w_i \le [\Delta_{+,i}]_+, \qquad [\Delta_{+,i}]_+ - w_i \le 2u_i.$$

If $w_i > 0$, then $\Delta_{+,i} \ge w_i > 0$. Proposition 1, applied to the current continuation policy, gives gain $\eta w_i \Delta_{+,i} \ge \eta w_i^2$. The conclusion concerns the gated intervention operator. It does not certify the NLL or joint neural update.

Direct Monte Carlo specialization. For a fixed set of m queried boundaries, independent Bernoulli trials within a boundary give $\Pr(|\widehat{V}_j - V_j| > e_j) \le 2\exp(-2K_j e_j^2)$. Taking $e_j = \sqrt{\log(2m/\delta)/(2K_j)}$ and a union bound gives simultaneous coverage at least $1-\delta$. Boundary independence is not needed for this union bound. Known terminal values require no samples and have $e_j = 0$. One may cap e_j at one using boundedness. If measured boundaries and counts are chosen from an independent pilot, condition on that pilot and use fresh main samples. Repeated peeking or outcome-adaptive stopping requires time-uniform bounds or an explicit error-budget schedule; the fixed-count formula cannot be reused unchanged. Across policy rounds, allocate failure probabilities whose sum is bounded by the desired total failure probability.

For two endpoints with K samples each and no drift, $u = \sqrt{2\log(2m/\delta)/K}$. On the coverage event a true gap $A > 2u$ is guaranteed to receive positive weight; the conservative sufficient condition is $K > 8\log(2m/\delta)/A^2$. This explains why $K = 16$ is not a universal certificate for small effects. No low-cost success claim follows from writing a confidence gate.

Drift specialization. Assume the supremum next-token total-variation distance between acquisition and use policies is at most ε at all reachable histories, with at most H_j remaining tokens. Couple the two policies maximally until their first different token. The probability of any mismatch is at most $1-(1-\varepsilon)^{H_j} \le H_j\varepsilon$. On paths without a mismatch, the deterministic verifier sees the same response. Since return is bounded by one, its expectation changes by at most $d_j = \min(1, H_j\varepsilon)$. Average observed KL or clipping fraction is not this supremum bound. If a uniform conditional KL bound κ is separately justified, Pinsker's inequality supplies $\varepsilon \le \sqrt{\kappa/2}$. In practice these quantities are drift diagnostics unless their required coverage has been established.

No automatic learned-Q certificate. Cross-fitting removes direct same-question label reuse but does not bound approximation bias. An ensemble can agree and still be wrong; marginal calibration is not a uniform bound on neighboring differences. At unqueried prefixes, an honest certificate therefore needs additional function-class or coverage assumptions. If these cannot be supported, report plug-in credit as such, and report abstention and empirical sign error on independent probes. An all-abstain gate has no false admissions but is not a useful improvement algorithm.

A.4 Proof of Lemma 3

For real a, $a = [a]_+ - [-a]_+$ and $|a| = [a]_+ + [-a]_+$. Summing over a fixed trace gives $P - D = \sum_i \Delta_i = V_n - V_0 = 1 - V_0$ and $P + D = W$. Rearrangement gives Equation 7. Because $1 - V_0$ and D are nonnegative, $P \to 0$ is equivalent to both of them tending to zero. For a terminal value other than one the identity becomes $P = V_n - V_0 + D$, and the stated equivalence need not hold. The verified-complete-trace assumption is therefore essential.

Since $V_i \le 1$, $[\Delta_i]_+ = [V_i - V_{i-1}]_+ \le 1 - V_{i-1}$. The triangle inequality and the per-token score-gradient bound give

$$\|\nabla L_T\| \le \frac{G}{N_T} \sum_{(x,y^T),i} |\mathcal{T}_i| w_i.$$

Use $w_i = [\Delta_i]_+$, or $w_i \le [\Delta_i]_+$ on the conservative coverage event, to obtain Equation 8. If $|\widehat{\Delta}_i - \Delta_i| \le \zeta$, then $[\widehat{\Delta}_i]_+ \le [\Delta_i]_+ + \zeta$, adding at most $G\zeta$. The coefficient λ_T must remain bounded for the same conclusion about the teacher contribution to the joint gradient. These bounds concern weights and a sufficient gradient bound; gradient cancellation can make a gradient vanish even when total credit is positive.

