SCRA: BALANCED GROUP RECONSTRUCTION FOR MATHEMATICAL REASONING FROM SPARSE SUCCESSES

Anonymous authors
Paper under double-blind review

ABSTRACT

Group-relative reinforcement learning depends on finding informative combinations of successful and unsuccessful responses. When successes are scarce, a verified sibling can seed further sampling. We introduce Success-Conditioned Reconstruction and Augmentation (SCRA): generate additional solutions from a short successful-sibling prefix, verify them, and reconstruct the maximum-available 1:1 group for training on the original no-hint question, then finish with ordinary GRPO. On Qwen3-8B, SCRA reaches 90.04 / 39.17 / 31.25 on MATH-500, AIME 2024, and AIME 2025 over three training seeds, and 71.48 on chemistry. It is strongest on both AIME suites and on every SciKnowEval split; HiPO remains slightly ahead on MATH-500. SCRA stays above GRPO on every reported split, including a GRPO run with 16 original samples that matches the extra sampling budget of triggered SCRA steps on the mathematical protocol. A same-protocol SDPO run collapsed on the mathematical suites after entropy dropped, but stayed above the base model on the four SciKnowEval splits. On the same full-benchmark protocol, 1:1 reconstruction is strongest on both AIME suites among the tested allocations, extra no-hint sampling does not match the sibling prefix on AIME, and withdrawing the scaffold after 81 of 135 steps outperforms a later switch.

1 INTRODUCTION

Reasoning policies can improve by learning from their own verified solutions. A successful response or useful feedback conditions the same model, producing training trajectories for a policy that must later solve the original problem alone (Zhao et al., 2026; Hübotter et al., 2026; Yang et al., 2026). This opportunity is particularly useful when ordinary sampling finds only a few correct solutions: a short prefix from one successful sibling can support additional attempts without an external teacher.

Generating more correct responses is only part of the problem. Group-relative advantages depend on which responses are placed together in the update. A pool dominated by failures provides a different learning signal from a balanced group, even when both contain the same verified successes. We therefore study how to generate additional positive candidates and how to allocate them against the original negative trajectories.

We introduce Success-Conditioned Reconstruction and Augmentation (SCRA); Figure 1 summarizes one training step and the later GRPO tail. During an initial training phase, the model samples ordinary responses, extracts a short prefix from a correct sibling on success-scarce problems, and generates further complete solutions. It combines verified original and conditional positives with original negatives through maximum-available 1:1 reconstruction. Selected conditional responses are then optimized under the original question context with a clipped behavior-probability weight. A later phase uses ordinary GRPO updates, so conditional augmentation is concentrated early in training. Eight instruction templates provide alternative ways to present the sibling prefix; the method selects one template per problem and training step.

The 1:1 allocation gives each reconstructed group equal positive and negative counts. For binary rewards and unnormalized group-centered advantages, it maximizes mean squared outcome contrast. This property motivates the construction; the resulting policy quality is evaluated empirically. Our development experiments first compare 1:1 with 1:2, 1:3, and 2:1 allocations, alongside no-hint addi-

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.

tional sampling and ordinary GRPO. Those frozen choices are then evaluated on the full mathematical protocol.

Our benchmark study compares GRPO, a GRPO run with $N{=}16$ that matches the extra sampling budget of triggered SCRA steps, HiPO, and SDPO on MATH-500, AIME 2024, AIME 2025, and separately trained chemistry, physics, biology, and materials splits. SCRA is strongest on both AIME suites and on chemistry, physics, biology, and materials, and remains above GRPO on every reported split; HiPO is slightly ahead on MATH-500. SDPO collapses on the mathematical protocol. On the SciKnowEval splits it stays above the base model and below both GRPO and SCRA. Table 2 reports reconstruction-ratio and schedule comparisons on the full mathematical protocol, and Table 3 records the prefix-length and failed-hint runs that selected the 192-token sibling prefix.

Our contributions are:

- A training procedure that couples successful-sibling sampling with balanced group reconstruction and a later GRPO phase.
- An analysis of binary-reward contrast under balanced reconstruction, together with empirical positive–negative ratio comparisons.
- No-hint results on MATH-500, AIME 2024, AIME 2025, chemistry, physics, biology, and materials against GRPO, HiPO, and SDPO, including a GRPO $N{=}16$ extra-sampling control, plus prefix-length, failed-hint, reconstruction-ratio, and scaffold-duration comparisons.

2 Related Work

Group-relative learning and group construction. GRPO estimates an outcome baseline from responses to the same problem (Shao et al., 2024). With binary rewards, uniform-outcome groups yield zero relative advantages. SEELE relates rollout accuracy to a local loss-descent upper envelope proportional to $p(1-p)$ and adjusts hint length to target approximately 50% success (Li et al., 2025). Reinforce-Ada's original formulation adaptively collects responses, downsamples toward balanced positive and negative counts, and computes the advantage baseline from the full collected pool (Xiong et al., 2025). We build on these motivations: SCRA selects a balanced group after generation and verification, then recomputes its baseline within the reconstructed group. Thus our intervention acts on the composition of the update rather than directly controlling the policy's rollout success probability.

Self-generated guidance and transfer. HiPO reuses successful siblings as prefix hints for groups with a success rate between zero and one half, retaining informative original groups and adding hinted groups (Deng et al., 2026). SCRA instead combines accepted hinted positives with original trajectories within a reconstructed group and trains them against the original no-hint prompt. HiLL distinguishes creating mixed-outcome groups from transferring hinted success to the no-hint setting through its hint-reliance analysis (Xia et al., 2026). This distinction also motivates evaluating SCRA without hints and treating the subsequent GRPO phase as part of the training schedule.

On-policy self-distillation. Generalized Knowledge Distillation learns from student-generated trajectories under teacher supervision (Agarwal et al., 2023). Recent reasoning methods use privileged contexts to obtain token-level targets or corrected solutions (Zhao et al., 2026; Hübotter et al., 2026; Yang et al., 2026). SCRA operates at the trajectory level: a verifier selects the policy's own conditional completions, which supply positive examples for a group-relative update. Its central design choices are success-conditioned generation, group reconstruction, and withdrawal of the scaffold during the final training phase.

