# SIMPLE: VERIFIABLE STUDENT-IMPROVEMENT FEEDBACK FOR PRETRAINING-TIME LEARNING BY TEACHING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Learning-by-Teaching (LbT) requires feedback on whether a teacher helps a student, yet raw pretraining text provides no such feedback. Neither token-level proxy gaps nor hint validity alone can provide this feedback signal, while raw student improvement can be inflated by answer leakage. We introduce **Student Improvement Mode of Pseudo Learning by Teaching** (SIMPLE), a framework that constructs source-grounded teaching events and derives teacher-training feedback from answer-masked likelihood gains in frozen students. SIMPLE measures teaching utility as answer-masked gains in frozen students' answer log-likelihood, and grounding and anti-hacking checks decide which hints may receive credit at all. Coverage correction assigns zero utility to events without a valid candidate, so the metric jointly accounts for valid-hint availability and the usefulness of selected hints. The teacher learns through supervised hint imitation followed by reward-filtered preference refinement that prioritizes more useful hints over valid but less useful alternatives. Across eight benchmark-derived evaluations and two small-model backbones, SIMPLE attains the highest coverage-corrected teaching utility among the evaluated methods.

## 1 INTRODUCTION

Learning-by-Teaching (LbT) improves a teacher through feedback about its effect on a student (Ning et al., 2024). Pretraining-style training breaks this principle: raw text provides material to teach, but neither explicit teaching tasks nor observations of student improvement. Next-token prediction supplies a learning signal for modeling the text, not for assessing a generated hint (Korbak et al., 2023; Cheng et al., 2024). We study *pretraining-time LbT* for small language models and construct teaching feedback from source-grounded question–answer–evidence events using frozen students' answer likelihoods.

One way to approximate the missing feedback is to use an internal weak proxy and minimize its excess token-level negative log-likelihood (NLL) relative to the full model. This objective measures a predictability gap, not the effect of a generated hint, and our pilot confirms that it is the wrong target: NLL-gap pLbT reduces its proxy gap, yet longer LM-only training at the same estimated compute yields lower held-out LM loss (Appendix F). The feedback must instead be tied to what teaching is for: a hint's effect on a student's answer likelihood (Ning et al., 2024; Zhou et al., 2026a).

Directly rewarding raw student improvement is equally inadequate (Gao et al., 2023). A hint that discloses the answer increases answer likelihood without teaching anything (Dinucu-Jianu et al., 2025), and a hint can pass every grounding and non-leakage check while contributing nothing to the student's answer. As Figure 1 illustrates, *apparent improvement does not establish valid teaching, and validity does not establish utility*. Evaluation must further account for whether valid hints can be produced at all: averaging only over successful selections silently discards every event without a valid candidate (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017). Reliable teaching feedback is therefore necessarily a joint problem of utility, validity, and coverage; a signal that omits any one of the three can be gamed, and Section 3 shows each failure mode occurring in practice.

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

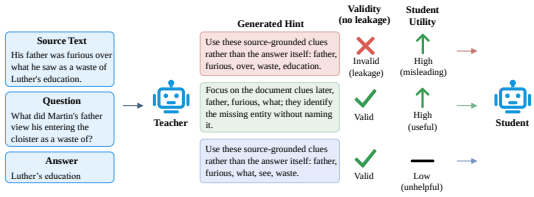



Figure 1: **Why teaching feedback requires both validity and student utility.** The same source-grounded question illustrates three cases: answer leakage that inflates apparent improvement, a valid hint that helps the student, and a valid hint that provides little benefit. Validity determines which hints can receive credit, while student improvement distinguishes their usefulness.

We introduce **Student Improvement Mode of Pseudo Learning by Teaching** (SIMPLE), a framework for verifiable teaching feedback on source-grounded teaching events. Each source-grounded teaching event contains a question, a reference answer, and grounding text. The teacher generates short hints, whose utility is measured by the gain in frozen students' reference-answer log-likelihood after masking answer strings in each hint. Grounding and anti-hacking checks decide which hints can receive credit at all; valid hints are then ranked by masked improvement together with verifier and source-cue terms, and events without a valid candidate contribute zero to coverage-corrected utility. The resulting signal credits a hint only when it is valid and only in proportion to how much it helps, while charging the teacher for every event on which it fails to produce a usable hint.

To learn from this feedback, SIMPLE combines supervised hint imitation with reward-filtered preference refinement using a SimPO-style objective (Rafailov et al., 2023; Meng et al., 2024). Refinement prioritizes valid but lower-utility negatives when available, so the teacher is trained to distinguish more useful hints from valid but less useful alternatives rather than merely to avoid violations. The objective uses validity-constrained student utility to construct supervision aligned with the evaluation metric. Our contributions are as follows:

- We formulate pretraining-time LbT through *validity-constrained student improvement*, shifting the teaching signal from token-level proxy gaps to the utility of generated hints. Our formulation separates utility, validity, and coverage, making both hint usefulness and valid-hint availability explicit in teaching credit.
- We propose SIMPLE, which learns hint generation on source-grounded teaching events using answer-masked feedback from frozen students. Its reward-filtered preference refinement prioritizes comparisons between high-utility valid hints and valid but less useful alternatives, directing teacher learning toward *usefulness within the valid set*.
- Across eight benchmark-derived evaluations and two small-model backbones, SIMPLE achieves the highest coverage-corrected teaching utility among the evaluated methods. Utility-based selection consistently outperforms verifier-only selection in same-student evaluation, demonstrating the value of student feedback beyond validity checks.

## 2 Method

**Overview.** SIMPLE trains a teacher to generate hints that help frozen students without disclosing the answer. As shown in Figure 2, student feedback selects hint-imitation targets and constructs preference pairs for offline teacher training. The feedback separates three questions: *validity* determines whether a hint is eligible for teaching credit; *utility* measures its effect on a student's answer likelihood; and *coverage* accounts for evaluation events without a valid hint. The teacher first learns to imitate selected hints, then learns to prefer more useful hints over valid but less useful alternatives. Only teacher parameters are updated.

### 2.1 Candidate Generation and Scoring

**Teaching events and hints.** Each teaching event is represented as $z = (d, q, a, h)$: source text $d$, question $q$, reference answer $a$, and grounding text $h$ used for validity checks. The representation accommodates answer spans and answer texts; source-specific construction, including the question-and-option context used for ARC, is described in Appendix E. The teacher samples $K$ hints, $e_1, \ldots, e_K \sim \pi_\theta(\cdot \mid \mathrm{TeachPrompt}(d, q))$, forming $\mathcal{C}_K(z) = \{e_1, \ldots, e_K\}$. The prompt requests a short source-grounded hint without naming the answer. The teacher receives the source and question, whereas the student receives only the question with or without the hint.

**Answer-masked student improvement.** To measure a hint's effect without exposing exact answer strings, we apply a deterministic answer-masking operator $M_a$ before student scoring. For a frozen

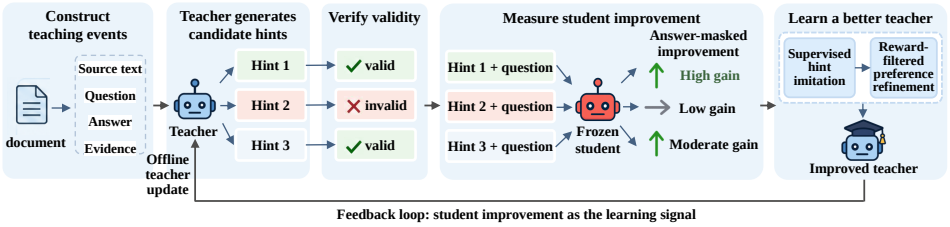



Figure 2: Overview of SIMPLE. Frozen students provide answer-masked feedback for validity-constrained supervision, which trains the teacher through hint imitation followed by preference refinement. Only teacher parameters are updated.

student $s$, masked improvement is

$$\Delta_s^{\text{mask}}(e; z) = \log p_s\big(a \mid q, M_a(e)\big) - \log p_s(a \mid q). \tag{1}$$

Both terms score the same reference answer under the same student; only the hint changes. This measures the immediate effect of a hint, not improvement from updating the student. For $M$ frozen *reward students* $\{s_j\}_{j=1}^{M}$, training uses the clipped mean

$$\overline{\Delta}^{\text{mask}}(e; z) = \frac{1}{M} \sum_{j=1}^{M} \text{clip}\big(\Delta_{s_j}^{\text{mask}}(e; z), -c, c\big), \tag{2}$$

where $c$ is the clipping bound. This statistic filters training candidates and enforces utility gaps between preference pairs; ranking uses a separate rank-normalized signal.

### 2.2 Validity-Constrained Preference Construction

**Eligibility for teaching credit.** Let $\mathcal{E}(z)$ denote the hint space and $\mathcal{A}(z) \subseteq \mathcal{E}(z)$ the hints passing the rule-based validity checks. These checks require no detected answer leakage, source grounding without excessive copying, informative non-answer cues, and a usable hint format (Appendix D). The valid sampled pool is $\mathcal{F}_K(z) = \mathcal{C}_K(z) \cap \mathcal{A}(z)$. For training, we further restrict this pool to $\mathcal{F}_K^{\text{train}}(z) = \{e \in \mathcal{F}_K(z) : \overline{\Delta}^{\text{mask}}(e; z) \geq \delta_{\text{mask}}\}$. This utility threshold is separate from validity and is not applied during evaluation.

**Choosing training targets.** The training reward combines student feedback with support and anti-hacking features:

$$R(e; z) = U_{\text{rank}}^{\text{train}}(e; z) + \alpha^\top \psi_{\text{support}}(e, z) - \lambda^\top \psi_{\text{hack}}(e, z). \tag{3}$$

Here $U_{\text{rank}}^{\text{train}}$ averages the reward students' normalized ranks of clipped masked gains, computed over the scored proposal pool before filtering. The feature terms favor source support and penalize leakage, copying, and format violations, with weights $\alpha$ and $\lambda$. Appendix D gives the feature expansion; Appendix D.5 specifies rank aggregation. When $\mathcal{F}_K^{\text{train}}(z) \neq \varnothing$, we select

$$e^+(z) \in \arg\max_{e \in \mathcal{F}_K^{\text{train}}(z)} R(e; z). \tag{4}$$

An invalid hint cannot become a chosen target regardless of its reward. Events with no eligible target supply no hint-imitation target or preference pair.

**Learning usefulness within the valid set.** Rejected hints include *unsafe negatives*, which violate validity constraints, and *valid-low-utility negatives* from $\mathcal{F}_K^{\text{train}}(z)$. SIMPLE prioritizes valid-low-utility negatives when available, so refinement learns which valid hints help more rather than only which hints pass the checks. Every retained pair satisfies $R(e^+; z) - R(e^-; z) \geq \delta_R$. Valid-low-utility pairs additionally require

$$\overline{\Delta}^{\text{mask}}(e^+; z) - \overline{\Delta}^{\text{mask}}(e^-; z) \geq \delta_{\text{utility}} > 0.$$

Thus, both hints in these pairs meet the training eligibility criteria, while the chosen hint has higher masked utility as well as the required reward advantage.

### 2.3 TEACHER OPTIMIZATION

**Hint imitation.** Let $\mathcal{D}_{\text{hint}}$ contain the selected targets $(z, e^+)$ and $\mathcal{D}_{\text{pref}}$ the retained triples $(z, e^+, e^-)$. The first stage learns to generate the selected hints through supervised imitation:

$$\mathcal{L}_{\text{hint}}(\theta) = -\mathbb{E}_{(z,e^+)\sim\mathcal{D}_{\text{hint}}} \log \pi_\theta(e^+ \mid d, q). \tag{5}$$

**Reward-filtered preference refinement.** Starting from the Hint-SFT checkpoint, we apply a reference-free SimPO-style objective (Meng et al., 2024). Let $r_\theta(e \mid z) = |e|^{-1} \log \pi_\theta(e \mid d, q)$ be the length-normalized hint log-probability, with $|e|$ the number of hint tokens. The preference loss is

$$\mathcal{L}_{\text{pref}}(\theta) = -\mathbb{E}_{(z,e^+,e^-)\sim\mathcal{D}_{\text{pref}}} \left[\log \sigma\big(\beta\big(r_\theta(e^+ \mid z) - r_\theta(e^- \mid z) - \gamma\big)\big)\right], \tag{6}$$

where $\sigma$ is the logistic function, $\beta$ the scale, and $\gamma$ the preference margin. Both stages retain causal language-modeling training:

$$\mathcal{L}_{\text{SIMPLE}}(\theta) = \mathcal{L}_{\text{LM}}(\theta) + \lambda_{\text{hint}}\mathcal{L}_{\text{hint}}(\theta) + \lambda_{\text{pref}}\mathcal{L}_{\text{pref}}(\theta). \tag{7}$$

Hint-SFT sets $\lambda_{\text{pref}} = 0$; refinement enables the preference loss while retaining a small hint-imitation anchor. Student scores are computed during data construction, before the corresponding teacher updates; gradients do not pass through the frozen students.

### 2.4 HINT SELECTION AND COVERAGE-CORRECTED UTILITY

**Evaluation-time hint selection.** Hard-pass reward-best selects the hint maximizing $R_{\text{sel}}(e; z)$ over $\mathcal{F}_K(z)$, returning $e^\star(z) = \varnothing$ when the valid pool is empty. The score combines normalized ranks from *selection students* with evaluation-specific support and penalty terms. Unlike training, it ranks *unclipped* masked gains within the valid pool and applies no training-only utility threshold (Appendix E.3). The selected hint is then scored by the *evaluation student*. These roles coincide in same-student evaluation; held-out-student evaluation excludes the evaluation student from candidate selection.