Root mastery is insufficient. Values $(V_0, V_1, V_2) = (1, 0, 1)$ are realizable when the student always solves from the question, but an off-support first teacher step sends it to a prefix from which it always fails; the full teacher response is correct. Here $P = D = 1$ despite $V_0 = 1$. Full-support policies can approach this example with $(1-\epsilon, \epsilon, 1)$ as $\epsilon \downarrow 0$. Thus the obstruction persists asymptotically: root mastery does not imply compatibility with teacher prefixes. A rising overall test score likewise does not make every teacher step's advantage decrease monotonically.

What matching the teacher does imply. For a well-defined next-step macro-action distribution, let $p_T = \pi(a_T \mid s)$ and $q_{\neg T}$ be the mean continuation value conditional on another action. For

$p_T < 1$, $A_T = (1 - p_T)(q_T - q_{\neg T})$, so $|A_T| \leq 1 - p_T$. The $p_T = 1$ case has $A_T = 0$. Therefore local action matching is another sufficient condition for disappearing exact credit, but not necessary. If $A_T \leq 0$, exact positive-part credit is already zero. Consistent gap estimates with vanishing allowances inherit these limits. Persistent noisy estimates can instead leave a positive-part noise floor. None of this proves convergence to a better-than-teacher student or nonzero RL gradients after handoff.

A.5 Proof of Lemma 4

Let $\widehat{V}_j$ be independent, unbiased boundary means, and reuse each mean in every adjacent gap. Then

$$\mathbb{E}[(\widehat{\Delta}_i - \Delta_i)^2] = \frac{\sigma_i^2}{K_i} + \frac{\sigma_{i-1}^2}{K_{i-1}}, \qquad \mathcal{R} = \sum_{j=0}^{n} \frac{a_j}{K_j}.$$

Although boundary errors are independent, neighboring gap errors are not: $\mathrm{Cov}(\widehat{\Delta}_i - \Delta_i, \widehat{\Delta}_{i+1} - \Delta_{i+1}) = -\sigma_i^2/K_i$. This covariance does not enter the chosen sum of individual MSEs, but would matter for other objectives. Cauchy–Schwarz gives

$$\left(\sum_j \sqrt{a_j c_j}\right)^2 \leq \left(\sum_j a_j/K_j\right)\left(\sum_j c_j K_j\right) = C\mathcal{R}.$$

Equality requires $K_j \propto \sqrt{a_j/c_j}$, giving Equation 9. Zero-a_j indices need no samples in the idealized objective; if all are zero, the risk is zero and the displayed allocation need not be evaluated. Known verified terminal boundaries also need no samples. Unknown variances require a pilot and minimum counts; integer rounding and minimum constraints change the exact optimum.

This is the classical variance-per-cost allocation argument specialized to adjacent contrasts. The useful design consequence is the incidence factor $\omega_j + \omega_{j+1}$ for a shared boundary, not a claim of a new optimization principle. The reference implementation remains fixed-K. An adaptive calibration arm must charge its pilot, freeze counts and priorities before fresh main outcomes, and keep separate the exploration budget whose purpose is RL rather than estimating these contrasts.

Telescoping does not conserve optimization credit. Neither P nor the NLL-weighted sum is the telescoping signed sum. Multiplying by likelihood gradients or adding GRPO does not decompose the unhinted-return gradient. The handoff identity is a property of values along a trace, not an allocation of a conserved teacher/student gradient.

B Implementation Details

B.1 Acquisition and Value Estimation

The acquisition mixture and eligibility rules are specified in Section 3. The fixed group budget counts both endpoints of a measured transition, with shared endpoints generated once.

Each group contains 16 continuations, with one terminal reward label per continuation. Q uses all 16 labels, and the same group supplies the suffix GRPO update. Decoding settings are given in Appendix C. Each group stores terminal rewards, per-token behavior log-probabilities, masks, termination reasons, and policy IDs. There is no sample-until-success rule. All-correct and all-incorrect groups supervise Q even when their GRPO advantages vanish. The implementation retains the same permanent question folds across rounds.