3 Success Scarcity and Group Reconstruction

3.1 Sparse successes in group-relative learning

For a problem x, the old policy samples $N = 8$ no-hint responses,

$$y_i^0 \sim \pi_{\theta_{\text{old}}}(\cdot \mid x), \qquad R(y_i^0) \in \{0, 1\}. \tag{1}$$

With centered outcome advantages, every response in an all-correct or all-incorrect group has zero advantage. A group with only one to three correct responses supplies a learning signal, but successful trajectories remain scarce. Such a group also contains its own source of guidance: a verifier-correct sibling whose reasoning prefix can seed further attempts at the same problem.

3.2 Turning a successful sibling into training examples

SCRA extracts a short prefix h from a correct sibling and samples conditional responses using a reflection template T:

$$y_j^h \sim \pi_{\theta_{\text{old}}}(\cdot \mid x, h, T). \tag{2}$$

Only verifier-correct conditional responses enter the positive pool. Reconstruction uses original successes as preferred positives, adds accepted conditional successes as needed, and pairs them with original incorrect responses. The selected trajectories are trained against the original no-hint prompt, with tokenwise behavior correction for the conditional samples.

This separates discovering additional successes from choosing the examples that define an update. The construction controls both the relative number of positive and negative responses and the source of each class. Our development study varies the reconstruction ratio and the duration of conditional augmentation. The selected schedule applies SCRA for 81 sampling steps and ordinary GRPO for the remaining 54 steps.

3.3 A local contrast argument for 1:1 reconstruction

Let a and b denote the available negative and positive counts after verification, and let C be the per-class cap. We select without replacement

$$m = \min(a, b, C), \qquad |G^-| = |G^+| = m. \tag{3}$$

This is the largest balanced subset permitted by the available pools and cap. The preference for equal counts is consistent with SEELE's $p(1-p)$ motivation (Li et al., 2025), but concerns post-selection composition rather than the underlying rollout success probability. The following elementary identity makes that distinction explicit.

Proposition (conditional contrast coefficient). Consider a reconstructed group of K trajectories with binary rewards, positive fraction $p \in (0, 1)$, and advantages $A_i = R_i - p$, without standard-deviation normalization. Under equal trajectory averaging and an unclipped local surrogate, let v_i be a trajectory's local score-gradient vector, including any fixed token weights or detached behavior corrections. Then

$$g_G = \frac{1}{K} \sum_{i \in G} A_i v_i = p(1-p)(\bar{v}_+ - \bar{v}_-), \tag{4}$$

$$\frac{1}{K} \sum_{i \in G} A_i^2 = p(1-p) \le \frac{1}{4}, \tag{5}$$

where $\bar{v}_+$ and $\bar{v}_-$ are the class-conditional mean vectors. Holding these vectors fixed, 1:1 maximizes the scalar contrast coefficient per retained trajectory.

Proof. The Kp positive trajectories have advantage $1-p$, and the $K(1-p)$ negatives have advantage $-p$. Substitution yields both identities; $p(1-p) = \frac{1}{4} - (p - \frac{1}{2})^2$.

Trajectory sources determine the remaining gradient. In a reconstructed 1:1 group, let n_0 of the m selected positives come from original sampling and the rest from conditional sampling. For nonempty source subsets, Equation 4 becomes

$$g_G = \frac{1}{4} \left(\frac{n_0}{m} \bar{v}_+^0 + \frac{m - n_0}{m} \bar{v}_+^h - \bar{v}_-^0 \right). \tag{6}$$

Terms with no selected trajectories are omitted. Balance fixes the scalar coefficient; reconstruction selects the mixture of original and conditional successes and compares it with original failures. Balancing a separate hinted group produces different class means. This distinction motivates the hinted-group control in Appendix D. Our ratio study evaluates complete reconstruction rules because changing the ratio also changes selected trajectories and group size. Appendix C gives the loss-aggregation and advantage-normalization details.

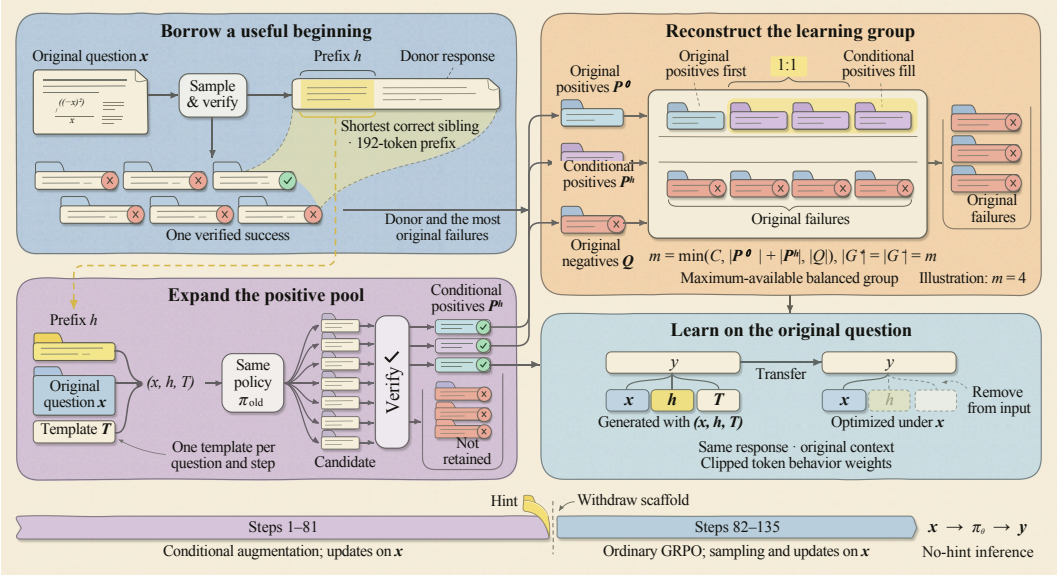


Figure 1: **SCRA in one training step, then the GRPO tail.** A verified sibling donates a 192-token prefix. The same policy expands the positive pool under one template $T(x, s)$. Maximum-available 1:1 reconstruction keeps original failures and fills positives with original successes first. Selected responses are rebased onto the original question. Conditional sampling runs through step 81; steps 82–135 are ordinary GRPO; evaluation is always no-hint.