**Accounting for hint availability.** For $n$ evaluation events, let $\mathcal{I} = \{i : e_i^\star \neq \varnothing\}$ and $\widehat{C} = |\mathcal{I}|/n$. When $|\mathcal{I}| > 0$, let $\widehat{\Delta}^{\text{mask}}_{\text{valid},s}$ denote the mean masked gain of evaluation student $s$ over valid selections. The coverage-corrected utility is

$$\widehat{\Delta}^{\text{mask}}_{\text{cov},s} = \frac{1}{n}\sum_{i\in\mathcal{I}} \Delta^{\text{mask}}_s(e_i^\star; z_i) = \widehat{C}\,\widehat{\Delta}^{\text{mask}}_{\text{valid},s}. \tag{8}$$

Every event remains in the denominator, so events without a valid hint contribute zero. If $\mathcal{I}$ is empty, coverage-corrected utility is zero and the valid-subset mean is undefined. We report the evaluation student's unclipped masked gain, not the composite selection score; negative gains from valid hints remain negative. The student subscript is omitted when clear from context.

## 3 EXPERIMENTS

### 3.1 EXPERIMENTAL SETUP

*Models and data.* We continue training Qwen3-1.7B (Yang et al., 2025) and Gemma-3-1B-IT (Gemma Team, 2025) teachers on SQuAD (Rajpurkar et al., 2016), the sole training source. Here, *pretraining-time LbT* denotes language-modeling continuation augmented with teaching-feedback losses. Training events use dataset-provided questions and answers, with a cloze fallback (Appendix E.1). The main evaluation uses 1,000 events from each of eight sources: SQuAD, AdversarialQA (Bartolo et al., 2020), answerable SQuAD-v2 (Rajpurkar et al., 2018), RACE-Middle/High (Lai et al., 2017), SciQ (Welbl et al., 2017), and ARC-Easy/Challenge (Clark et al., 2018). The seven non-SQuAD sources are used only for evaluation.

*Comparisons.* We distinguish *training-objective* comparisons from *selection-rule* comparisons. The former compares SIMPLE with Base, LM-only, LM-Matched, Hint-SFT, Hint-SFT+AR, and SimPO under the same selection rule (Table 1). The latter fixes the teacher and compares reward-best with verifier-best, policy-first, faithful-reward-best, and faithful-cascade (Figures 3 and 4). Loss weights, baseline definitions, and training budgets are given in Appendix E.2.

Table 1: **Coverage-corrected teaching utility across eight sources and two backbones.** Answer-masked improvement $\Delta_{\text{cov}}$ (higher is better), with 1,000 events per source and hard-pass reward-best selection for trained teachers. Avg. is the unweighted source mean; bold marks the best trained result within each backbone and column. Colors show differences from the corresponding Base: $-6.5$ $-3.25$ $0$ $+3.25$ $+6.5$. RACE-M/H denote RACE-Middle/High. Compare within backbone, as evaluation-student log-likelihood scales differ. Baseline definitions appear in Appendix E.

| Model | Method | AdvQA | SQuAD | SQuAD2 | SciQ | RACE-M | RACE-H | ARC-E | ARC-C | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|
| Qwen3-1.7B | Base | -0.034 | -0.043 | -0.109 | -0.042 | -0.148 | -0.126 | 0.126 | 0.128 | -0.031 |
| | LM-only | -0.138 | -0.048 | -0.095 | -0.088 | -0.140 | -0.157 | 0.119 | 0.163 | -0.048 |
| | LM-Matched | -0.122 | -0.032 | -0.083 | -0.075 | -0.089 | -0.188 | 0.148 | 0.176 | -0.033 |
| | Hint-SFT | 1.350 | 1.601 | 2.177 | 3.894 | 0.323 | 0.002 | 1.001 | 0.818 | 1.396 |
| | Hint-SFT+AR | 1.320 | 1.582 | 2.033 | 3.714 | 0.241 | -0.006 | 0.950 | 0.788 | 1.328 |
| | SimPO | 1.291 | 1.614 | 2.027 | 3.628 | 0.278 | 0.050 | 0.803 | 0.663 | 1.294 |
| | **SIMPLE** | **1.585** | **1.970** | **2.438** | **4.065** | **0.667** | **0.112** | **1.051** | **0.843** | **1.591** |
| Gemma-3-1B-IT | Base | 0.133 | 0.066 | 0.143 | 0.061 | 0.033 | 0.020 | 0.070 | 0.062 | 0.074 |
| | LM-only | 0.084 | 0.071 | 0.081 | 0.005 | 0.120 | 0.086 | 0.025 | 0.041 | 0.064 |
| | LM-Matched | 0.135 | 0.072 | 0.081 | 0.008 | 0.166 | 0.048 | 0.077 | 0.098 | 0.086 |
| | Hint-SFT | 4.711 | 3.956 | 4.470 | 4.317 | 2.400 | 1.378 | 5.479 | 3.805 | 3.814 |
| | Hint-SFT+AR | 3.044 | 2.748 | 3.151 | 2.665 | 1.760 | 1.079 | 4.247 | 2.871 | 2.696 |
| | SimPO | 4.206 | 3.708 | 4.118 | 3.731 | 2.159 | 1.070 | 5.334 | 3.515 | 3.480 |
| | **SIMPLE** | **5.075** | **4.624** | **5.209** | **4.555** | **3.004** | **1.505** | **6.134** | **4.351** | **4.307** |

*Evaluation.* The main results use hard-pass reward-best selection and report coverage-corrected, answer-masked teaching utility, $\Delta_{\text{cov}}^{\text{mask}}$ (Section 2.4). For each backbone, a frozen student of that backbone both selects hints and evaluates utility; held-out evaluations exclude the evaluation student from selection. Training-objective comparisons share masking, validity filters, and coverage correction; Section 3.3 varies these components explicitly. We also report selection coverage, detected leakage, source-copy ratio, and faithfulness. Reported 95% intervals use row-level bootstrapping (Appendix E.5).

### 3.2 Teaching Utility and Hint Selection

*Teaching Utility Improves beyond the Training Source.* SIMPLE achieves the highest coverage-corrected teaching utility on every evaluation source for both backbones (Table 1). Its mean gains over Hint-SFT, the strongest baseline by average utility, are 0.195 on Qwen and 0.493 on Gemma. More importantly, the advantage extends from SQuAD, the sole training source, to all seven evaluation-only sources spanning different QA settings. These results support transfer of the learned hint-generation behavior beyond the training source under the same evaluation settings.

*Utility Gains Do Not Require Higher Selection Coverage.* On Qwen, SIMPLE improves coverage-corrected utility over Hint-SFT on every source despite slightly lower valid-selection coverage (Figure 3; Appendix Table 3). By the decomposition in Eq. (8), this combination implies higher average masked utility among the selected valid hints.

*Validity Does Not Guarantee Helpfulness.* Reward-best consistently outperforms verifier-best in the Qwen evaluations (Figure 3). The shared-pool comparison in Figure 4 holds the teacher, valid candidates, and coverage fixed, isolating the effect of the ranking rule. On both RACE sources, verifier-best yields negative utility, whereas reward-best yields positive utility. Passing validity checks therefore does not ensure that a selected hint helps the student: student feedback is needed to distinguish usefulness among eligible candidates. This conclusion concerns same-student selection; Section 3.4 examines transfer when the evaluation student is excluded from selection.

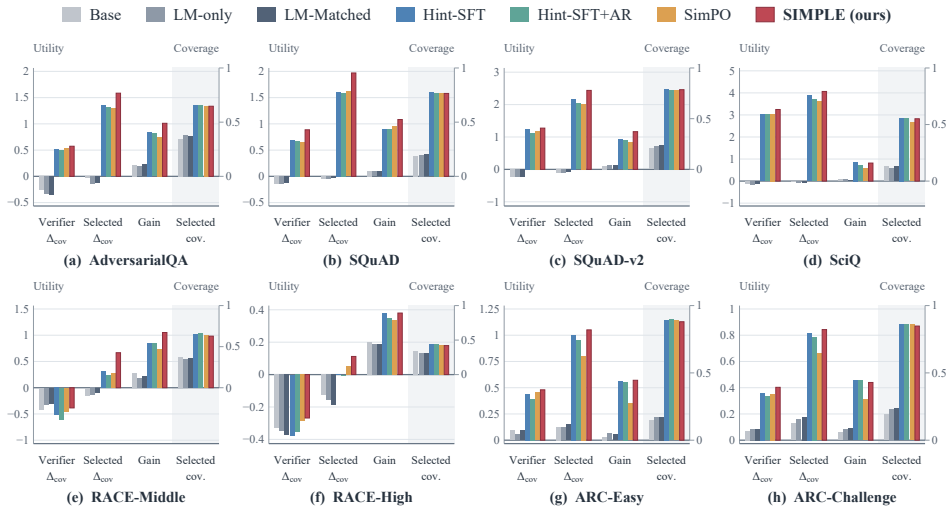



Figure 3: **Reward-best selection beats verifier-best in all 56 method–source combinations.** Selection results for seven Qwen3-1.7B teachers across eight evaluation sources (1,000 events per source). The left axis reports verifier-best utility, reward-best utility, and their difference; the right axis reports reward-best selection coverage. All utilities are answer-masked and coverage-corrected, and no answer leakage was detected among valid selections in any combination. Red bars identify SIMPLE, which has the highest reward-best utility on every source. Utility scales vary by source. Full numerical results appear in Table 3.

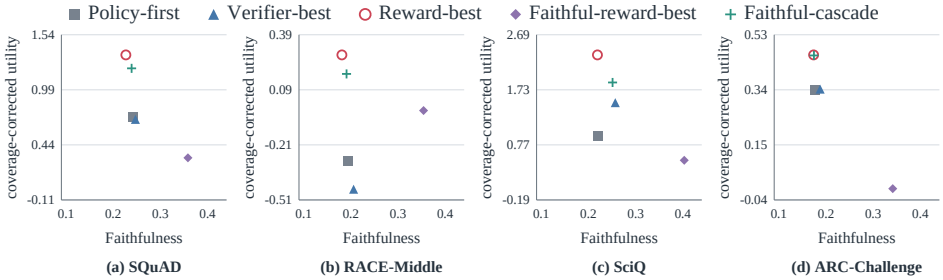



Figure 4: **Reward-best (red circles) attains the highest utility on every source at the same coverage as verifier-best.** Four ranking rules sharing one candidate pool, plus the original Policy-first reference; fixed pre-RFT SimPO teacher; horizontal axis: faithfulness of valid hints; vertical axis: $\Delta_{\text{cov}}$. All eight sources: Appendix Figure 10, Table 4.

*Selection Trade-offs and Search Efficiency.* Stricter grounding does not necessarily improve overall teaching utility: tightening the faithfulness threshold increases measured faithfulness but sharply reduces valid-hint coverage, lowering coverage-corrected utility in the evaluated settings. Candidate search also shows diminishing returns: most of the observed utility gain is attained by $K = 16$, with smaller improvements at larger budgets. Reward-best exceeds verifier-best at every evaluated $K > 1$, indicating that its selection advantage is not confined to a single candidate budget (Figure 5; Appendix G).

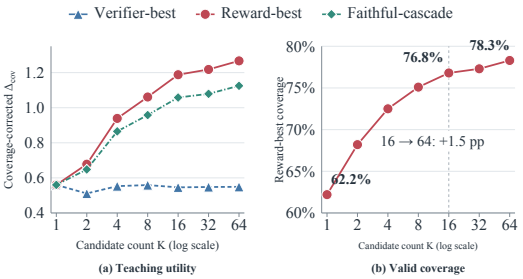



Figure 5: **More candidates benefit utility-aware selection, with diminishing gains at larger budgets.** Candidate-budget sensitivity under hard-pass selection. Panel (a) reports coverage-corrected masked utility for three selection strategies; panel (b) reports reward-best valid-selection coverage. Curves connect point estimates at the evaluated budgets on a base-2 logarithmic $K$ axis.

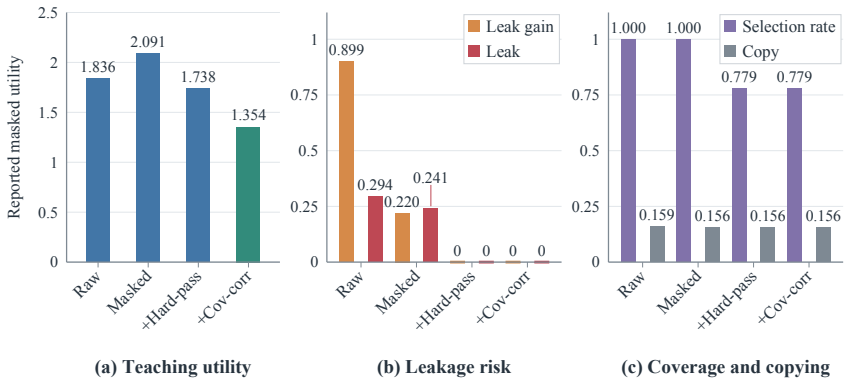



Figure 6: **Each control removes a distinct source of reward inflation: masking alone leaves 24.1% detected leakage, hard-pass filtering brings leakage and leakage gain to zero.** Controls sharing the teacher, Qwen3-1.7B student, and 1,000-event SQuAD candidate pool of Table 4. Panels report (a) reported masked utility, (b) leakage gain and detected answer-leakage rate, and (c) hint-selection rate and source-copy ratio. Raw and Masked use unmasked and answer-masked rewards, respectively; +Hard-pass restricts selection to valid hints; +Cov-corr assigns zero utility to events without valid selections. The last two settings select identical hints but use different averaging rules.

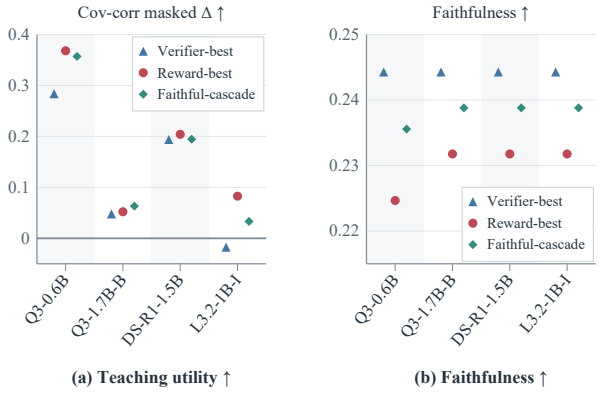



Figure 7: **Hints selected without the evaluation student still help it.** Pre-RFT SimPO teacher, 300 SQuAD events per held-out student. (a) Coverage-corrected masked utility; (b) faithfulness of valid hints. Q3: Qwen3; B: Base; DS-R1: DeepSeek-R1-Distill-Qwen; L3.2-1B-I: Llama-3.2-1B-Instruct. Full results: Table 5.