B.2 Loss Masks and Update Bookkeeping

Natural EOS tokens are supervised; truncation does not introduce an EOS target. Question and supplied prefix tokens condition the student but are masked from the RL loss. No additional NLL anchor or reference-policy KL is introduced. A fixed negative teacher-NLL coefficient would make the scalar objective unbounded below as the teacher probability approaches zero, motivating the positive-part rule. This avoids that pathology without guaranteeing safe neural optimization.

Each round permits at most M joint updates in one pass over the fresh data. Snapshot divergence triggers refresh, and checkpoints include the policy, Q, budget, and data cursor. The two losses are complementary; they are not an exact partition of conserved policy-gradient credit. The mathematical setting uses a verifier, but its learned state-value predictor means the full system is not critic-free. These implementation definitions complement the experimental settings in Appendix C.

C Experimental details

C.1 Qwen3.5-4B training-trajectory study

The trajectory study uses `Qwen/Qwen3.5-4B-Instruct` on A800-40G-PCIE hardware, with one policy-training seed and a fixed budget of 69 updates. We first specify this study's settings, then describe their reuse in the credit-assignment and full-dataset experiments below.

Training data. We use 2,200 questions from OpenR1-Math-220k (Open R1 Team, 2025). The dataset identifier is `open-r1/OpenR1-Math-220k`, with revision

`e4e141ec9dea9f8326f4d347be56105859b2bd68`.

The set combines 508 existing pilot questions and 1,692 hash-selected extension questions. It is not a fresh uniform sample from the full dataset. The data-selection seed is 11403.

Randomness and value supervision. The policy-training seed is 11419. Q-supervision question selection uses seed 11503, cross-fitting uses 11505, Q-label rollouts use 11507, evaluation rollouts use 11423, and credit shuffling uses 11517. Q-head initialization uses base seed 11511, with fold/member-specific derivations; these heads do not constitute additional policy-training seeds. Each queried prefix receives a group of $K = 16$ continuations. Each continuation receives one terminal reward label, and Q uses all 16 labels. The same 16 continuations and rewards supply the within-prefix GRPO objective. This training group size is separate from the 32 answers per question used for AMC evaluation below.

Decoding and AMC evaluation. Generation uses vLLM 0.8.5 in FP32, temperature 0.8, top-p 0.95, at most 1,536 generated tokens, a maximum context length of 8,192, and prefix caching. Evaluation uses 83 questions from AMC 2022–2023 with question-only prompts and 32 sampled answers per question. Writing $R_{q,j} \in \{0, 1\}$ for answer correctness, the plotted metric is

$$\text{avg@32} = \frac{100}{83 \times 32} \sum_{q=1}^{83} \sum_{j=1}^{32} R_{q,j}. \tag{10}$$

Thus each checkpoint averages 2,656 answer outcomes to estimate single-answer success under the stated sampling policy. It is not pass@32, which would count a question as solved if any answer is correct. The ten checkpoints are indexed by 10%, 20%, …, 100% of the update budget; the figure retains these relative-progress coordinates.

Answer verification. We use Math-Verify 0.9.0 with `first_match` answer extraction and priority given to boxed answers. Parsing failures and returned verification-exception states are scored as incorrect, yielding reward zero in the mean-accuracy calculation.

Policy optimization. Training uses full-parameter SGD with momentum zero, weight decay zero, and fixed learning rate 5×10^{-5}, without warmup or a scheduler. Parameters are FP32 and forward computation uses BF16 autocast. A batch contains 32 questions, accumulated through one-question microbatches. One epoch over 2,200 questions comprises 68 complete batches and a final batch of 24, giving 69 optimizer updates. After accumulation, the global gradient norm is clipped to 1.0 before the optimizer step. SFT selects among three predefined learning rates using the internal 256-question development set only; the selected rate is 5×10^{-5}. The reported test curves use the 83-question AMC 2022–2023 set.

Teacher-loss normalization. Let L_q be the number of supervised teacher-response tokens in question q and $\bar{L} = N^{-1} \sum_{q=1}^{N} L_q$ the fixed mean over the training corpus, where $N = 2,200$ for this study. For a batch $\mathcal{B}$, the implemented teacher objective is

$$L_T^{\text{impl}} = -\frac{1}{|\mathcal{B}|\bar{L}} \sum_{q \in \mathcal{B}} \sum_{v=1}^{L_q} \text{sg}(w_{q,v}) \log \pi_\theta(y_{q,v}^T \mid x_q, y_{q,<v}^T). \tag{11}$$

The denominator uses the actual number of questions in the current batch, including the final partial batch. Weighted token NLLs are summed; the loss is neither a mean computed separately for each response nor divided by the total weight. Gradients are accumulated across microbatches before clipping. This defines the teacher objective's reduction, separately from the plotted branch-gradient magnitudes.