4 Conditional Sampling and Balanced Reconstruction

4.1 Generate candidates from sparse successes

For each question x, SCRA samples $N = 8$ ordinary responses and evaluates their binary rewards. During the augmentation phase, questions with at least one correct sibling and an original success fraction below 0.5 receive additional sampling. We select the shortest correct sibling, remove an explicit final-answer suffix where possible, and retain a prefix of $B = 192$ tokens. The same policy generates $M = 8$ complete candidate solutions from this prefix. Questions with no verified sibling keep the original group: a prefix cannot be formed, and forcing additional sampling on an all-wrong problem is unlikely to yield a usable scaffold.

The prefix is presented using one of eight instruction templates. A hash of the problem identifier and training step selects a template $T(x, s)$, shared by all conditional responses for that question at that step. We call this per-question (per-UID) assignment:

$$y_j^h \sim \pi_{\theta_{\text{old}}}(\cdot \mid x, h, T(x, s)), \qquad j = 1, \ldots, M. \tag{7}$$

The exact templates are listed in Appendix B. They vary the wording of the request while keeping the question and sibling prefix fixed.

4.2 Reconstruct the maximum-available 1:1 group

Let P^0 contain original-correct responses, P^h conditional-correct responses, and Q original-incorrect responses. We set the pair cap to $C = N = 8$ and choose

$$m = \min\big(C, |Q|, |P^0| + |P^h|\big). \tag{8}$$

Since $|Q| \le N$, this cap does not reduce the available balanced group: $m = \min(|Q|, |P^0| + |P^h|)$. When P^h is nonempty, we keep m negatives without replacement and m positives. Positive selection gives priority to original-correct responses in their existing order, then fills the remaining slots with conditional-correct responses. The resulting group is the maximum-available balanced group. If no conditional response succeeds, the original group is used.

Equation 5 characterizes binary outcome contrast for the constructed group. The default rule uses all available pairs at a 1:1 ratio. Experimental settings and the distributed batch handling of completed runs are specified in Appendix C.

4.3 Update the original-question policy

A conditional response is generated with the question, the sibling prefix, and the selected template in context. The response itself is a complete solution, not a continuation that still depends on the prefix. Before the update, the prefix and the template are removed from the conditioning context, and the same full response is scored under the original question alone. Selected conditional responses are therefore rebased from (x, h, T) to x. Original trajectories receive behavior weight one. Conditional tokens use the clipped ratio

$$w_{j,t} = \mathrm{clip}\left(\frac{\pi_{\theta_{\mathrm{old}}}(y_{j,t} \mid x, y_{j,<t})}{\pi_{\theta_{\mathrm{old}}}(y_{j,t} \mid x, h, T, y_{j,<t})}, 0, c\right), \tag{9}$$

where $c = 1.2$. This tokenwise correction adjusts conditional trajectories toward the original-question policy. Let

$$\rho_{j,t} = \frac{\pi_\theta(y_{j,t} \mid x, y_{j,<t})}{\pi_{\theta_{\mathrm{old}}}(y_{j,t} \mid x, y_{j,<t})}, \qquad A(y_j) = R(y_j) - \bar{R}_x, \tag{10}$$

where $\bar{R}_x$ is the mean reward of the group used for the update. The reported SCRA configuration uses unnormalized advantages and the clipped objective

$$\mathcal{L}_{j,t} = -w_{j,t} \min(\rho_{j,t} A(y_j), \mathrm{clip}(\rho_{j,t}, 1-\epsilon, 1+\epsilon) A(y_j)). \tag{11}$$

The behavior weight is a local token correction; it does not account for every change in the trajectory distribution introduced by verification and selection.

4.4 Finish with ordinary GRPO

The main mathematical configuration trains for 135 sampling steps. SCRA is active through step 81; from step 82 through step 135, sampling and updates use the original questions without conditional generation or group reconstruction. The optimizer continues across this transition. Inference throughout evaluation uses the original question alone.

5 Experiments

The experiments evaluate final no-hint reasoning quality and the effect of positive–negative allocation.

Main mathematical comparison. The main protocol uses Qwen3-8B on DAPO-Math-17k with prompt batch 128, eight original responses per question, and 135 sampling steps. The training file contains 17,393 entries after removing five entries that exactly match evaluation questions. Final evaluation uses all 500 MATH500 questions and all 30 questions each from AIME24 and AIME25, with 16 no-hint responses per question, temperature 1.0, top-p 0.95, and a 4,096-token response limit. Chemistry, physics, biology, and materials are trained and evaluated separately on the official held-out SciKnowEval splits (100 sampling steps, one seed; augmentation for the first 60 steps, then GRPO). Mathematical GRPO, HiPO, and SCRA rows report the mean and standard deviation over three training seeds. Science columns and SDPO are single runs.

SCRA configuration. SCRA uses $C = N = 8$ for maximum-available 1:1 reconstruction, a partial-only trigger $0 < n^+/N < 0.5$, a 192-token sibling prefix, eight conditional responses, and one of eight templates selected per UID. Conditional augmentation is active for steps 1–81, followed by GRPO for steps 82–135. Advantages are group-centered without standard-deviation normalization. The configuration and implementation details of completed development runs are given in Appendix C.

Table 1: **No-hint evaluation with Qwen3-8B.** Mean@16 accuracy (%). Mathematical GRPO, HiPO, and SCRA entries are means over three training seeds (standard deviations in Appendix D). Chemistry, physics, biology, and materials are single runs trained separately for 100 steps under one protocol. SDPO is a single run and collapsed on the mathematical suites. GRPO $N{=}16$ matches the extra sampling budget of triggered SCRA steps on the mathematical protocol, not wall-clock cost. Bold marks the best trained result in each column.