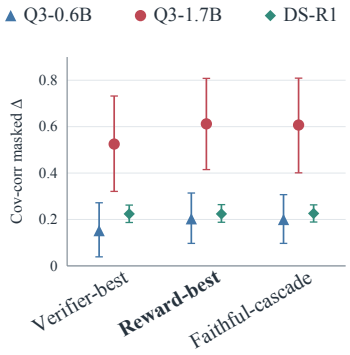



Figure 8: **All three held-out students gain positive utility.** Leave-one-student-out transfer on 1,000 SQuAD events; points are coverage-corrected masked utility, bars 95% bootstrap intervals; colors and markers identify students. Q3: Qwen3; DS-R1: DeepSeek-R1-Distill-Qwen-1.5B.

### 3.3 Effects of Masking, Filtering, and Coverage Correction

*Masking and Filtering Serve Different Roles.* Answer masking reduces leakage-related gain but does not ensure that selected hints are valid: nearly a quarter of the selected hints still contain detected leakage without hard-pass filtering. With the teacher and candidate pool fixed, hard-pass filtering excludes flagged candidates while retaining positive masked utility (Figure 6). Masking determines the information used to measure usefulness, whereas filtering determines which hints are eligible for teaching credit. The two controls therefore address different aspects of reliable feedback.

*Teaching Utility Must Account for Hint Availability.* Hard-pass filtering leaves some events without a usable hint. Averaging only over successful selections evaluates usefulness conditional on a valid hint being available, rather than improvement over the full evaluation set. Figure 6 isolates this distinction: the final two settings select identical hints but use different denominators, with valid-subset reporting overstating overall utility by 28% in this experiment. Coverage correction therefore makes the inability to produce a valid hint part of the reported teaching performance.

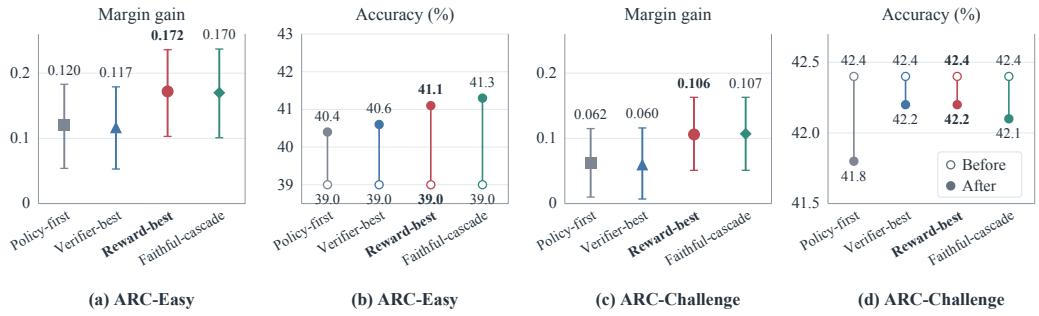



Figure 9: **Positive margin gains do not consistently translate into higher answer accuracy.** ARC-Easy and ARC-Challenge, 1,000 events each, from the original separately sampled runs. Panels (a) and (c) show coverage-corrected correct-option margin gains; bars are 95% bootstrap intervals for individual estimates. Panels (b) and (d) show accuracy before (open) and after (filled) the hint; invalid or missing hints retain the no-hint prediction. The shared-pool comparison appears in Table 4; full metrics appear in Appendix Figure 12.

### 3.4 Transfer to Held-Out Students

*Hint Utility Can Transfer beyond the Selection Student.* Using the pre-refinement SimPO teacher, we evaluate SQuAD hints selected without using the evaluation student for candidate selection (Figures 7 and 8). The larger leave-one-student-out study yields positive utility with confidence intervals excluding zero for every evaluated student. The smaller four-student study also gives positive reward-best point estimates, but the evidence is less conclusive for some models, including Llama. These results support transfer of hint usefulness beyond the selection student, while showing that the strength of this evidence varies across models and evaluation settings.

### 3.5 Training-Objective Analysis

*Additional Training Does Not Automatically Improve Teaching.* The tested LM-only continuation settings produce little teaching utility compared with hint supervision. Hint-SFT provides a strong starting point, but adding answer reconstruction reduces utility, and the reported SimPO baseline does not reliably improve on Hint-SFT. In contrast, SIMPLE's reward-filtered refinement improves utility across the evaluated sources for both backbones (Table 1). These comparisons show that additional objectives or preference optimization alone do not guarantee better teaching; the refinement configuration matters. In the recorded Qwen run, refinement improves teaching utility while reducing validation PPL from 12.98 to 12.83, with 27.4 additional training minutes (Table 2). Full training details appear in Appendix H.

### 3.6 Statistical Evidence and Evaluation Scope

*Statistical Evidence Supports the Aggregate RACE Gain.* Paired row-level bootstrap analysis of the two Qwen RACE sources supports an aggregate teaching-utility advantage for SIMPLE over both SimPO and Hint-SFT. Against Hint-SFT, the gain is $+0.227$ (95% CI $[0.147, 0.309]$); the interval against SimPO also excludes zero. The RACE-High difference alone is not statistically significant, so this conclusion applies to the aggregate result.

*No Leakage or High-Copy Cases Are Flagged in the Inspected Hints.* The recorded detector fields flag no answer leakage, template artifacts, answer-support failures, or high-copy cases among the 320 inspected hard-pass reward-best hints (Appendix J). This documents consistency with the implemented filtering criteria.

Table 2: **Recorded training cost and validation perplexity on Qwen3-1.7B.** Final SIMPLE combines Hint-SFT and 600 refinement steps. Time sums the recorded training stages, excluding candidate generation and student scoring. Full training details appear in Appendix H.

| Training objective | Total steps | Time (min) | Final val. PPL |
|---|---|---|---|
| LM-only | 1,000 | 26.7 | 12.97 |
| LM-Matched | 1,400 | 37.5 | 13.10 |
| Hint-SFT | 1,000 | 32.8 | 12.98 |
| Hint-SFT+AR | 1,000 | 38.3 | 12.98 |
| SimPO | 1,000 | 44.4 | 12.93 |
| **SIMPLE** | 1,600 | 60.2 | **12.83** |

*Likelihood Gains Do Not Consistently Translate into Accuracy Gains.* In the SciQ and ARC selection analyses, reward-best improves likelihood-based utility and correct-option margins, whereas top-1 accuracy gains are small or inconsistent (Figure 9; Appendix G). These results provide stronger evidence for improvements in likelihood-based teaching utility than in discrete answer accuracy.

## 4 Related Work

**Small-model supervision and pretraining.** MiniLLM trains compact generative students through reverse-KL distillation, adapting supervision to limited model capacity (Gu et al., 2024). Pretraining methods also incorporate human preferences (Korbak et al., 2023), instruction–response augmentation (Cheng et al., 2024), and teacher-informed data selection (Gu et al., 2025). These approaches enrich the supervision a model receives, but do not by themselves specify how to reward the teaching value of the hints it generates.

**Learning by teaching and student feedback.** Ning et al. (2024) show that student feedback, iterative teaching, and student diversity can improve language-model reasoning. Student feedback also guides training-data generation: Montessori-Instruct optimizes a teacher with DPO using synthetic examples' effects on student reference loss (Li et al., 2025), while DataEnvGym uses student errors to guide data generation (Khan et al., 2025). SOAR rewards curricula through students' post-training accuracy gains (Sundaram et al., 2026). For teacher evaluation and hint-based learning, Teach2Eval measures weaker students' gains during interaction (Zhou et al., 2026b), and HiLL jointly trains a hinter and a reasoner using hinted versus unhinted trajectory log-likelihood differences to reduce hint dependence and improve no-hint reasoning (Xia et al., 2026). Multi-turn analyses further distinguish feedback-specific gains from repeated-attempt gains (Cupiał et al., 2026). A distinct question for teacher-only continued training is how to derive supervision from the immediate effect of an individual source-grounded hint while keeping the student fixed.

**Utility-based preference learning.** DPO (Rafailov et al., 2023) and reference-free SimPO (Meng et al., 2024) optimize language models using preference pairs, while Self-Rewarding Language Models derive preferences from model judgments (Yuan et al., 2024). PUM uses lightweight students' solve-rate gains to construct preferences and train a scalar utility model for mathematical-reasoning prefixes (Zhou et al., 2026a). The preference objective itself does not specify which hints are eligible for credit or ensure that comparisons distinguish more useful valid hints from less useful ones.

**Verification and reward hacking.** Dinucu-Jianu et al. (2025) combine simulated student–tutor interactions with online reinforcement learning to balance pedagogical support and student solving accuracy. Verification can improve learning from synthetic data (Feng et al., 2025), and robust reward-model training can reduce length and format artifacts (Liu et al., 2025). Incomplete verifiers nevertheless admit shortcuts (Helff et al., 2026). Damani et al. (2026) combine verifiable rewards with an adversarial signal from human demonstrations to preserve properties missed by correctness checks. For teaching, validity checks alone do not establish usefulness, and averaging only over accepted hints can conceal failures to produce valid guidance.

These gaps motivate SIMPLE, which learns source-grounded hints for pretraining-time learning by teaching through answer-masked feedback from frozen students and preference refinement that prioritizes usefulness within the valid set, while accounting for valid-hint availability in evaluation.

## 5 Conclusion

We introduced SIMPLE, a framework that derives verifiable teaching feedback from student improvement on source-grounded teaching events. Its central insight is that hint validity, student utility, and valid-hint availability play distinct roles: grounding and anti-hacking checks determine eligibility for credit, answer-masked gains in frozen students' answer log-likelihood measure usefulness, and coverage correction assigns zero utility to events without valid candidates. Supervised hint imitation followed by reward-filtered preference refinement directs teacher learning toward more useful hints within the valid set. Across eight benchmark-derived evaluation sources and two small-model backbones, SIMPLE achieves the highest coverage-corrected teaching utility in all 16 backbone–source comparisons.

## AI USE STATEMENT

We used generative AI tools for language polishing and code assistance. The authors take full responsibility for the final content of this work.

## REPRODUCIBILITY STATEMENT

All code used to run the experiments described in the main paper and appendix is available as a part of the supplemental material.

## ETHICS STATEMENT

This research was conducted in compliance with all applicable ethical guidelines and institutional regulations. Since the study did not involve human participants, animals, or sensitive data, no specific ethical approvals were required. All data used in this research were obtained from publicly available sources, ensuring full transparency and reproducibility of the results.

## REFERENCES

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. Beat the AI: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678, 2020.

Daixuan Cheng, Yuxian Gu, Shaohan Huang, Junyu Bi, Minlie Huang, and Furu Wei. Instruction pre-training: Language models are supervised multitask learners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2529–2550, 2024.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Bartłomiej Cupiał, Jan Łojek, Mikołaj Garstecki, Szymon Pobłocki, Alicja Ziarko, and Piotr Miłoś. What drives interactive improvement from feedback? In *ICML 2026 Workshop on Reinforcement Learning from World Feedback (RLxF)*, 2026.

Mehul Damani, Isha Puri, Idan Shenfeld, and Jacob Andreas. Right in the right way: LM training with verifiable rewards and human demonstrations. *arXiv preprint arXiv:2607.01181*, 2026.

David Dinucu-Jianu, Jakub Macina, Nico Daheim, Ido Hakimi, Iryna Gurevych, and Mrinmaya Sachan. From problem-solving to teaching problem-solving: Aligning LLMs with pedagogy using reinforcement learning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 272–292, 2025.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.

Yunzhen Feng, Elvis Dohmatob, Pu Yang, François Charton, and Julia Kempe. Beyond model collapse: Scaling up with synthesized data requires verification. In *International Conference on Learning Representations*, 2025.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 10835–10866, 2023.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 2017.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024.

Yuxian Gu, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. MiniPLM: Knowledge distillation for pre-training language models. In *International Conference on Learning Representations*, 2025.

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

Lukas Helff, Quentin Delfosse, David Steinmann, Ruben Härle, Hikaru Shindo, Patrick Schramowski, Wolfgang Stammer, Kristian Kersting, and Felix Friedrich. LLMs gaming verifiers: RLVR can lead to reward hacking. *arXiv preprint arXiv:2604.15149*, 2026.

Zaid Khan, Elias Stengel-Eskin, Jaemin Cho, and Mohit Bansal. DataEnvGym: Data generation agents in teacher environments with student feedback. In *International Conference on Learning Representations*, 2025.

Tomasz Korbak, Kejian Shi, Angelica Chen, Rasika Vinayak Bhalerao, Christopher Buckley, Jason Phang, Samuel R Bowman, and Ethan Perez. Pretraining language models with human preferences. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 17506–17533, 2023.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 785–794, 2017.

Xiaochuan Li, Zichun Yu, and Chenyan Xiong. Montessori-Instruct: Generate influential training data tailored for student learning. In *International Conference on Learning Representations*, 2025.

Tianqi Liu, Wei Xiong, Jie Ren, Lichang Chen, Junru Wu, Rishabh Joshi, Yang Gao, Jiaming Shen, Zhen Qin, Tianhe Yu, Daniel Sohn, Anastasiia Makarova, Jeremiah Z. Liu, Yuan Liu, Bilal Piot, Abe Ittycheriah, Aviral Kumar, and Mohammad Saleh. RRM: Robust reward model training mitigates reward hacking. In *International Conference on Learning Representations*, 2025.

Yu Meng, Mengzhou Xia, and Danqi Chen. SimPO: Simple preference optimization with a reference-free reward. In *Advances in Neural Information Processing Systems*, 2024.