Equation 2 uses the reference token denominator $N_T = \sum_{q \in \mathcal{B}} L_q$. For a fixed batch, the implemented loss and its pre-clipping gradient equal that reference multiplied by $r_{\mathcal{B}} = N_T/(|\mathcal{B}|\bar{L})$. Consequently, Equation 8's upper bound and its plug-in error term are multiplied by $r_{\mathcal{B}}$. Since $r_{\mathcal{B}} \leq \max_q L_q/\bar{L}$ for the fixed corpus, this factor is bounded and the gradient-vanishing implication is preserved. The credit identity in Equation 7 does not depend on this loss denominator.

Figure 2 contains five performance trajectories and five mechanism trajectories. Table 3 reports their final performance values and direct percentage-point contrasts. All configurations are compared at the end of their stated update budget; this convention does not establish matched token counts, wall-clock time, or GPU-hours. The first recorded checkpoint occurs at 10% progress, so changes from that point do not measure gains over the initial model. The metric averages individual answer outcomes, not pass@32.

Mechanism interpretation. The reported root-deficit series is the complement of the full method's AMC accuracy. Positive credit equals the sum of root deficit and negative prefix variation at all ten checkpoints. These numerical identities are consistent with Equation 7, but do not establish that the three quantities were measured independently. Finite-sample rectification can bias estimates of positive and negative variation. The aggregate data cannot distinguish this effect from credit-estimation error. Gradient curves describe changes within each learning branch. Their interpretation follows the separate teacher and RL objectives rather than a decomposition of a conserved gradient.

C.2 Separate teacher-credit assignment comparison

Table 4 studies teacher-credit assignment with Qwen 3B on 2,200 OpenR1 training questions, approximately 1% of the nominal 220,000-question dataset. This focused experiment tests whether aligning teacher weights with individual steps improves learning at a limited data scale. All four arms retain online Q and suffix GRPO under a common Q-independent prefix schedule while varying teacher weights. The aligned reference attains higher success than shuffled, uniform-token, and question-level weighting.

This comparison uses the same data construction, seeds, decoding, verification, SGD settings, precision, and teacher-loss reduction specified above. All four arms train for one epoch, with 32-question batches and one-question microbatches, giving 69 updates on the 2,200-question subset. Each training group contains 16 continuations and Q uses all 16 reward labels; only the teacher-weight assignment differs between arms.

For question q, let $\ell_{q,i} > 0$ count supervised teacher tokens in step i, let $w_{q,i} = [\widehat{\Delta}_{q,i}]_+$, and define

$$L_q = \sum_i \ell_{q,i}, \qquad M_q = \sum_i \ell_{q,i} w_{q,i}.$$

The reference uses the step-aligned weights $w_{q,i}$. The control transformations are

$$w_{q,i}^{\text{shuffle}} = \frac{M_q}{\sum_j \ell_{q,j} w_{q,\sigma_q(j)}} w_{q,\sigma_q(i)}, \tag{12}$$

$$w_{q,i}^{\text{uniform}} = \frac{\sum_{q' \in \mathcal{B}} M_{q'}}{\sum_{q' \in \mathcal{B}} L_{q'}}, \tag{13}$$

$$w_{q,i}^{\text{question}} = \frac{M_q}{L_q}. \tag{14}$$

Here $\mathcal{B}$ denotes the logical teacher minibatch. The shuffled control uses a within-question derangement for questions with at least two steps; single-step questions remain unchanged. If $M_q = 0$, all shuffled weights are zero. Positive step lengths ensure that the shuffled denominator is positive whenever $M_q > 0$. Equal weights can remain numerically unchanged after permutation.

Shuffling and question-level averaging preserve each question's token-weighted credit mass M_q; uniform weighting preserves the minibatch total. Rescaling the shuffle need not preserve the original coefficient histogram, and none of the controls normalizes mean credit to one. Matching weight mass does not match gradient norms. Because policies diverge during training, each arm's transformation applies to its own current untransformed weights and on-policy continuations.