Method	MATH-500	AIME 2024	AIME 2025	Chemistry	Physics	Biology	Materials
Base	84.29	25.00	21.67	42.98	59.77	30.75	60.57
GRPO $N{=}8$ (Shao et al., 2024)	89.51	36.67	28.54	70.60	69.69	54.75	74.20
GRPO $N{=}16$	89.88	37.50	29.17	–	–	–	–
HiPO (Deng et al., 2026)	**90.18**	37.71	29.58	71.04	72.89	61.13	74.67
SDPO (Hübotter et al., 2026)	64.91	4.58	2.29	70.42	64.69	49.38	73.07
SCRA	90.04	**39.17**	**31.25**	**71.48**	**73.28**	**61.63**	**75.20**

Baselines. The comparison includes the initial model, GRPO with $N{=}8$, GRPO with $N{=}16$ original samples, HiPO, and SDPO (Hübotter et al., 2026). The $N{=}16$ GRPO run matches the extra sampling budget used when SCRA triggers ($N{+}M$ rather than N original responses). HiPO is adapted to the same pure-text setting, with standard-deviation-normalized advantages and an upper PPO clip of 0.28; GRPO and SCRA use unnormalized advantages and symmetric 0.2 clipping. SDPO is a single run: on the mathematical protocol it collapsed after entropy dropped, while the separately trained science splits stayed above the base model. It stays below GRPO on all four SciKnowEval splits (Table 1).

Ablations. Table 2 reuses the Table 1 mathematical protocol: the 17,393-entry file and the full MATH-500, AIME 2024, and AIME 2025 suites. GRPO $N{=}8$ and $N{=}16$ reuse the Table 1 means as compute references; the remaining rows, including SCRA, are matched single runs at seed 1234/42. A development split on the earlier 17,398-entry file was used only to freeze the reconstruction ratio, template rule, prefix length, and 81/54 schedule; those scores are in Table 3 and Appendix D.

6 Main Results

6.1 Full-benchmark comparison

Table 1 compares the base model, GRPO, a GRPO run with $N{=}16$ original samples, HiPO, SDPO, and SCRA. SCRA is strongest on AIME 2024, AIME 2025, and the four SciKnowEval splits. HiPO is strongest on MATH-500. Every SCRA entry remains above GRPO and above HiPO on chemistry, physics, biology, and materials. The margin over HiPO is smallest on physics. GRPO $N{=}16$ narrows the mathematical gap but does not match SCRA on either AIME suite; it matches the extra sampling budget of triggered SCRA steps, not wall-clock cost, and is not repeated on the science splits. SDPO is a single run: it collapses on the mathematical suites, and on all four SciKnowEval splits it stays below GRPO. Mathematical GRPO, HiPO, and SCRA scores are means over three training seeds; the science columns are single runs. Figure 3(a) shows periodic in-training AIME 2024, separate from the Table 1 evaluation.

6.2 Balanced reconstruction is strongest on AIME

On the Table 1 mathematical protocol, 1:1 reconstruction is strongest on AIME 2024 and AIME 2025 among the matched single runs (Table 2). Extra no-hint sampling plus 1:1, and GRPO $N{=}16$, both beat GRPO $N{=}8$, but neither matches the sibling prefix on AIME. Keeping the original group size N after sibling sampling is also weaker than 1:1 reconstruction on both AIME suites. Unguided extra sampling remains below SCRA on both AIME suites, so we do not treat MATH-500 as the ablation criterion. Extending the scaffold through step 108 remains above GRPO $N{=}8$ but trails the 81/54 cutoff on both AIME suites. Section 7 discusses these allocations together with prefix length and failed hint presentations.

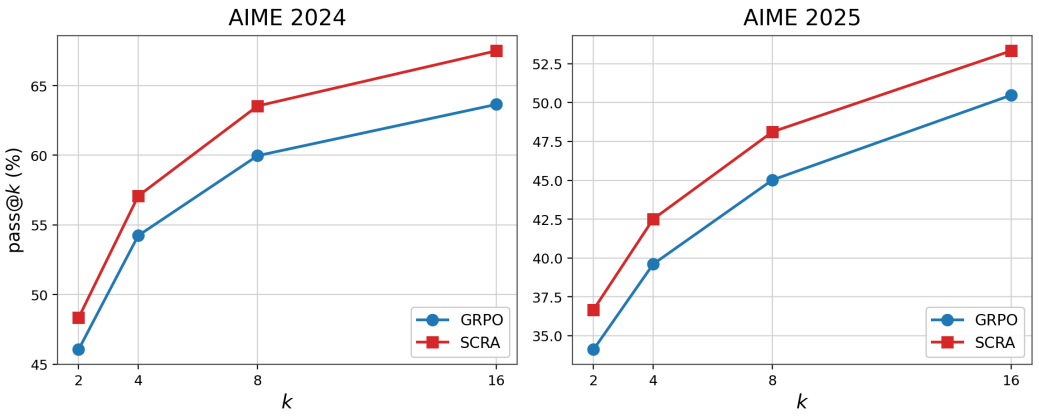


Figure 2: Pass@k (coverage of at least one success among k of 16 no-hint samples) on AIME 2024 and AIME 2025 for GRPO and SCRA. This is not mean@16: at k=16 the values sit above Table 1. On both sparse suites SCRA remains above GRPO as k grows. HiPO comparisons remain those in Table 1.