Xuefei Ning, Zifu Wang, Shiyao Li, Zinan Lin, Peiran Yao, Tianyu Fu, Matthew B. Blaschko, Guohao Dai, Huazhong Yang, and Yu Wang. Can LLMs learn by teaching for better reasoning? a preliminary study. In *Advances in Neural Information Processing Systems*, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, pp. 53728–53741, 2023.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, 2016.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 784–789, 2018.

Shobhita Sundaram, John Quan, Ariel Kwiatkowski, Kartik Ahuja, Yann Ollivier, and Julia Kempe. Teaching models to teach themselves: Reasoning at the edge of learnability. In *International Conference on Machine Learning*, 2026.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 94–106, 2017.

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

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E Weston. Self-rewarding language models. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Yuhang Zhou, Yixin Cao, and Guangnan Ye. From correctness to utility: Gain-based prefix evaluation for LLM reasoning. *arXiv preprint arXiv:2606.07190*, 2026a.

Yuhang Zhou, Xutian Chen, Yixin Cao, Yuchen Ni, Yu He, Siyu Tian, Xiang Liu, Yunwen Chen, Guangnan Ye, Xipeng Qiu, and Hongfeng Chai. Teach2Eval: An interaction-driven LLMs evaluation method via teaching effectiveness. In *International Conference on Learning Representations*, 2026b.

# Appendix

## Table of Contents


**A Full Qwen Selection Results** **14**

**B Full Task-Free Selection-Mode Results** **16**

**C Full Held-Out-Student Results** **18**

**D Method Implementation Details** **19**
- D.1 Optional Proxy-Gap Warm-Up . . . 19
- D.2 Answer-Reconstruction Ablation . . . 19
- D.3 Verifier Fields and Valid Hint Set . . . 20
- D.4 Selection Rules for Comparison . . . 20
- D.5 Masked Scoring and Rank Aggregation . . . 21

**E Additional Experimental Setups** **21**
- E.1 Data and Event Formats . . . 21
- E.2 Training Objectives and Budgets . . . 22
- E.3 Candidate Selection and Reward-Hacking Controls . . . 22
- E.4 Held-Out-Student and Leave-One-Student-Out Evaluation . . . 23
- E.5 Metrics and Statistical Analysis . . . 23

**F Proxy-Gap Pilot** **24**

**G Candidate Budget, Cross-Task, and Task-Format Evaluations** **24**

**H Language-Modeling Evaluation** **26**

**I Reward-Hacking Control Ablation** **28**

**J Rule-Based Hint Validity Analysis** **28**

**K Qualitative Examples** **30**

**L Exploratory Two-Student Refinement Ablation** **30**
- L.1 Preference-Pair Statistics . . . 31
- L.2 Same-Student Selection Results . . . 31
- L.3 Held-Out Results . . . 31

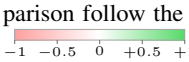
**M Shared-Candidate-Pool Sensitivity** **33**

# A Full Qwen Selection Results

The tables below give the complete Qwen results shown in Figure 3.

Table 3: Full Qwen3-1.7B selection results underlying Figure 3, with 1,000 events per source; Gain is positive in all 56 method–source cells, and SIMPLE has the highest Selected utility on every source. Both utility columns are answer-masked and coverage-corrected. Verifier reports verifier-best utility; for trained teachers, Selected reports hard-pass reward-best utility. Gain is Selected minus Verifier, and Selected cov. is valid-selection coverage. No answer leakage was detected among valid selections. Colors encode $z = (x - x_{\text{Base}})/M_j$, where the baseline is the same-source Base and $M_j$ is the largest absolute baseline difference in column $j$ across all eight sources. Green/red indicate higher/lower values, with zero in white: $-1$ $-0.5$ $0$ $+0.5$ $+1$.

| Dataset | Method | Verifier $\Delta_{\text{cov}}$ ↑ | Selected $\Delta_{\text{cov}}$ ↑ | Gain ↑ | Selected cov. ↑ |
|---|---|---|---|---|---|
| AdversarialQA | Base | -0.254 | -0.034 | 0.220 | 0.343 |
| | LM-only | -0.327 | -0.138 | 0.189 | 0.380 |
| | LM-Matched | -0.346 | -0.122 | 0.224 | 0.367 |
| | Hint-SFT | 0.511 | 1.350 | 0.839 | 0.656 |
| | Hint-SFT+AR | 0.503 | 1.320 | 0.817 | 0.660 |
| | SimPO | 0.544 | 1.291 | 0.748 | 0.649 |
| | SIMPLE | 0.574 | 1.585 | 1.011 | 0.647 |
| SQuAD | Base | -0.143 | -0.043 | 0.100 | 0.188 |
| | LM-only | -0.139 | -0.048 | 0.090 | 0.199 |
| | LM-Matched | -0.122 | -0.032 | 0.090 | 0.201 |
| | Hint-SFT | 0.694 | 1.601 | 0.907 | 0.773 |
| | Hint-SFT+AR | 0.677 | 1.582 | 0.904 | 0.768 |
| | SimPO | 0.654 | 1.614 | 0.960 | 0.768 |
| | SIMPLE | 0.887 | 1.970 | 1.083 | 0.765 |
| SQuAD-v2 | Base | -0.223 | -0.109 | 0.114 | 0.205 |
| | LM-only | -0.218 | -0.095 | 0.123 | 0.228 |
| | LM-Matched | -0.220 | -0.083 | 0.136 | 0.243 |
| | Hint-SFT | 1.257 | 2.177 | 0.920 | 0.789 |
| | Hint-SFT+AR | 1.133 | 2.033 | 0.900 | 0.786 |
| | SimPO | 1.196 | 2.027 | 0.832 | 0.785 |
| | SIMPLE | 1.274 | 2.438 | 1.164 | 0.786 |
| SciQ | Base | -0.103 | -0.042 | 0.061 | 0.132 |
| | LM-only | -0.142 | -0.088 | 0.054 | 0.114 |
| | LM-Matched | -0.119 | -0.075 | 0.045 | 0.126 |
| | Hint-SFT | 3.031 | 3.894 | 0.863 | 0.558 |
| | Hint-SFT+AR | 3.013 | 3.714 | 0.702 | 0.553 |
| | SimPO | 3.044 | 3.628 | 0.585 | 0.518 |
| | SIMPLE | 3.247 | 4.065 | 0.818 | 0.551 |

Table 3: Full Qwen3-1.7B selection results (continued). Metric definitions and the same-source Base comparison follow the preceding page. Colors use the same column-normalized scale across all eight sources: 


| Dataset | Method | Verifier $\Delta_{\text{cov}}$ ↑ | Selected $\Delta_{\text{cov}}$ ↑ | Gain ↑ | Selected cov. ↑ |
|---|---|---|---|---|---|
| RACE-Middle | Base | -0.418 | -0.148 | 0.270 | 0.377 |
| | LM-only | -0.318 | -0.140 | 0.178 | 0.345 |
| | LM-Matched | -0.300 | -0.089 | 0.210 | 0.357 |
| | Hint-SFT | -0.522 | 0.323 | 0.845 | 0.647 |
| | Hint-SFT+AR | -0.603 | 0.241 | 0.845 | 0.661 |
| | SimPO | -0.453 | 0.278 | 0.731 | 0.637 |
| | SIMPLE | -0.385 | 0.667 | 1.052 | 0.628 |
| RACE-High | Base | -0.329 | -0.126 | 0.203 | 0.338 |
| | LM-only | -0.346 | -0.157 | 0.189 | 0.308 |
| | LM-Matched | -0.373 | -0.188 | 0.186 | 0.313 |
| | Hint-SFT | -0.376 | 0.002 | 0.377 | 0.441 |
| | Hint-SFT+AR | -0.356 | -0.006 | 0.350 | 0.447 |
| | SimPO | -0.288 | 0.050 | 0.338 | 0.428 |
| | SIMPLE | -0.268 | 0.112 | 0.380 | 0.420 |
| ARC-Easy | Base | 0.094 | 0.126 | 0.032 | 0.148 |
| | LM-only | 0.057 | 0.119 | 0.062 | 0.173 |
| | LM-Matched | 0.093 | 0.148 | 0.055 | 0.173 |
| | Hint-SFT | 0.439 | 1.001 | 0.562 | 0.893 |
| | Hint-SFT+AR | 0.394 | 0.950 | 0.556 | 0.896 |
| | SimPO | 0.454 | 0.803 | 0.349 | 0.889 |
| | SIMPLE | 0.480 | 1.051 | 0.571 | 0.880 |
| ARC-Challenge | Base | 0.071 | 0.128 | 0.057 | 0.193 |
| | LM-only | 0.080 | 0.163 | 0.083 | 0.232 |
| | LM-Matched | 0.085 | 0.176 | 0.091 | 0.237 |
| | Hint-SFT | 0.361 | 0.818 | 0.457 | 0.859 |
| | Hint-SFT+AR | 0.332 | 0.788 | 0.457 | 0.858 |
| | SimPO | 0.352 | 0.663 | 0.311 | 0.858 |
| | SIMPLE | 0.403 | 0.843 | 0.440 | 0.848 |

The tables include every method, source, and reported metric evaluated using the same validity filters, including both selection modes for final SIMPLE on all eight sources. The Gemma rows in Table 1 summarize the backbone replication; its full logs and bootstrap results are in the supplementary material. Each colored appendix table defines its baseline and scale in the caption. Color intensity is not a measure of statistical significance. Tables 11 and 13 are uncolored.

## B Full Task-Free Selection-Mode Results

Figure 10 extends the four-source view in Figure 4 to all eight evaluation sources. The tables below report coverage, valid masked improvement, coverage-corrected masked improvement, faithfulness, and copying for every source and selection strategy.

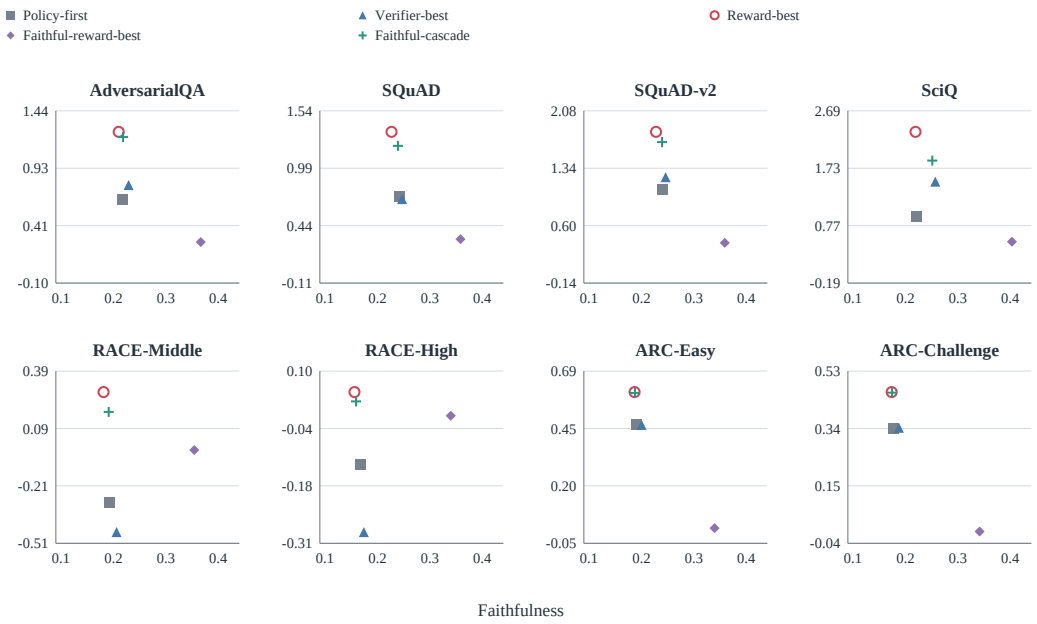



Figure 10: Complete eight-source utility–faithfulness comparison for the five selection strategies in Figure 4, using the same pre-RFT task-free SimPO teacher under same-student evaluation. The horizontal axis reports mean faithfulness of valid selected hints, and the vertical axis reports coverage-corrected masked utility, $\Delta_{\mathrm{cov}}$. Red circles identify reward-best. Faithfulness scales are shared, while utility scales vary by source. The four ranking rules share the archived reward-best pool for each event; Policy-first retains its original separately sampled hints. All methods and sources are retained, including overlapping points. Full numerical results and coverage appear in Table 4.