The common Q-independent prefix schedule keeps adaptive frontier selection outside this comparison, so the experiment specifically examines teacher-weight allocation within the fixed acquisition design. Its three controls distinguish step-specific assignment from preserving question-level or minibatch-level supervision mass.

C.3 Main benchmark training and task-specific rewards

The EMBER rows in Tables 1 and 2 use the full `open-r1/OpenR1-Math-220k` dataset at the revision specified above, trained on six NVIDIA A100 80GB PCIe GPUs. Their optimization and reproducibility settings match the Table 4 study: one policy-training seed (11419), the same auxiliary seeds, full-parameter SGD at a fixed learning rate of 5×10^{-5}, zero momentum and weight decay, no warmup or scheduler, FP32 parameters with BF16 autocast, and global gradient clipping at 1.0. The logical batch contains 32 questions across the GPUs, with one-question microbatches. Training covers one epoch of the full dataset; its update count follows the number of training questions and is not the 69-update subset budget.

The decoding configuration remains vLLM 0.8.5, FP32, temperature 0.8, top-p 0.95, 1,536 generated tokens, an 8,192-token context limit, and prefix caching. Each selected prefix supplies 16 continuations, each with one reward label; Q uses all 16 labels, which are shared with suffix GRPO. Teacher-loss normalization follows Equation 11, with $\bar{L}$ recomputed over the full training corpus. The full dataset replaces the subset's 508-plus-1,692 question selection.

The benchmark tables retain their listed student families and evaluation metrics. Mathematical rewards use the Math-Verify configuration above. Published baselines are attributed to Chen et al. (2026b) and Huang et al. (2025), respectively, and retain their source studies' training conditions; they are not presented as new replications.

For instruction following, frozen Skywork-Reward-V2-Qwen3-1.7B (Liu et al., 2025a) supplies the training reward described in Section 3.5, while GPT-4o scores evaluate response quality. Reward-model supervision differs from a binary mathematical verifier and adds both preference information and inference cost. The reward and evaluation settings therefore form part of the access and compute conditions needed to interpret cross-study comparisons.

C.4 Complete cost accounting

For N teacher examples, an acquisition pass with $G = bN$ groups has bNK continuations. At the method's reference budget $b = 0.10$, $K = 16$, this is $1.6N$, or 3,520 continuations when $N = 2,200$. This is a per-acquisition-pass arithmetic count under the reference configuration, not a measured total over the 69-update trajectory study. The full training cost is

$$C_{\text{total}} = C_T + C_{\text{generate}} + C_{\text{oldlogp}} + C_R + C_V + C_{\text{operational}}, \tag{15}$$

where the terms include teacher updates, generation, behavior-policy log-probabilities, suffix RL, value-model features and fitting, and operational overhead. Reusing suffix groups avoids duplicate acquisition for Q supervision; it does not remove generation or backpropagation.

The available cost entries in Table 4 are 1.79 GPU-hours for the reference, shuffled, and question-level arms, and 0.33 GPU-hours for the uniform arm. Without a common accounting boundary and hardware record, these values do not identify a fair speedup or an end-to-end cost ratio. The new trajectory study has a hardware identifier but no accompanying resource ledger. Efficiency conclusions therefore require costs that include startup, synchronization, rejected and constant-reward groups, Q computation, and allocated idle time, with evaluation and tuning expense reported separately.

D THEORY-TO-EXPERIMENT DIAGNOSTICS

Figure 2 connects observed learning dynamics to the teacher-credit envelope: root deficit, negative prefix variation, and positive credit decline overall, and the teacher-gradient diagnostic decreases. This pattern supports a descriptive partial-handoff account. The algebraic identity among the plotted credit terms is a consistency check, while residual positive credit shows that the vanishing-credit limit is not reached.

The other theoretical results have narrower empirical coverage. The aggregate curves do not measure the false-positive rate of the uncertainty-and-drift gate, verify pointwise coverage of learned Q, or compare shared-boundary allocation against equal-cost alternatives. Independent current-policy probes would be needed for these quantities, with uncertainty in the signed intervention effects accounted for and admission rate reported alongside false-positive rate. The fixed-schedule credit-assignment comparison concerns training outcomes, not direct verification of the sign certificate. These distinctions separate the proved conditional results from the behavior measured in the present experiments.