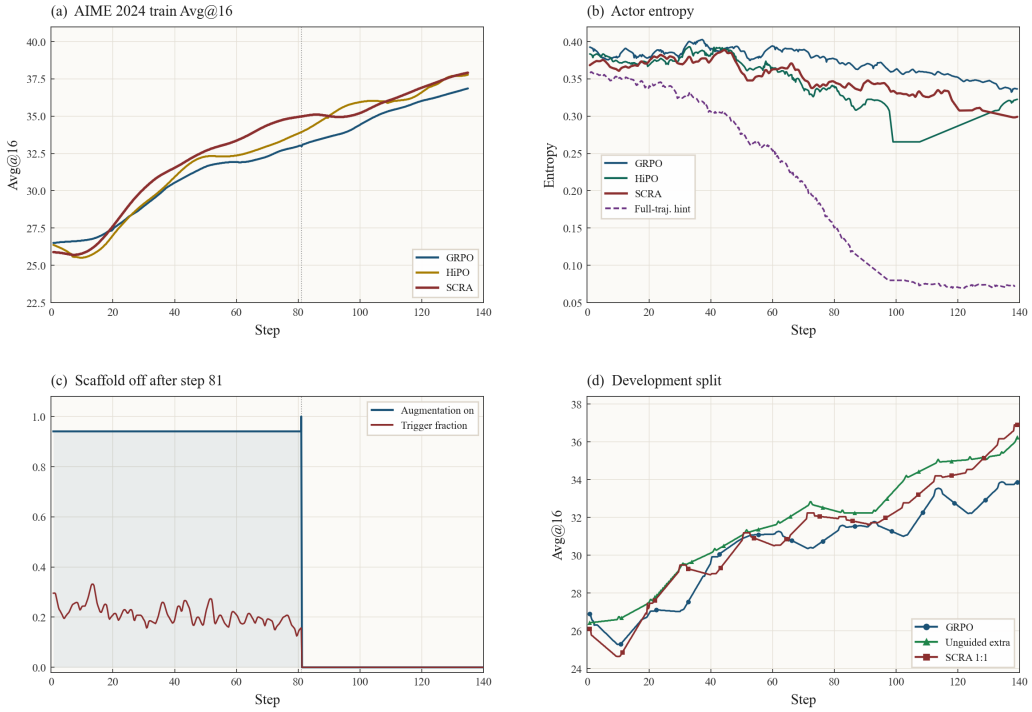


Figure 3: Training dynamics. (a) In-training AIME 2024 Avg@16. Faint traces are the per-step series and solid lines are a moving average. The dotted line marks step 81. (b) Actor entropy; the dashed curve is a full-trajectory hint run from the development study. (c) Trigger fraction and the on/off schedule; augmentation stops after step 81. (d) Development split used to freeze 1:1 reconstruction (Table 5).

6.3 Prefix length and what not to put in the hint

We compared 128, 192, 256, and 512-token sibling prefixes on the development split used to freeze the method (Table 3). 192 tokens is best: shorter prefixes give a weaker scaffold, while 512 tokens starts to look like copying the donor trajectory and the score drops. Replacing the prefix with the full

Table 2: **Ablations on the Table 1 mathematical protocol.** No-hint avg@16 (%) at step 135. Every row uses the 17,393-entry training file and the full MATH-500 / AIME 2024 / AIME 2025 suites. GRPO N=8 and N=16 reuse the Table 1 three-seed means as compute references. Remaining rows, including SCRA, are matched single runs (seed 1234/42). Bold marks the strongest AIME scores among those matched runs. Appendix D records the development split used to freeze these settings.

Configuration	MATH-500	AIME 2024	AIME 2025
GRPO N=8 (Table 1 mean)	89.51	36.67	28.54
GRPO N=16 (Table 1 mean)	89.88	37.50	29.17
Unguided extra 8 + 1:1	89.79	38.33	30.00
Sibling prefix, keep original N	89.72	37.92	29.79
Sibling prefix, 1:2	90.18	38.12	30.00
Sibling prefix, 1:3	89.61	37.08	28.96
Sibling prefix, 2:1	89.36	36.46	28.54
Sibling prefix, per rollout	90.12	37.92	29.38
Sibling prefix, 108/27	89.62	37.08	29.17
SCRA: 1:1**, per question, 81/54**	89.86	**38.96**	**31.04**

Table 3: **Prefix length and failed hint presentations** on the development split. Prefix lengths use prompt batch 32 and 100 steps; failed presentations use prompt batch 128 and 100 steps. A 192-token sibling prefix is strongest among the lengths; copying a full trajectory or asking the model to self-check collapses training.

Setting	Development avg@16 (%)
Sibling-prefix length	
128 tokens	30.75
192 tokens	**31.75**
256 tokens	31.06
512 tokens	28.31
Failed presentations (same split, batch 128)	
Full successful trajectory as hint	26.81
Strict self-check instruction	18.00
Final-answer self-check instruction	0.75

successful trajectory, or with a self-check instruction, collapses training. These runs are why SCRA conditions on a short verified prefix rather than on a complete solution or a verbal critique. The same split's reconstruction-ratio numbers are in Appendix D.

7 Reconstruction and Scaffold Duration

Balanced reconstruction is the strongest tested allocation on AIME. The 1:1 configuration is strongest on AIME in Table 2. On AIME 2024 the matched single run is 0.84 points above the best non-balanced alternative among those rows. The 1:2 and 2:1 rules share the coefficient $p(1-p) = 2/9$ but diverge on all three benchmarks. Their difference reinforces the role of selected trajectories and group size alongside outcome balance. MATH-500 is not the ablation criterion.

Unguided extra sampling is a strong but incomplete control. Eight additional no-hint samples followed by 1:1 reconstruction already beat GRPO. The sibling prefix adds 0.63 points on AIME 2024 and a similar margin on AIME 2025 at the matched seed. Extra sampling and reconstruction therefore explain part of the gain; the prefix is what remains on the success-scarce AIME suites. A keep-original-N control that uses the same sibling samples without forcing 1:1 is 1.04 points below 1:1 on AIME 2024, so reconstruction itself still contributes at this seed. On MATH-500 the unguided control remains below the matched SCRA run. Per-question template assignment outperforms per-rollout assignment on both AIME columns, so we keep the per-question rule.

Returning to unaided sampling improves the tested schedule. Switching to GRPO after 81 steps is strongest among the two tested cutoffs. Continuing augmentation through step 108 drops

AIME 2024 by 1.88 points at the common step-135 endpoint, while remaining above GRPO $N{=}8$. This experiment changes both the length of the scaffold and the length of the unaided tail, so the comparison is between the two tested schedules. The development split already froze this cutoff (Table 5); Table 2 is the same schedule decision on the full protocol. Figure 3(c) records the cut in the training logs: the trigger fraction falls to zero after step 81. The injected-positive count does the same and is shown with the reward traces in Appendix D.

Cross-source reconstruction is stronger than keeping the sources apart. The ratio rows change how many positives and negatives are kept. They do not ask whether successes sampled under the sibling prefix should share an update group with the original failures. The separate-group rule holds that fixed. It uses the same candidate generation, the same original-question update, the same tokenwise behavior weights, unnormalized advantages, symmetric 0.2 clipping, and the same $81/54$ schedule as SCRA. Original successes are balanced against original failures, and conditional successes are balanced inside their own group. The two sources are never merged, so each update sees only one class of positives.