Table 4: Same-student selection results for the pre-RFT task-free SimPO teacher across eight sources. The four ranking rules share the archived reward-best candidate pool for each event; Policy-first is retained as a separately sampled reference. Coverage is the fraction of events with a valid selection. Valid masked Δ averages these selections only; coverage-corrected masked Δ assigns zero to events without a valid hint. Faith. and Copy are mean faithfulness and source-copy ratio among valid selections. No leakage was detected among valid selections. Colors show direction-adjusted differences from the same-source Policy-first, normalized by each column's largest absolute difference across all eight sources. Green/red indicate better/worse values, reversing Copy because lower is better; zero is white: −1 −0.5 0 +0.5 +1.

| Dataset | Selection mode | Coverage ↑ | Valid masked Δ ↑ | Cov-corr masked Δ ↑ | Faith. ↑ | Copy ↓ |
|---|---|---|---|---|---|---|
| AdversarialQA | Policy-first | 0.465 | 1.389 | 0.646 | 0.217 | 0.161 |
| | Verifier-best | 0.643 | 1.208 | 0.777 | 0.229 | 0.163 |
| | Reward-best | 0.643 | 1.947 | 1.252 | 0.210 | 0.155 |
| | Faithful-reward-best | 0.136 | 1.961 | 0.267 | 0.366 | 0.173 |
| | Faithful-cascade | 0.643 | 1.874 | 1.205 | 0.218 | 0.157 |
| SQuAD | Policy-first | 0.646 | 1.117 | 0.722 | 0.241 | 0.163 |
| | Verifier-best | 0.779 | 0.895 | 0.698 | 0.247 | 0.164 |
| | Reward-best | 0.779 | 1.722 | 1.342 | 0.227 | 0.156 |
| | Faithful-reward-best | 0.217 | 1.445 | 0.313 | 0.358 | 0.161 |
| | Faithful-cascade | 0.779 | 1.550 | 1.207 | 0.239 | 0.159 |
| SQuAD-v2 | Policy-first | 0.652 | 1.634 | 1.066 | 0.239 | 0.164 |
| | Verifier-best | 0.782 | 1.562 | 1.221 | 0.246 | 0.164 |
| | Reward-best | 0.782 | 2.310 | 1.806 | 0.228 | 0.156 |
| | Faithful-reward-best | 0.236 | 1.587 | 0.375 | 0.359 | 0.161 |
| | Faithful-cascade | 0.782 | 2.141 | 1.674 | 0.239 | 0.159 |
| SciQ | Policy-first | 0.357 | 2.583 | 0.922 | 0.221 | 0.173 |
| | Verifier-best | 0.590 | 2.556 | 1.508 | 0.257 | 0.169 |
| | Reward-best | 0.590 | 3.965 | 2.339 | 0.219 | 0.163 |
| | Faithful-reward-best | 0.221 | 2.279 | 0.504 | 0.403 | 0.177 |
| | Faithful-cascade | 0.590 | 3.151 | 1.859 | 0.251 | 0.168 |

Table 4: Same-student selection results for the pre-RFT task-free SimPO teacher (continued). Metrics and color normalization follow the preceding page. Colors compare each strategy with Policy-first on the same source; green/red indicate better/worse values, reversing Copy: −1 −0.5 0 +0.5 +1.

| Dataset | Selection mode | Coverage ↑ | Valid masked Δ ↑ | Cov-corr masked Δ ↑ | Faith. ↑ | Copy ↓ |
|---|---|---|---|---|---|---|
| RACE-Middle | Policy-first | 0.409 | -0.734 | -0.300 | 0.193 | 0.148 |
| | Verifier-best | 0.637 | -0.712 | -0.453 | 0.206 | 0.149 |
| | Reward-best | 0.637 | 0.436 | 0.278 | 0.181 | 0.141 |
| | Faithful-reward-best | 0.096 | -0.263 | -0.025 | 0.354 | 0.158 |
| | Faithful-cascade | 0.637 | 0.273 | 0.174 | 0.191 | 0.144 |
| RACE-High | Policy-first | 0.233 | -0.531 | -0.124 | 0.168 | 0.161 |
| | Verifier-best | 0.428 | -0.672 | -0.288 | 0.174 | 0.164 |
| | Reward-best | 0.428 | 0.118 | 0.050 | 0.156 | 0.152 |
| | Faithful-reward-best | 0.016 | -0.422 | -0.007 | 0.340 | 0.157 |
| | Faithful-cascade | 0.428 | 0.065 | 0.028 | 0.159 | 0.153 |
| ARC-Easy | Policy-first | 0.824 | 0.559 | 0.461 | 0.190 | 0.163 |
| | Verifier-best | 0.873 | 0.528 | 0.461 | 0.200 | 0.158 |
| | Reward-best | 0.873 | 0.689 | 0.601 | 0.187 | 0.162 |
| | Faithful-reward-best | 0.016 | 1.069 | 0.017 | 0.339 | 0.133 |
| | Faithful-cascade | 0.873 | 0.684 | 0.597 | 0.187 | 0.162 |
| ARC-Challenge | Policy-first | 0.769 | 0.444 | 0.341 | 0.177 | 0.160 |
| | Verifier-best | 0.832 | 0.412 | 0.343 | 0.187 | 0.159 |
| | Reward-best | 0.832 | 0.553 | 0.460 | 0.174 | 0.165 |
| | Faithful-reward-best | 0.010 | 0.191 | 0.002 | 0.341 | 0.142 |
| | Faithful-cascade | 0.832 | 0.551 | 0.458 | 0.174 | 0.165 |

## C Full Held-Out-Student Results

The table below gives the values behind Figure 7, including every student–selection combination and its coverage, utility, faithfulness, and valid-hint copy ratio.

Table 5: Held-out-student results for the pre-RFT SimPO teacher on 300 SQuAD events per student. Verifier-best and reward-best share identical candidate lists; faithful-cascade uses a separate candidate pool on the same evaluation events. Evaluation students are excluded from selection: Qwen3-0.6B uses hints selected by Qwen3-1.7B alone, while the other students use both Qwen3-1.7B and Qwen3-0.6B. Utility is coverage-corrected and answer-masked; faithfulness and copying average valid selections. No leakage was detected among valid selections. Colors show direction-adjusted differences from verifier-best for the same student, normalized by each column's largest absolute difference across all four students. Green/red indicate better/worse values, reversing Copy; zero is white: $-1 \quad -0.5 \quad 0 \quad +0.5 \quad +1$.

| Held-out student | Selection | Coverage ↑ | Cov-corr masked Δ ↑ | Faith. ↑ | Copy ↓ |
|---|---|---|---|---|---|
| Qwen3-0.6B | Verifier-best | 0.770 | 0.284 | **0.244** | 0.164 |
| | Reward-best | 0.770 | **0.368** | 0.225 | 0.158 |
| | Faithful-cascade | 0.777 | 0.357 | 0.236 | 0.162 |
| Qwen3-1.7B-Base | Verifier-best | 0.770 | 0.048 | **0.244** | 0.164 |
| | Reward-best | 0.770 | 0.052 | 0.232 | 0.156 |
| | Faithful-cascade | 0.777 | **0.063** | 0.239 | 0.161 |
| DeepSeek-R1-Distill-Qwen-1.5B | Verifier-best | 0.770 | 0.194 | **0.244** | 0.164 |
| | Reward-best | 0.770 | **0.204** | 0.232 | 0.156 |
| | Faithful-cascade | 0.777 | 0.194 | 0.239 | 0.161 |
| Llama-3.2-1B-Instruct | Verifier-best | 0.770 | -0.017 | **0.244** | 0.164 |
| | Reward-best | 0.770 | **0.083** | 0.232 | 0.156 |
| | Faithful-cascade | 0.777 | 0.033 | 0.239 | 0.161 |

# D Method Implementation Details

## D.1 Optional Proxy-Gap Warm-Up

Our main method uses sequence-level student-improvement feedback. Here we define the token-level proxy-gap objective used only for optional warm-up and the pilot in Appendix F. Let $\mathcal{D} = \{x^{(i)}\}_{i=1}^{N}$ be a pretraining corpus. Standard causal language modeling minimizes

$$\mathcal{L}_{\mathrm{LM}}(\theta) = \mathbb{E}_{x\sim\mathcal{D}}\left[-\sum_{t=1}^{T}\log p_\theta(x_t \mid x_{<t})\right]. \tag{9}$$

Proxy-gap pseudo-LbT introduces a weak internal model $q_{\theta,\phi}$ and measures its excess token loss relative to the full model:

$$\Delta_{\mathrm{NLL}}(c, y; \phi) = -\log q_{\theta,\phi}(y \mid c) - \mathrm{sg}\left[-\log p_\theta(y \mid c)\right]. \tag{10}$$

The corresponding gated hinge loss is

$$\mathcal{L}_{\mathrm{gap}}(\theta) = \mathbb{E}_{(c,y),\phi}\left[m(c, y)\max(0, \Delta_{\mathrm{NLL}}(c, y; \phi) - \tau)\right]. \tag{11}$$

When enabled, warm-up adds $\lambda_{\mathrm{gap}}\mathcal{L}_{\mathrm{gap}}$ to the training objective; $\lambda_{\mathrm{gap}} = 0$ in the main Hint-SFT and preference-refinement stages. The proxy reduces its NLL gap in the pilot, but not held-out LM loss under matched compute. This is why the main method measures sequence-level student improvement instead.

## D.2 Answer-Reconstruction Ablation

Answer reconstruction is evaluated separately in the Hint-SFT+AR ablation through

$$\mathcal{L}_{\mathrm{task}}(\theta) = -\mathbb{E}_{(z,e^+)\sim\mathcal{D}_{\mathrm{hint}}}\log p_\theta(a \mid q, e^+). \tag{12}$$

Table 6: Validity constraints for source-grounded teaching feedback. The four constraint categories summarize the checks used to determine hint validity, covering answer leakage, source grounding and copying, informative non-answer cues, and hint format. Verifier fields and threshold conditions are specified in Appendix D.

| Constraint category | Role in the feedback signal |
|---|---|
| No answer leakage | Prevents exact, fuzzy, and template-mediated answer disclosure from being rewarded. |
| Source-grounded but not copied | Requires support from the document while penalizing long source-span copying and question restatement. |
| Informative non-answer cue | Requires the hint to contain source cues that can guide the student without naming the answer. |
| Usable hint format | Removes unsuitable hint lengths, cloze templates, and boilerplate text that can satisfy surface checks without teaching. |

### D.3 Verifier Fields and Valid Hint Set

The main text groups validity constraints by function. Table 6 summarizes the validity checks and their roles in hint selection. The implementation records the following verifier vector:

$$\nu(e, z) = \big(V, F, S, C, A, U, T, I, L\big), \tag{13}$$

where $V$ is overall validity, $F$ source faithfulness, $S$ answer support, $C$ source-copy ratio, $A$ exact or fuzzy answer leakage, $U$ question-copy ratio, $T$ template artifacts, $I$ the informativeness of non-answer source cues, and $L$ hint length. Applying the thresholds gives

$$\begin{aligned}\mathcal{A}(z) = \{e \in \mathcal{E}(z) :& V(e, z) \geq v_{\min}, \quad F(e, z) \geq f_{\min}, \quad S(e, z) \geq s_{\min}, \\ & C(e, d) \leq c_{\max}, \quad U(e, q) \leq u_{\max}, \quad A(e, a) = 0, \quad T(e) = 0, \\ & I(e, d, a) \geq i_{\min}, \quad L_{\min} \leq L(e) \leq L_{\max}\}.\end{aligned} \tag{14}$$

The masked-utility threshold $\delta_{\text{mask}}$ belongs to the training candidate pool $\mathcal{F}_K^{\text{train}}(z)$ defined in Section 2.2, not to $\mathcal{A}(z)$. In the pair-building implementation, `min_chosen_masked_delta` is applied to the clipped-and-averaged reward-student improvement when screening both chosen targets and valid-low-utility negatives; its value is specified by the pair-construction configuration. Evaluation-time hard-pass filtering does not apply this utility threshold. The pair-construction reward in Eq. 3 is implemented as

$$\begin{aligned}R(e; z) =& U_{\text{rank}}^{\text{train}}(e; z) + \rho_v V(e, z) + \rho_f F(e, z) + \rho_s S(e, z) + \rho_i I(e, d, a) + \rho_p P(e) \\ & - \lambda_c C(e, d) - \lambda_a A(e, a) - \lambda_u U(e, q) - \lambda_t T(e) - \eta \frac{L(e)}{L_0}.\end{aligned} \tag{15}$$

Here $P(e)$ is the implementation's lexical-specificity score (`specificity`). The coefficients correspond to `rho`, `rho_faith`, `rho_support`, `rho_cue`, `rho_specificity`, `copy_penalty`, `answer_leak_penalty`, `question_copy_penalty`, `artifact_penalty`, and `eta`; $L_0$ is `length_norm`. The training ranking pool contains deduplicated generated and enabled rule-based proposals after decoder-hygiene exclusions, before hard-pass and utility-threshold filtering. The implementation stores the clipped mean as `student_improvement_masked` and the rank-normalized reward signal as `student_reward_signal`; only the latter enters the composite reward. We report these fields separately when they may explain a gain, especially leakage, copying, faithfulness, and coverage.

### D.4 Selection Rules for Comparison

Hard-pass reward-best is the primary evaluation mode. The following alternatives test which parts of the selection procedure matter:

- **Policy-first**: evaluate the first generated hint. This tests whether the policy can produce valid teaching behavior without searching over candidates.
- **Hard-pass verifier-best**: sample multiple candidates, retain valid ones, and select the highest verifier score. This tests whether validity alone is enough.

- **Hard-pass reward-best**: select the valid candidate with the highest composite selection score in Eq. (17). This is our primary utility-based rule.
- **Hard-pass faithful-reward-best**: impose a stronger faithfulness threshold before reward selection, testing the trade-off among utility, faithfulness, and coverage.
- **Faithful-cascade reward-best**: try faithful-reward selection first. If no candidate meets the stronger faithfulness threshold, use ordinary hard-pass reward-best. If none passes hard filtering, return no hint.

For all hard-pass modes, events with no valid candidate contribute zero to the coverage-corrected estimator.

### D.5 Masked Scoring and Rank Aggregation

**Raw and masked utility.** For a frozen student $s$, raw improvement is $\Delta_s^{\text{raw}}(e; z) = \log p_s(a \mid q, e) - \log p_s(a \mid q)$. Equation (1) instead scores the hint after applying the deterministic answer-masking operator $M_a$. If $M_a(e) = M_a(e')$, then $\Delta_s^{\text{mask}}(e; z) = \Delta_s^{\text{mask}}(e'; z)$ for every frozen student $s$. The clipped mean in Eq. (2) is consequently unchanged as well. This property concerns masked utility; validity checks separately determine eligibility for teaching credit.

**Rank-normalized student feedback.** For a nonempty ordered candidate pool $\mathcal{C}$ and nonempty student set $\mathcal{S}$, define

$$U_{\text{rank}}(e; z, \mathcal{C}, \mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \left(1 - \frac{r_s(e; z, \mathcal{C})}{\max(1, |\mathcal{C}| - 1)}\right), \tag{16}$$

where $r_s(e; z, \mathcal{C})$ is the zero-based descending rank of the candidate's masked gain for student $s$. Ties follow the preserved candidate order rather than receiving averaged ranks; a singleton pool receives score one.