Over three seeds the separate-group rule scores 89.92±0.19 on MATH-500, 37.92±0.74 on AIME 2024, and 30.63±0.72 on AIME 2025. It remains below SCRA on both AIME suites. The prefix and the extra candidates are already present in this control, so the gap is not explained by sampling more solutions. What changes is the group the update sees: a merged 1:1 group puts original failures next to successes from both sources, and that composition is the stronger of the two rules we tested.

8 Conclusion and Discussion

SCRA turns a short verified sibling prefix into additional candidates, reconstructs a maximum-available 1:1 group, and trains that group on the original question before returning to ordinary GRPO. It stays above GRPO on MATH-500, AIME 2024, AIME 2025, and the four SciKnowEval splits, and it is strongest on both AIME suites and on those science splits; HiPO remains slightly ahead on MATH-500. Among the allocations we tested, merged 1:1 reconstruction is what holds up on AIME, and ending the scaffold after 81 of 135 steps is stronger than keeping it on longer. Group composition and how long the scaffold stays on are therefore part of the method, not just the decision to sample more solutions.

The separate-group comparison makes the composition claim concrete. The prefix and the extra candidates are already there; what changes is whether successes from both sources share one update with the original failures. On both AIME suites, the merged group is the stronger of those two rules.

Two further comparisons keep the same conclusion from being read as "sample more." Eight extra no-hint samples followed by 1:1 reconstruction already beat GRPO, but they do not match the sibling prefix on either AIME suite. Keeping the original group size after those sibling samples is also weaker than forcing 1:1. The 1:2 and 2:1 rules share a binary-outcome coefficient and still diverge, which is what we would expect if the retained trajectories and the group size mattered along with the positive–negative balance. Per-question template assignment stays ahead of per-rollout assignment on both AIME columns, so the wording of the request is tied to the problem rather than redrawn for every sample.

The scaffold is temporary on purpose. Conditional augmentation runs for 81 steps and ordinary GRPO for the remaining 54; the training logs show the trigger fraction and the injected-positive count falling to zero at that cut. Continuing the scaffold through step 108 still beats GRPO $N{=}8$ at the shared endpoint, and it trails the earlier cutoff on both AIME suites. That comparison moves the length of the scaffold and the length of the unaided tail together. Between the two schedules we tested, returning to the original question earlier is the stronger one. The same pattern holds on the four SciKnowEval splits, which use 60 steps of reconstruction and then 40 of GRPO: SCRA stays above GRPO and above HiPO on each split, while a same-protocol SDPO run collapses on contest math and stays below GRPO on the science splits. After that cutoff, the update no longer conditions on the hint: the retained group is trained as a sample from the original question. Evaluation uses that same no-hint question, so the scaffold is a training device rather than part of the reported policy. The comparisons above are what justify treating the merged group and the early cutoff as part of that device.

AI USE STATEMENT

Generative AI tools assisted with literature discovery, methodological discussion, code development, experimental analysis, mathematical exposition, manuscript drafting and editing, and LaTeX preparation.

REPRODUCIBILITY STATEMENT

Appendix A gives the training procedure, Appendix B lists the exact templates, Appendix C records the model, data, hyperparameters, evaluation, and compute settings, and Appendices D–E provide additional results and metric definitions. Anonymized training code is included as supplementary material. A public repository will be released after the review.

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. *arXiv preprint arXiv:2306.13649*, 2023. URL https://arxiv.org/abs/2306.13649.

Qiyuan Deng, Kehai Chen, Min Zhang, and Zhongwen Xu. HiPO: Self-hint policy optimization for RLVR. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/0c0936a505992ce0d45169cd595b9755-Abstract-Conference.html.

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

Ziheng Li, Zexu Sun, Jinman Zhao, Erxue Min, Yongcheng Zeng, Hui Wu, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, Xu Chen, and Zhi-Hong Deng. Staying in the sweet spot: Responsive reasoning evolution via capability-adaptive hint scaffolding. *arXiv preprint arXiv:2509.06923*, 2025. URL https://arxiv.org/abs/2509.06923.

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.

Yu Xia, Canwen Xu, Zhewei Yao, Julian McAuley, and Yuxiong He. Learning to hint for reinforcement learning. *arXiv preprint arXiv:2604.00698*, 2026. URL https://arxiv.org/abs/2604.00698.

Wei Xiong, Chenlu Ye, Baohao Liao, Hanze Dong, Xinxing Xu, Christof Monz, Jiang Bian, Nan Jiang, and Tong Zhang. Reinforce-Ada: An adaptive sampling framework for reinforce-style LLM training. *arXiv preprint arXiv:2510.04996v1*, 2025. URL https://arxiv.org/abs/2510.04996v1.

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

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

A FULL ALGORITHM

Algorithm 1 describes SCRA with a pair budget C and augmentation cutoff S_a. The default mathematical configuration uses $C = N = 8$, $S_a = 81$, and $S = 135$, giving maximum-available 1:1 reconstruction.

Algorithm 1 SCRA with maximum-available 1:1 reconstruction and a GRPO tail

Require: Policy π_θ, verifier R, original count N, conditional count M, trigger f, prefix budget B, pair cap C, templates $\mathcal{T}$, cutoff S_a, horizon S

for sampling step $s = 1, \ldots, S$ **do**
  for question x in the prompt batch **do**
    Sample $Y^0 = \{y_i^0\}_{i=1}^N$ from the original question and compute rewards.
    $n^+ \leftarrow \sum_i R(y_i^0)$; $G_x \leftarrow Y^0$.
    if $s \le S_a$ and $0 < n^+ < fN$ **then**
      Form a B-token prefix h from the shortest verified sibling.
      Select one template $T \leftarrow T(x, s)$ for all M conditional responses.
      Sample and verify $Y^h = \{y_j^h\}_{j=1}^M$ under (x, h, T).
      $P^0 \leftarrow \{y_i^0 : R(y_i^0) = 1\}$; $P^h \leftarrow \{y_j^h : R(y_j^h) = 1\}$; $Q \leftarrow \{y_i^0 : R(y_i^0) = 0\}$.
      if $P^h \neq \varnothing$ **then**
        $m \leftarrow \min(C, |Q|, |P^0| + |P^h|)$.
        Keep m negatives without replacement.
        Keep up to m original positives, then fill remaining positive slots from P^h.
        Form G_x from the selected m positives and m negatives.
        Rebase selected conditional responses to the original context x.
  Concatenate the groups and apply the experiment's distributed batch handling.
  Compute group-centered advantages on the retained trajectories.
  Apply clipped behavior weights to selected conditional responses.
  Update the policy using the configured PPO minibatches.

Template assignment. Per-UID assignment hashes the problem identifier and global step into one template and uses it for all M conditional responses. Views vary across questions and steps.

Fallback and phase transition. The partial-only trigger requires $0 < n^+ < fN$. Questions without a verified sibling, questions at or above the trigger threshold, and triggered questions with no successful conditional response retain their ordinary groups. After S_a, all questions use ordinary groups.

B Reflection Templates

The implementation uses the following eight instruction templates. {`prompt`} is the original problem and {`hint`} is the prefix extracted from the shortest verifier-correct sibling. One template is selected for each problem and training step and shared by its conditional responses.

1. {`prompt`}
   Here is partial scratchwork from a verifier-correct attempt:
   {`hint`}
   Use it as a possible starting point and produce a complete solution. Follow the answer format required by the original problem.

2. A solver who reached the correct answer began with these notes:
   {`hint`}
   Now solve the following problem completely:
   {`prompt`}
   Follow the answer format required by the original problem.

3. Problem:
   {`prompt`}
   Potentially useful setup from a correct attempt:
   {`hint`}
   Complete the derivation carefully and give a self-contained solution.

4. {`prompt`}
   Incomplete derivation from a verifier-correct attempt:

{hint}
Finish the problem carefully and use the original answer format.

5. Candidate setup from a correct attempt:
{hint}
Original problem:
{prompt}
Use the setup if helpful and write a complete solution.

6. Solve the following problem:
{prompt}
Notes from a verifier-correct attempt:
{hint}
Give a complete derivation in the original answer format.

7. {prompt}
A verified solver started with these notes:
{hint}
Continue or revise the setup as needed, then solve the problem fully.

8. Reference beginning from a correct attempt:
{hint}
Question:
{prompt}
Produce a complete self-contained solution and respect the requested format.

The template assignment changes the phrasing and organization of the request. The reported configuration uses per-UID assignment with the same 192-token sibling prefix budget and eight conditional responses.

C Experimental Details

Table 4: Default SCRA configuration and mathematical evaluation settings. Completed development runs use the earlier training file described below.

Field	Value
Policy model	Qwen3-8B, thinking disabled
Training data, main protocol	DAPO-Math-17k, 17,393 entries
Hardware	one node, eight H800 GPUs; rollout TP=2
Prompt batch / original responses	128 / $N = 8$
Conditional responses	$M = 8$
Total sampling steps	135
Augmentation / GRPO tail	steps 1–81 / 82–135
Pair budget / reconstruction	$C = N = 8$ / maximum-available 1:1
Template assignment	one of eight templates per UID and step
Hint source / prefix budget	successful sibling / 192 tokens
Trigger configuration	partial-only, success fraction in $(0, 0.5)$
All-wrong groups	ordinary-group fallback
Behavior correction	token ratio, upper clip 1.2
Group advantage	centered, no standard-deviation normalization
Learning rate / warmup	10^{-6} / 10 steps
Weight decay / gradient clipping	0.01 / 1.0
PPO mini / micro batch	128 prompts / one trajectory per GPU
PPO epochs / clipping	1 / lower 0.2, upper 0.2
Original prompt / response limits	2,048 / 4,096 tokens
Conditional prompt limit	3,072 tokens
Training temperature / top-p	1.0 / 1.0
Evaluation responses / temperature / top-p	16 / 1.0 / 0.95
Data / model / rollout seeds	1234/42/42, 1235/43/42, 1236/44/42

Selection and distributed updates. The implementation gives original-correct trajectories priority within the positive slots and then fills the remaining slots with conditional successes. With $C = N = 8$, the pair cap does not bind: the available pools determine the maximum balanced group. The completed development runs concatenate reconstructed groups and trim a trailing remainder to make the batch divisible by the eight actor ranks. Advantages are computed after this operation. The 1:1 construction in the method specifies groups before this batch-level trimming. The stock actor uses a nominal global minibatch capacity of $128 \times 8 = 1024$ trajectories and one PPO epoch. With microbatch size one and `token-mean`, each trajectory's token loss is averaged before the fixed-capacity gradient accumulation.

Main evaluation. Final mathematical evaluation reloads the step-135 checkpoint and uses the original questions without hints. Each of MATH500, AIME24, and AIME25 contributes its full 500, 30, and 30 questions. Chemistry, physics, biology, and materials are each trained separately on a SciKnowEval split for 100 sampling steps with the same N=8 and decoding, then evaluated on a held-out portion of that split. The verifier searches the final 1,024 characters for the last boxed answer and applies the shared mathematical answer check; science uses the dataset's official checker. Reported mathematical GRPO, HiPO, and SCRA scores are the mean and standard deviation over the three seeds above. Science columns and SDPO are the single runs in Table 1.

Development protocol. A development split on the earlier 17,398-entry file was used to freeze reconstruction ratio, template assignment, and the 81/54 schedule (Table 5). Table 2 then evaluates those frozen choices on the main mathematical protocol; the two tables are not interchangeable.

D Additional Completed Results

Seed variation. Mathematical GRPO, HiPO, and SCRA in Table 1 are means over three training seeds. The standard deviations are 0.18 / 0.83 / 0.75 for GRPO N=8, 0.20 / 0.72 / 0.83 for GRPO N=16, 0.21 / 0.75 / 0.72 for HiPO, and 0.22 / 0.75 / 0.83 for SCRA, on MATH-500, AIME 2024, and AIME 2025. Science columns are single runs and have no seed standard deviation.