For training, each student's gain is clipped before ranking in the scored proposal pool $\mathcal{C}_{\text{train}}(z)$, before hard-pass and utility-threshold filtering. The pool contains deduplicated generated and enabled rule-based proposals after decoder-hygiene exclusions. The resulting $U_{\text{rank}}^{\text{train}}(e; z)$ enters the composite reward in Eq. (3). It is distinct from the clipped mean $\overline{\Delta}^{\text{mask}}(e; z)$ used for candidate filtering and utility-gap checks. Evaluation instead ranks unclipped gains within the pool eligible for the selection rule; ordinary hard-pass reward-best uses $\mathcal{C} = \mathcal{F}_K(z)$ (Appendix E.3).

**Training-pair margins.** The three training thresholds have different roles: $\delta_{\text{mask}}$ screens eligible hints, $\delta_R$ screens the composite-reward gap of every retained pair, and $\delta_{\text{utility}}$ screens the masked-utility gap of valid-low-utility pairs. The Qwen preference-refinement run uses $\delta_{\text{utility}} = 0.10$. Both the chosen hint and a valid-low-utility negative satisfy $\delta_{\text{mask}}$; this threshold is not part of evaluation-time validity filtering.

## E Additional Experimental Setups

### E.1 Data and Event Formats

Both backbone experiments use SQuAD as the sole training source and evaluate on 1,000 events from each source listed in Section 3.1. The seven additional benchmarks supply evaluation events only, testing transfer beyond the SQuAD training source without dataset-specific teacher training. Events follow the source-grounded representation $z = (d, q, a, h)$ in Section 2, containing a document, a question, a reference answer, and grounding text.

For SQuAD training, event construction first uses the dataset-provided questions and reference answers, with evidence extracted from answer-containing source sentences. When a QA record fails the event checks, the pipeline attempts to construct a cloze event from its source text.

For SQuAD, SQuAD-v2, and AdversarialQA evaluation, questions and reference answers are inherited from benchmark annotations, and evidence is extracted from answer-containing context

sentences. SQuAD-v2 uses answerable examples. Exam-style and multiple-choice sources are represented as answer-text events. RACE-Middle/High events use the original article and question, with the annotated correct option text as the reference answer. Evidence is selected from article sentences using lexical overlap with the question and answer options. SciQ events use the benchmark question and annotated correct answer, with the supplied support passage serving as both the document and evidence. For RACE and SciQ, option lists are withheld from the teaching question and retained for choice-margin analysis. ARC questions retain their option lists. ARC uses a question-and-option-conditioned setting with the benchmark question and annotated correct option text. The source contains the question and all options, and the grounding text contains the question and distractor options. Faithfulness measures answer-removed lexical overlap with the source or grounding text; the answer-support flag checks overlap with the grounding text or the presence of non-answer source cues.

### E.2 Training Objectives and Budgets

*Base* denotes the original checkpoint without additional training in this study. *LM-only* applies language-modeling continuation, and *LM-Matched* extends that continuation under the original estimated-compute schedule. *Hint-SFT* trains the teacher through supervised hint imitation; *Hint-SFT+AR* additionally reconstructs the reference answer. *SimPO* is the reference-free preference-learning baseline. Final SIMPLE starts from Hint-SFT and adds 600 reward-filtered preference-refinement steps, prioritizing valid-low-utility negatives and retaining a small Hint-SFT anchor.

**Qwen loss weights.** For Qwen, the target hint-imitation weight is 1.0 during Hint-SFT. During preference refinement, the target hint-imitation and preference weights are 0.3 and 0.1, respectively. Auxiliary loss weights follow the warm-up schedule, while the LM loss has unit weight throughout. Proxy-gap and answer-reconstruction losses are disabled in both stages.

In the recorded Qwen training-budget comparison, LM-only uses 1,000 steps and LM-Matched uses 1,400 steps. The extended continuation run controls the original estimated training budget; final SIMPLE has an additional refinement stage whose cost is recorded separately. That stage processes 2.46M tokens in 27.4 minutes in the recorded run. Appendix H reports the training-cost and language-modeling results.

### E.3 Candidate Selection and Reward-Hacking Controls

The main same-student evaluations use frozen Qwen3-1.7B and Gemma-3-1B-IT students for their corresponding teacher backbones. For each backbone, the same model selects hints and evaluates their utility. Training objectives are compared using the same answer masking and validity filters, with reward-best as the primary selection rule.

**Selection-score implementation.** For rank-mean selection, each selection student ranks unclipped masked gains within the candidate pool eligible for the current rule, and the normalized ranks are averaged as in Eq. (16). With this pool denoted by $\mathcal{C}$, the composite selection score is

$$\begin{aligned} R_{\mathrm{sel}}(e; z, \mathcal{C}) = U_{\mathrm{rank}}(e; z, \mathcal{C}, \mathcal{S}_{\mathrm{sel}}) + b_v V(e, z) + b_f F(e, z) \\ + b_s S(e, z) + b_i I(e, d, a) - b_c C(e, d) - b_a A(e, a). \end{aligned} \tag{17}$$

These coefficients are the evaluation settings `select_reward_verifier_weight`, `select_reward_faithfulness_weight`, `select_reward_answer_support_weight`, `select_reward_cue_weight`, `select_reward_copy_penalty`, and `select_reward_leak_penalty`, respectively; they are distinct from the training coefficients in Eq. (15). Hard-pass reward-best uses $\mathcal{C} = \mathcal{F}_K(z)$. The faithful branch ranks only candidates passing the stronger faithfulness condition and substitutes `faithful_reward_faithfulness_weight` for $b_f$. If that pool is empty, faithful-cascade recomputes ranks over $\mathcal{F}_K(z)$ with the ordinary reward-best weights. Ties use the preserved candidate order, without averaging tied ranks; a singleton pool receives rank score one. Verifier-best uses only the verifier score and policy-first uses candidate order, so neither uses rank-mean. Reported utility remains the evaluation student's masked likelihood improvement, rather than the normalized rank score.

The selection-rule comparison holds the teacher fixed and compares five rules. *Policy-first* evaluates the first sampled hint after validity filtering. *Verifier-best* chooses the valid hint with the highest

verifier score. *Reward-best* ranks valid hints using Eq. (17). *Faithful-reward-best* imposes a stronger faithfulness threshold before reward selection. *Faithful-cascade* first applies the stronger threshold and falls back to ordinary hard-pass reward-best when no candidate satisfies it. Events without a valid candidate contribute zero under hard-pass evaluation. With the same candidate pool and hard-pass criteria, faithful-cascade and ordinary reward-best return a valid hint on exactly the same events. The detailed definitions appear in Appendix D.

The SimPO baseline in the main training-objective comparison and the pre-refinement SimPO teacher used in the selection and reward-hacking experiments are distinct checkpoints. The detailed five-rule comparison uses the pre-refinement SimPO teacher. For the four ranking rules in Table 4, we reuse the complete ordered candidate lists archived by the original reward-best run, with 1,000 matched events per source. Teacher provenance, event records, answer masking, hard-pass criteria, student checkpoints, and each rule's scoring settings are held fixed. We recompute candidate utilities with Qwen3-1.7B, BF16 precision, the original student prompts, and rank-mean selection aggregation. Policy-first retains its original separately sampled hints as a reference. Verifier-best, reward-best, and faithful-cascade therefore have identical coverage within each shared pool. Appendix M repeats the comparison using the independently archived cascade pools. The reward-hacking control ablation reuses this checkpoint, Qwen3-1.7B student, and the complete ordered SQuAD candidate pool from Table 4, with no new hint generation. The 1,000 events come from `data/cross_task/squad/val.jsonl`; each archived event retains 1–22 candidates, all of which enter the ablation. It holds the candidate pool fixed while comparing raw unmasked reward, answer-masked reward, masked reward with hard-pass filtering, and coverage-corrected reporting.

### E.4 Held-Out-Student and Leave-One-Student-Out Evaluation

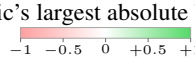
The held-out-student study uses the pre-refinement SimPO teacher and 300 SQuAD events per evaluation student. The evaluation students are Qwen3-0.6B, Qwen3-1.7B-Base, DeepSeek-R1-Distill-Qwen-1.5B, and Llama-3.2-1B-Instruct. The logged selection pool contains Qwen3-1.7B and Qwen3-0.6B, and the evaluation student is removed from that pool when present. Consequently, Qwen3-0.6B receives hints selected by Qwen3-1.7B alone, while the other three students receive hints selected by both pool members. The study compares verifier-best, reward-best, and faithful-cascade using answer masking, validity filtering, and coverage correction. Verifier-best and reward-best share identical candidate lists; faithful-cascade uses a separate candidate pool on the same evaluation events. Appendix C reports the complete results. Reward-best utility is 0.368 (95% CI $[0.172, 0.592]$) for Qwen3-0.6B and 0.204 ($[0.135, 0.275]$) for DeepSeek-R1-Distill-Qwen-1.5B; the Llama-3.2-1B-Instruct interval includes zero.

A separate leave-one-student-out experiment uses 1,000 SQuAD evaluation events and a pool of Qwen3-0.6B, Qwen3-1.7B, and DeepSeek-R1-Distill-Qwen-1.5B. Each student is evaluated using hints selected by the other two, keeping the evaluation student outside the selection ensemble. The pre-refinement teacher was trained with preference pairs scored by Qwen3-1.7B. In leave-one-student-out evaluation, this model is excluded from candidate selection when it serves as the evaluation student. Qwen3-1.7B-Base and Qwen3-1.7B denote distinct checkpoints, yielding five distinct evaluation models across the two evaluations.

### E.5 Metrics and Statistical Analysis

The primary metric is coverage-corrected answer-masked student improvement, defined in Section 2.4. Its denominator includes all evaluation events, with zero contribution from events without a valid hint. Additional metrics include selection coverage, valid-subset masked improvement, detected answer leakage, source-copy ratio, faithfulness, and verifier score. Raw answer-likelihood improvement is reported in the reward-hacking ablation.

For multiple-choice evaluations, we score the available answer options before and after the selected hint and measure the change in normalized log-likelihood margin between the correct option and the best distractor. Choice coverage accounts for recoverable option lists and valid selections. Reported 95% confidence intervals are estimated by bootstrapping row-level evaluation logs. Figure 5 reports candidate-budget results; additional analysis and task-format results appear in Appendix G. The analysis of recorded validity checks and the qualitative comparisons are reported in Appendices J and K, respectively.

Table 7: Proxy-gap and language-modeling metrics in the NLL-gap pseudo-LbT pilot. NLL-gap pLbT reduces the proxy gap, whereas longer LM-only training yields the lowest held-out loss at the same estimated cost. Est. cost denotes the reported estimated training budget. Colors show reductions relative to LM baseline, normalized by each metric's largest absolute baseline difference. Green/red indicate lower/higher loss or NLL gap, with zero in white: 

. Training-budget columns are uncolored.

| Method | Steps | Est. cost | Val loss ↓ | Held-out loss ↓ | NLL gap ↓ |
|---|---|---|---|---|---|
| LM baseline | 200 | 200 | 2.9321 | 2.8911 | 7.3883 |
| Compute-matched LM | 200 | 300 | 2.9313 | 2.8896 | 9.4669 |
| NLL-gap pLbT | 200 | 300 | 2.9295 | 2.8917 | 4.8577 |
| NLL-gap pLbT ($\lambda = 0.01$) | 200 | 300 | 2.9297 | 2.8875 | 5.4808 |
| LM baseline, same estimated FLOPs | 300 | 300 | **2.9121** | **2.8720** | 7.4339 |

## F Proxy-Gap Pilot

Table 7 reports the original NLL-gap pilot. The objective improves the proxy metric, but longer LM-only training gives better LM loss at the same estimated compute. This result led us to use sequence-level student improvement in place of token-level proxy gaps.

## G Candidate Budget, Cross-Task, and Task-Format Evaluations

The following three analyses were summarized in Section 3.

*Coverage Correction Exposes the Cost of Over-Strict Grounding.* On ARC-Easy (Appendix Table 4), faithful-reward-best raises faithfulness from 0.187 to 0.339 and valid-subset utility from 0.689 to 1.069 relative to reward-best, yet coverage falls from 87.3% to 1.6%, so coverage-corrected utility drops 35-fold, from 0.601 to 0.017. Faithful-cascade restores ordinary hard-pass coverage through fallback, while trading some utility for stronger grounding (on SciQ, faithfulness 0.251 vs. 0.219 at identical 59.0% coverage, but utility 1.859 vs. 2.339). With candidates held fixed, reward-best has the highest coverage-corrected utility on all eight sources. Its advantage over cascade is small on ARC-Easy and ARC-Challenge; Appendix M reports paired differences and intervals.

*Most of the Search Benefit Arrives by $K = 16$.* In the candidate-budget sweep (Figure 5), reward-best utility more than doubles from 0.560 at $K = 1$ to 1.188 at $K = 16$, then rises only to 1.218 at $K = 32$ and 1.267 at $K = 64$: 89% of the total gain is realized by $K = 16$, and quadrupling the budget to 64 adds just 0.079 while coverage plateaus at 0.77–0.78. Reward-best exceeds verifier-best at every $K > 1$, so the selection advantage is not tied to one budget.

*Clearer Gains in Likelihood-Based Utility Than in Top-1 Accuracy.* On choice-matched SciQ (Figure 11), reward-best increases the correct-option margin by 0.608 [0.424, 0.789], versus 0.428 for verifier-best and 0.254 for policy-first, while its top-1 accuracy change is $+0.5$ points. On ARC (Figure 12), reward-best delivers positive utility and margin gains on both sources (0.601/0.172 on ARC-Easy; 0.460/0.106 on ARC-Challenge), raises ARC-Easy accuracy from 39.0% to 41.1%, and leaves ARC-Challenge accuracy essentially unchanged (42.4% to 42.2%). The evidence is thus strong for likelihood-based utility, the quantity the objective optimizes, and modest for discrete accuracy.

More candidates help reward-based selection, but the gains taper off (Figure 5). Reward-best rises from 0.560 at $K = 1$ to 1.188 at $K = 16$, 1.218 at $K = 32$, and 1.267 at $K = 64$, while verifier-best stays flat between 0.511 and 0.560 at every budget, so the reward-best advantage widens monotonically with $K$ and reaches 0.718 at $K = 64$. Coverage levels off around 0.77–0.78. Most of the search benefit is reached by $K = 16$–$32$; $K = 64$ adds only a small gain.

Table 8 uses the same evaluation procedure for three 1,000-event extractive-QA validation sets: AdversarialQA (Bartolo et al., 2020), SQuAD (Rajpurkar et al., 2016), and SQuAD-v2 (Rajpurkar et al., 2018). Reward-best exceeds both policy-first and verifier-best on each task, with margins over verifier-best of 0.475 (AdversarialQA), 0.644 (SQuAD), and 0.585 (SQuAD-v2) at identical coverage, and no leakage detected among valid hints in any of the 12 settings. All three datasets share

Table 8: **Reward-best has the highest utility on all three extractive-QA datasets, at the same coverage as verifier-best.** Selection robustness across three extractive-QA datasets (1,000 events per dataset). Coverage reports the valid hard-pass rate, utility is coverage-corrected and answer-masked, and faithfulness averages valid selected hints. Reward-best exceeds verifier-best by 0.475–0.644 with no detected leakage among valid selections in any of the 12 settings. The ranking rules share the archived reward-best candidate pool; Policy-first is a separately sampled reference. Colors show differences from the same-dataset Policy-first, normalized by each column's largest absolute baseline difference across all three datasets. Green/red indicate higher/lower values, with zero in white: −1 −0.5 0 +0.5 +1.

| Dataset | Selection | Coverage ↑ | Cov-corr masked Δ ↑ | Faith. ↑ |
|---|---|---|---|---|
| AdversarialQA | Policy-first | 0.465 | 0.646 | 0.217 |
| | Verifier-best | 0.643 | 0.777 | **0.229** |
| | Reward-best | 0.643 | **1.252** | 0.210 |
| | Faithful-cascade | 0.643 | 1.205 | 0.218 |
| SQuAD | Policy-first | 0.646 | 0.722 | 0.241 |
| | Verifier-best | 0.779 | 0.698 | **0.247** |
| | Reward-best | 0.779 | **1.342** | 0.227 |
| | Faithful-cascade | 0.779 | 1.207 | 0.239 |
| SQuAD-v2 | Policy-first | 0.652 | 1.066 | 0.239 |
| | Verifier-best | 0.782 | 1.221 | **0.246** |
| | Reward-best | 0.782 | **1.806** | 0.228 |
| | Faithful-cascade | 0.782 | 1.674 | 0.239 |

a source-grounded answer-span format but differ in construction (standard, adversarially written, and unanswerable-augmented), so this tests robustness across extractive-QA distributions; the answer-text sources below extend the same comparison to non-extractive formats.

*Cross-task results.* Reward-best has positive coverage-corrected masked improvement on all three 1,000-event QA sets, with bootstrap 95% CIs excluding zero: 1.252 [1.032, 1.472] on AdversarialQA, 1.342 [1.107, 1.550] on SQuAD, and 1.806 [1.590, 2.043] on SQuAD-v2. Verifier-best selects from the same valid pool at identical coverage (0.643, 0.779, 0.782) yet reaches only 0.777, 0.698, and 1.221, so on SQuAD it realizes roughly half of the utility that reward-best obtains from the same candidates. Faithful-cascade has exactly the same coverage and positive utility on every task, but yields less student improvement than reward-best. The selection effect therefore extends beyond one SQuAD validation split, with the same ordering on all three sets, within the shared extractive-QA setting.

*Multiple-choice evaluation.* Table 9 tests a different answer format on 1,000 SciQ multiple-choice science QA events (Welbl et al., 2017). The answer is the correct option text, not an extracted source span. Under same-student evaluation, reward-best has the highest coverage-corrected masked improvement: 2.339 with bootstrap CI [2.023, 2.684], versus 1.508 [1.224, 1.810] for verifier-best and 0.922 [0.681, 1.176] for policy-first. The reward-best interval lies entirely above the verifier-best interval, and the point estimate is 1.55 times verifier-best and 2.5 times policy-first. No answer leakage is detected among valid hints.

Figure 11 tests the subset with recoverable option lists and valid hard-pass selections. We score every option before and after the hint, then compute the coverage-corrected change in the correct-option margin over the best distractor. Reward-best produces the largest margin gain, 0.608 [0.424, 0.789], compared with 0.428 [0.263, 0.593] for verifier-best and 0.254 [0.129, 0.399] for policy-first; all three are positive with intervals excluding zero, and the ordering policy-first $<$ verifier-best $<$ reward-best matches the utility ordering in Table 9. Correct-option likelihood gains follow the same ordering. Top-1 accuracy changes little: the clearer effect is on confidence margins, not answer accuracy. SciQ

Table 9: **Reward-best yields 1.55 times the utility of verifier-best on SciQ at the same coverage.** Teaching utility on SciQ answer-text events, with Qwen3-1.7B serving as both selection and evaluation student (1,000 events). Reference answers are correct option texts, with option lists withheld from the teaching question. All modes use answer masking, hard-pass validity checks, and coverage correction. Reward-best achieves the highest utility among the compared rules (2.339 vs. 1.508 for verifier-best and 0.922 for policy-first), with no detected leakage among valid selections. The ranking rules share the archived reward-best candidate pool; Policy-first is a separately sampled reference. Colors show differences from Policy-first, normalized by each column's largest absolute baseline difference; green/red indicate higher/lower values and zero is white: −1 −0.5 0 +0.5 +1.

| Selection | Coverage ↑ | Cov-corr masked Δ ↑ | Faith. ↑ |
|---|---|---|---|
| Policy-first | 0.357 | 0.922 | 0.221 |
| Verifier-best | 0.590 | 1.508 | **0.257** |
| Reward-best | 0.590 | **2.339** | 0.219 |
| Faithful-cascade | 0.590 | 1.859 | 0.251 |

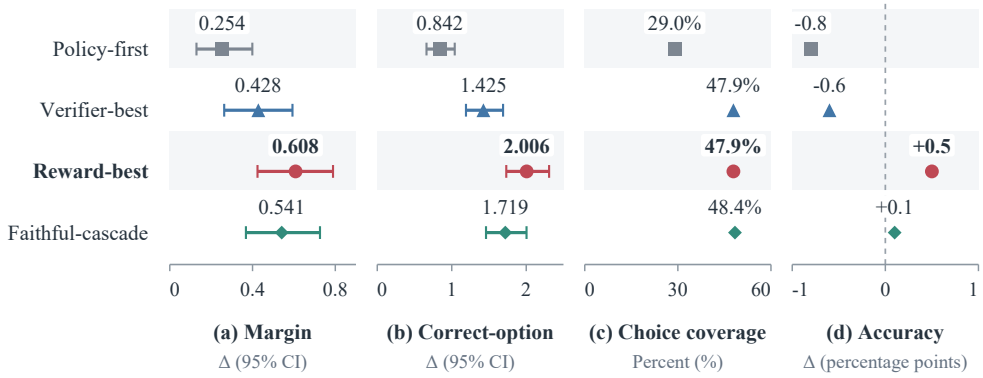



Figure 11: **Reward-best produces the largest correct-option margin gain on SciQ (0.608 vs. 0.428 for verifier-best).** Choice-matched SciQ results from the original separately sampled runs, rather than the shared-pool re-evaluation in Table 9. Panels show (a) coverage-corrected change in the normalized log-likelihood margin between the correct option and the best distractor, (b) correct-option log-likelihood gain, (c) the fraction of events with recoverable options and a valid selected hint, and (d) top-1 accuracy change in percentage points. Gains compare performance after versus before the hint. Horizontal bars in (a)–(b) indicate 95% bootstrap confidence intervals for individual estimates. Reward-best yields the largest likelihood and margin gains, while accuracy changes remain small.

thus extends the reward-selection result beyond extractive spans to an option-text answer format; together with RACE and ARC, five of the eight evaluation sources use non-extractive answers.

*ARC multiple-choice extension.* Figure 12 evaluates ARC-Easy and ARC-Challenge with validity filtering (Clark et al., 2018), which use multiple-choice science questions rather than extractive answer spans. Reward-best improves both coverage-corrected masked utility and the correct-option margin over the best distractor, with no detected leakage among valid hints. Its utility and margin gains are 0.601 and 0.172 on ARC-Easy, and 0.460 and 0.106 on ARC-Challenge, compared with utility of 0.461 and 0.343 for verifier-best on the two sources (Table 4). Top-1 accuracy rises from 39.0% to 41.1% on ARC-Easy and changes little on ARC-Challenge. As with SciQ, these results show teaching-utility and confidence-margin gains on multiple-choice science questions unseen during teacher training.

## H Language-Modeling Evaluation

Table 10 reports training costs and language-modeling metrics for the training-objective baselines.

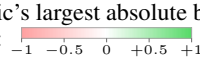



Figure 12: **Reward-best improves utility and correct-option margin on both ARC sources with zero detected leakage.** ARC multiple-choice evaluations from the original separately sampled runs, with 1,000 events from each source; the shared-pool utility comparison is in Table 4. Panels show (a) coverage-corrected masked utility, (b) coverage-corrected change in the normalized log-likelihood margin between the correct option and the best distractor, (c) valid coverage (the fraction of events with a hint that passes validity checks), and (d) top-1 accuracy before (open circles) and after (filled circles) using valid answer-masked hints. Invalid or missing hints retain the no-hint prediction in (d), for all methods. Horizontal bars in (a)–(b) indicate 95% bootstrap confidence intervals for individual estimates. Scales are shared across datasets within each metric. Coverage and accuracy are reported as percentages.

**Detailed baseline comparisons.** On Qwen, extending LM-only continuation from 1,000 to 1,400 steps changes mean utility from $-0.048$ to $-0.033$, versus $-0.031$ for Base, 1.396 for Hint-SFT, and 1.591 for final SIMPLE (Table 1). On SQuAD, Hint-SFT reaches 77.3% coverage versus 20.1% for LM-Matched. On Gemma, Hint-SFT and final SIMPLE reach mean utilities of 3.814 and 4.307, versus 0.086 for LM-Matched. Hint-SFT+AR reaches 1.328 on Qwen and 2.696 on Gemma, below Hint-SFT on all 16 backbone–source combinations; in the recorded Qwen run, it increases training time from 32.8 to 38.3 minutes with unchanged final perplexity of 12.98.

SimPO underperforms Hint-SFT on six of eight Qwen sources and all eight Gemma sources, with means of 1.294 versus 1.396 and 3.480 versus 3.814, respectively. Final SIMPLE increases mean utility over SimPO by 0.297 on Qwen and 0.827 on Gemma. The recorded Qwen refinement reduces validation perplexity from 12.98 to 12.83.

LM-Matched extends LM-only continuation according to the original estimated-compute schedule: 1,400 steps versus 1,000 for LM-only in this Qwen evaluation. The matching criterion is the estimated training budget; final SIMPLE adds the refinement-stage cost reported below. Hint-SFT+AR adds answer reconstruction to hint supervision. Across Hint-SFT and preference refinement, final SIMPLE uses 1,600 training steps, processes 6.55M LM training tokens, and takes 60.2 minutes in the recorded Qwen run. Starting from Hint-SFT, final SIMPLE adds 600 preference-refinement steps: 2.46M tokens and 27.4 minutes in the recorded run. Its final validation perplexity is 12.83, below every baseline in Table 10 (12.93–13.10) and 0.15 below its Hint-SFT starting point, so the refinement stage that yields the utility gains in Table 1 does not trade away language-modeling quality.

Table 10: **The five baseline objectives yield final validation perplexities of 12.93–13.10.** Training cost and language-model preservation for the Qwen training-objective baselines. Columns report training steps, processed tokens, training time, best validation loss, and final validation perplexity (PPL). Token counts refer to the language-modeling batches processed during training. The additional refinement stage of SIMPLE, which reaches 12.83 PPL, is reported separately in Appendix H. Colors show reductions relative to LM-only, normalized by each metric's largest absolute baseline difference. Green/red indicate lower/higher loss or perplexity, with zero in white: 

. Training-cost columns are uncolored.

| Training objective | Steps | LM training tokens (M) | Time (min) | Best val loss ↓ | Final val PPL ↓ |
|---|---|---|---|---|---|
| LM-only | 1000 | 4.10 | 26.7 | **2.552** | 12.97 |
| LM-Matched | 1400 | 5.73 | 37.5 | **2.552** | 13.10 |
| Hint-SFT | 1000 | 4.10 | 32.8 | 2.563 | 12.98 |
| Hint-SFT+AR | 1000 | 4.10 | 38.3 | 2.563 | 12.98 |
| SimPO | 1000 | 4.10 | 44.4 | 2.559 | **12.93** |

# I Reward-Hacking Control Ablation

Figure 6 compares the reward-hacking controls described here. All variants reuse the pre-RFT SimPO teacher checkpoint, Qwen3-1.7B evaluation student, and exact ordered candidate pool from the 1,000 SQuAD events in Table 4. We compare four settings: raw unmasked student improvement, answer-masked student improvement, answer-masked improvement after hard-pass filtering, and the full coverage-corrected metric. In the hard-pass settings, valid candidates are ranked by masked student improvement; the reward-best selection experiments use the composite score $R_{\text{sel}}$. Raw and Masked report means over all selected hints; hard-pass settings additionally report coverage-corrected means with zero contribution from events without valid selections. The comparison explains our use of answer-masked, coverage-corrected metrics rather than raw student improvement.

Raw reward yields high apparent improvement, but also a leakage gain of 0.899 and a 29.4% answer-leakage rate. Masking raises masked utility from 1.836 to 2.091 and cuts leakage gain to 0.220, yet 24.1% of the selected hints still leak when hard-pass constraints are absent. Hard-pass filtering reduces both detected leakage and leakage gain to zero while leaving the source-copy ratio essentially unchanged (0.156–0.159 across all four settings), so it removes leakage specifically rather than shifting the selection toward copying. The trade-off is coverage: 77.9% of events have a valid candidate in this evaluation. Counting the remaining events as zero lowers the reported value from the valid-subset mean of 1.738 to the coverage-corrected 1.354. Each control therefore addresses a different way the reported reward can be inflated, and the four settings together move the estimate from an unfiltered masked utility of 1.836 to a coverage-corrected masked utility of 1.354, with zero detected leakage under the implemented checks.