SDPO. We adapted SDPO (Hübotter et al., 2026) to the same Qwen3-8B protocol. Table 1 reports the single runs: on contest math it collapsed after entropy dropped, while the separately trained chemistry, physics, biology, and materials splits stayed above the base model and below GRPO. We therefore treat SDPO as a same-family control rather than a competitive mathematical baseline. The science columns are single 100-step runs under one protocol and are not the three-seed mathematical protocol. On materials, reconstruction fires on a smaller fraction of questions than on biology or chemistry, and SCRA is the strongest trained run.

HiPO hinted-group variant. Section 7 compares merged reconstruction with the separate-group rule under the same candidate generation, original-question update, and $81/54$ schedule. The separate-group scores are reported there. A hinted group that is balanced on its own has different class means from a group that mixes original failures with successes from both sources; that is the distinction the separate-group comparison is meant to test.

Development split. Table 5 is a development split used only to freeze the 192-token prefix, the per-question template rule, and the 81/54 schedule. It uses the earlier 17,398-entry training file. Table 2 then evaluates those frozen choices on the full mathematical protocol; the two tables are not the same evaluation, and we do not claim that the development ranking copies onto the full suites.

Ratio-study budgets. The 1:2 and 1:3 configurations use ratio-based maximum-available reconstruction. The 2:1 configuration additionally caps the reconstructed group at the original per-question sample budget. The keep-original-N row in Table 2 uses the same sibling samples and original-question rebase, but retains the original group size instead of forcing 1:1. The ratio rows therefore compare complete reconstruction rules with different retained sample counts; they do not use a common fixed candidate pool.

Table 5: **Preliminary development split** used to freeze reconstruction ratio, template assignment, and scaffold duration. Scores are step-135 avg@16 (%) on a contest subset from the earlier 17,398-entry training file. This split is not the Table 1 AIME 2024 protocol, and its ranking is not assumed to copy onto Table 2.

Configuration	Development avg@16 (%)
SCRA reference: 1:1, per question, 81/54	**36.50**
Positive:negative reconstruction	
1:2	35.00
1:3	33.63
2:1 (total-count cap)	33.25
Sampling and schedule controls	
GRPO	33.63
Unguided extra 8 + 1:1, 81/54	35.88
Sibling prefix, per rollout, 81/54	34.88
Sibling prefix, per question, 108/27	34.75

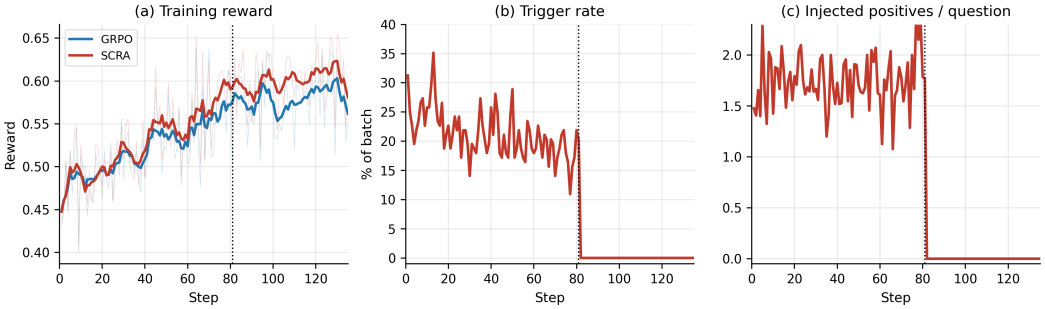


Figure 4: Training reward on original-question groups, fraction of the batch that triggers reconstruction, and mean injected positives per triggered question. The dotted line is the step-81 switch to ordinary GRPO. Reward is a training diagnostic, not the Table 1 evaluation.

E Metric Definitions

No-hint avg@16. For Q evaluation questions and 16 sampled responses per question, the reported percentage is

$$\text{avg@16} = \frac{100}{16Q} \sum_{q=1}^{Q} \sum_{j=1}^{16} R(y_{q,j}). \tag{12}$$

Each question has equal weight in the full benchmark evaluations. This is the mean single-response success rate under the evaluation sampler. Pass@16, which asks whether any of the 16 responses succeeds, is a different quantity. The mathematical main table uses avg@16.

Development scores. The development score is computed on the development split and is not substituted for Table 2.

Group composition and reward variance. The positive fraction p is measured within a group of trajectories used together to compute advantages. For binary rewards and $A_i = R_i - p$, $|G|^{-1} \sum_i A_i^2 = p(1-p)$. An ideal 1:1 reconstruction has $p = 1/2$. A pair cap limits group size, while distributed trimming can change the composition actually used by an evaluated implementation; construction-level formulas and actual update groups are distinguished accordingly.

Conditional success and generation cost. Conditional success is the verifier-correct fraction among the additional responses. Original and conditional response tokens are counted separately when measuring generation cost. A sampling step denotes one prompt batch and its generated trajectories; its number of optimizer updates depends on the resulting actor minibatches. GRPO

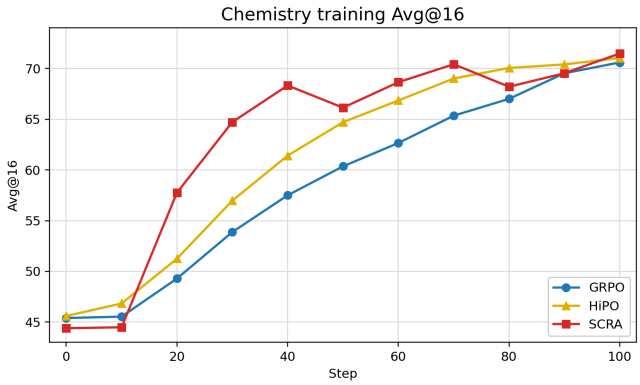


Figure 5: In-training chemistry Avg@16 on SciKnowEval. SCRA rises earlier than GRPO. The held-out ranking is Table 1, where SCRA is slightly ahead of HiPO.

N=16 matches the extra sampling budget of a triggered SCRA step ($N+M$), not total tokens, GPU time, or optimizer updates: reconstruction fires only on a subset of questions and only during steps 1–81.