# J Rule-Based Hint Validity Analysis

We inspect 40 reward-best hints from each of the eight benchmark-derived evaluations, for 320 hints in total. The original samples come from 500-event AdversarialQA, SQuAD, and SQuAD-v2 evaluation runs. We apply the same procedure to 1,000-event SciQ, RACE-Middle, RACE-High, ARC-Easy, and ARC-Challenge runs. In each archived log, hard-pass selections retain their original order; we sample positions $\text{round}(i(M-1)/39)$ for $i = 0, \ldots, 39$, with $M$ the number of hard-pass selections and positions indexed from zero. This spreads the sample across the log but is neither random nor blinded. We aggregate rule-based detector fields and retain file hashes and sample indices for reproducibility. As shown in Table 11, every sampled hint has zero logged leakage and template flags, passes answer-support checks, and falls below the high-copy threshold.

Table 11: **Zero leakage, template, answer-support, or high-copy flags across all 320 inspected hints.** Rule-based validity checks for 320 hard-pass reward-best hints, with 40 from each evaluation source. Template and Unsupp. report rates of flagged template artifacts and failed answer-support checks. High copy and Low cue denote fractions with source-copy ratio at least 0.5 and fewer than three source cues, respectively. Copy, Faith., and Spec. are mean logged copy, faithfulness, and specificity scores. Overall pools all 320 hints. No answer leakage was flagged in the inspected sample.

| Dataset | $N$ | Template | Unsupp. | High copy | Low cue | Copy | Faith. | Spec. |
|---|---|---|---|---|---|---|---|---|
| AdversarialQA | 40 | 0.000 | 0.000 | 0.000 | 0.075 | 0.166 | 0.236 | 0.931 |
| SQuAD | 40 | 0.000 | 0.000 | 0.000 | 0.075 | 0.174 | 0.239 | 0.887 |
| SQuAD-v2 | 40 | 0.000 | 0.000 | 0.000 | 0.100 | 0.200 | 0.278 | 0.925 |
| SciQ | 40 | 0.000 | 0.000 | 0.000 | 0.200 | 0.170 | 0.239 | 0.916 |
| RACE-Middle | 40 | 0.000 | 0.000 | 0.000 | 0.050 | 0.156 | 0.197 | 0.997 |
| RACE-High | 40 | 0.000 | 0.000 | 0.000 | 0.000 | 0.150 | 0.162 | 0.997 |
| ARC-Easy | 40 | 0.000 | 0.000 | 0.000 | 0.000 | 0.167 | 0.185 | 1.000 |
| ARC-Challenge | 40 | 0.000 | 0.000 | 0.000 | 0.000 | 0.155 | 0.187 | 1.000 |
| Overall | 320 | 0.000 | 0.000 | 0.000 | 0.062 | 0.167 | 0.216 | 0.957 |

## K Qualitative Examples

Table 12: Examples of hints selected by reward-best, verifier-best, and faithful-cascade for the same teaching events. Each group gives the question, reference answer, and selected hints, illustrating differences in source cues and phrasing. Faithful-cascade and reward-best select identical hints in three of the four examples.

| | |
|---|---|
| **Example 1.** How many primary affiliates did ABC have in 1954?<br>*Reference answer:* 14 | |
| *Reward-best* | Focus on the document clues however, year, abc, primary; they identify the missing entity without naming it. |
| *Verifier-best* | Use these source-grounded clues rather than the answer itself: however, year, abc, only, primary. |
| *Faithful-cascade* | Focus on the document clues however, year, abc, primary; they identify the missing entity without naming it. |
| **Example 2.** For which ABC Movie of the Week film did Steven Spielberg first gain success?<br>*Reference answer:* Duel | |
| *Reward-best* | Focus on the document clues movie, week, broadcast, feature; they identify the missing entity without naming it. |
| *Verifier-best* | Use these source-grounded clues rather than the answer itself: movie, week, broadcast, feature, length. |
| *Faithful-cascade* | Focus on the document clues movie, week, broadcast, feature; they identify the missing entity without naming it. |
| **Example 3.** Tesla incorrectly believed that [MASK] were longitudinal waves, such as those produced in waves in plasmas.<br>*Reference answer:* X-rays | |
| *Reward-best* | Use these source-grounded clues rather than the answer itself: incorrectly, believed, waves, produced, plasma. |
| *Verifier-best* | Use these source-grounded clues rather than the answer itself: incorrectly, believed, longitudinal, waves, plasma. |
| *Faithful-cascade* | Use these source-grounded clues rather than the answer itself: incorrectly, believed, longitudinal, waves, produced. |
| **Example 4.** When was a patent-sharing agreement signed between Westinghouse Electric and General Electric?<br>*Reference answer:* 1896 | |
| *Reward-best* | Focus on the document clues although, patent, sharing, agreement; they identify the missing entity without naming it. |
| *Verifier-best* | Use these source-grounded clues rather than the answer itself: although, patent, sharing, agreement, signed. |
| *Faithful-cascade* | Focus on the document clues although, patent, sharing, agreement; they identify the missing entity without naming it. |

## L Exploratory Two-Student Refinement Ablation

This exploratory two-student refinement ablation tests cross-student transfer by adding a second reward student and a separate preference-refinement stage. The run produces many high-quality chosen hints and restores reward-best $>$ verifier-best in same-student evaluation, but does not clearly outperform the earlier single-student SimPO evaluation.

### L.1 Preference-Pair Statistics

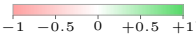
Table 13 summarizes the two-student preference pairs. Chosen hints are safe and useful, but many rejected hints are unsafe: their leakage rate is high and hard-pass rate is low. The loss therefore still learns many "safe > unsafe" comparisons rather than "high-utility valid hint > low-utility valid hint" comparisons. Adding another reward student does not, by itself, turn these into better ranking examples.

Table 13: Preference-pair statistics for the exploratory two-student refinement ablation. The table summarizes 11,385 chosen–rejected pairs using masked improvement, the logged reward signal, hard-pass rate, detected leakage rate, and source-copy ratio. All chosen hints pass the recorded hard-pass checks, while rejected hints include both valid and constraint-violating alternatives. This ablation is separate from final SIMPLE.

| Pair side | Count / rate | Masked Δ | Reward signal | Hard-pass | Leak | Copy |
|---|---|---|---|---|---|---|
| Chosen | 11,385 | 3.542 | 0.588 | 1.000 | 0.000 | 0.157 |
| Rejected | 11,385 | 1.859 | 0.443 | 0.262 | 0.679 | 0.276 |

### L.2 Same-Student Selection Results

The two-student run restores a reward-best advantage over verifier-best on its original 200-event evaluation (Table 14). In Tables 14 and 15, verifier-best and reward-best share identical candidate lists, while faithful-cascade uses a separate pool on the same 200 events. The main results in Table 1 use separate 1,000-event sets. These runs are not matched comparisons; the former is an ablation of multi-student refinement.

Table 14: Same-student selection results for the exploratory two-student refinement checkpoint on its original 200-event evaluation set. Verifier-best and reward-best share identical candidate lists; faithful-cascade uses a separate candidate pool on the same evaluation events. Valid masked improvement averages valid selections; coverage-corrected masked improvement assigns zero to events without a valid hint. Reward-best achieves the highest coverage-corrected utility, with no detected leakage among valid selections. Colors show direction-adjusted differences from Policy-first, normalized by each column's largest absolute baseline difference. Green/red indicate better/worse values, reversing Copy because lower is better; zero is white: 

.

| Selection mode | Coverage ↑ | Valid masked Δ ↑ | Cov-corr masked Δ ↑ | Faith. ↑ | Copy ↓ |
|---|---|---|---|---|---|
| Policy-first | 0.630 | 0.944 | 0.595 | 0.225 | 0.149 |
| Verifier-best | 0.745 | 1.122 | 0.836 | 0.236 | 0.157 |
| Reward-best | 0.745 | **1.826** | **1.361** | 0.225 | 0.151 |
| Faithful-reward-best | 0.180 | 2.280 | 0.410 | **0.353** | **0.146** |
| Faithful-cascade | 0.760 | 1.561 | 1.186 | 0.232 | 0.151 |

### L.3 Held-Out Results

Table 15 reports the two-student run on diverse held-out models. It improves on an earlier naive multi-student run for Qwen3-0.6B (Yang et al., 2025), DeepSeek-R1-Distill-Qwen-1.5B (Guo et al., 2025), and Llama-3.2-1B-Instruct (Grattafiori et al., 2024).

Table 15: Held-out-student results for the exploratory two-student refinement checkpoint. Three selection strategies are compared on four evaluation students using coverage, coverage-corrected masked utility, and valid-hint faithfulness. Verifier-best and reward-best share identical candidate lists; faithful-cascade uses a separate candidate pool on the same evaluation events. No answer leakage was detected among valid selections. Colors show differences from verifier-best for the same student, normalized by each column's largest absolute baseline difference across all four students. Green/red indicate higher/lower values, with zero in white: 


| Held-out student | Selection | Coverage ↑ | Cov-corr masked Δ ↑ | Faith. ↑ |
|---|---|---|---|---|
| Qwen3-0.6B | Verifier-best | 0.745 | 0.355 | 0.236 |
| | Reward-best | 0.745 | **0.412** | 0.219 |
| | Faithful-cascade | 0.760 | 0.368 | 0.228 |
| Qwen3-1.7B-Base | Verifier-best | 0.745 | 0.059 | 0.236 |
| | Reward-best | 0.745 | 0.048 | 0.225 |
| | Faithful-cascade | 0.760 | **0.059** | 0.232 |
| DeepSeek-R1-Distill-Qwen-1.5B | Verifier-best | 0.745 | 0.208 | 0.236 |
| | Reward-best | 0.745 | 0.227 | 0.225 |
| | Faithful-cascade | 0.760 | **0.239** | 0.232 |
| Llama-3.2-1B-Instruct | Verifier-best | 0.745 | 0.018 | 0.236 |
| | Reward-best | 0.745 | **0.148** | 0.225 |
| | Faithful-cascade | 0.760 | 0.094 | 0.232 |

*Implications for pair construction.* Future multi-student runs should use rejected hints that are mostly valid but less useful. Otherwise, SimPO spends much of its training on safety comparisons rather than learning to rank useful hints within the valid set.

## M SHARED-CANDIDATE-POOL SENSITIVITY

The primary selection-rule comparison reuses the archived reward-best candidate lists. We repeat all four ranking rules on the candidate lists archived by faithful-cascade, retaining the same 1,000 events per source and all scoring settings. The two pools differ on six sources; the RACE-Middle and RACE-High pools are identical, so their repeated results are not independent replications. No new hints are generated. All 32 original source–rule logs pass the hard-pass-count and selected-verifier-field checks. Replaying reward-best and cascade on their own original pools exactly reproduces all selected hints, validity states, and selected-hint utilities.

Within either pool, verifier-best, reward-best, and cascade have identical event-level coverage. Cascade falls back to reward-best when the strict faithful subset is empty and matches faithful-reward-best otherwise. Reward-best has the highest coverage-corrected utility on every source in both pools. Table 16 gives the alternative-pool results; Table 17 reports paired reward-best-minus-cascade differences. Intervals use 2,000 event-level paired bootstrap resamples, seed 7, and percentile 95% bounds, without multiplicity correction.

Table 16: Alternative archived cascade pools. $C$ is shared by verifier-best (V), reward-best (R), and cascade (Casc.); $C_f$ is strict faithful-reward-best (F) coverage. Remaining columns report coverage-corrected masked utility.

| Source | $C$ | $C_f$ | V | R | F | Casc. |
|---|---|---|---|---|---|---|
| AdversarialQA | 0.646 | 0.135 | 0.725 | 1.257 | 0.266 | 1.200 |
| SQuAD | 0.783 | 0.216 | 0.663 | 1.353 | 0.311 | 1.215 |
| SQuAD-v2 | 0.784 | 0.237 | 1.215 | 1.824 | 0.371 | 1.684 |
| SciQ | 0.597 | 0.238 | 1.493 | 2.389 | 0.547 | 1.896 |
| RACE-Middle | 0.637 | 0.096 | -0.453 | 0.278 | -0.025 | 0.174 |
| RACE-High | 0.428 | 0.016 | -0.288 | 0.050 | -0.007 | 0.028 |
| ARC-Easy | 0.873 | 0.017 | 0.461 | 0.605 | 0.014 | 0.601 |
| ARC-Challenge | 0.834 | 0.010 | 0.343 | 0.460 | 0.001 | 0.459 |

Table 17: Paired reward-best minus cascade utility: primary reward-best pool (A) and alternative cascade pool (B). Brackets give unadjusted 95% bootstrap intervals.

| Source | A: difference [95% CI] | B: difference [95% CI] |
|---|---|---|
| AdversarialQA | 0.04668 [0.02550, 0.07216] | 0.05697 [0.03257, 0.08537] |
| SQuAD | 0.13423 [0.09495, 0.17562] | 0.13731 [0.09850, 0.17824] |
| SQuAD-v2 | 0.13199 [0.09009, 0.18206] | 0.13990 [0.09596, 0.19152] |
| SciQ | 0.48023 [0.36094, 0.61543] | 0.49311 [0.36282, 0.63457] |
| RACE-Middle | 0.10349 [0.06829, 0.14415] | 0.10349 [0.06829, 0.14415] |
| RACE-High | 0.02253 [0.00776, 0.04112] | 0.02253 [0.00776, 0.04112] |
| ARC-Easy | 0.00389 [0.00042, 0.00908] | 0.00418 [0.00069, 0.00923] |
| ARC-Challenge | 0.00188 [0.00016, 0.00420] | 0.00117 [0.00012, 0.00273] |