# EDIT: Evidence-Diagnosed Intervention Training for Rule-Faithful LLM Grading

**Zhihao Wu**[1*] **Linhai Zhang**[1*] **Taiyi Wang**[2*] **Runcong Zhao**[1]
**Peter Andrews**[3] **Cesare Aloisi**[3] **Yulan He**[1,4†]

[1]King's College London [2]University of Cambridge [3]AQA [4]The Alan Turing Institute

{zhihao.2.wu, linhai.zhang, runcong.zhao, yulan.he}@kcl.ac.uk
taiyi.wang@cl.cam.ac.uk, {peter.andrews, caloisi}@aqa.org.uk

## Abstract

Reliable rubric grading requires more than accurate score prediction. Each judgement must be grounded in the mark scheme and evidence from the student answer. Existing credit-assignment and intervention methods, primarily designed for self-contained reasoning tasks such as mathematics reasoning, struggle in this setting because they do not identify where grading reasoning goes wrong or how the model's belief about the final mark changes during reasoning. We propose Evidence-Diagnosed Intervention Training (EDIT), a two-phase framework for training more rubric-faithful LLM graders. First, EDIT-SFT locates problematic reasoning steps using internal model signals: posterior belief over the final mark and input-grounding scores. It then revises only these local steps with help from a rubric checklist. Second, EDIT-RL calibrates the grader with belief-guided reward shaping, penalising large harmful belief drifts while still allowing helpful exploration. Experiments on two real-world, multi-subject grading benchmarks demonstrate that EDIT consistently outperforms strong supervised fine-tuning and reinforcement learning baselines on both in-domain and out-of-domain splits, with ablation studies confirming that internal-state diagnostics drive these gains. Under deterministic rubric-edit interventions, EDIT-SFT is the most rule-responsive of all evaluated systems, and EDIT-RL largely retains this responsiveness while improving accuracy.

## 1 Introduction

Grading a student's exam response is a rule application task. A grader needs to read the student answer, compare it with an explicit mark scheme, and assign a holistic mark. Large language models (LLMs) are increasingly deployed as automatic graders (Grévisse, 2024; Li et al., 2025; Lai et al., 2025; Wang et al., 2026). However, training them to grade reliably remains difficult. The final mark depends on a sequence of intermediate judgements, such as which rubric criteria are satisfied, what evidence supports them, which mark band applies, and how partial credits compose.

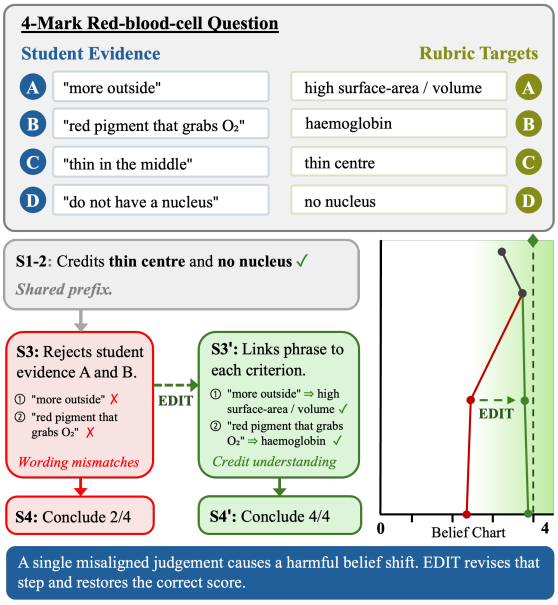


Figure 1: An illustration of a 4-mark grading question. The shared prefix (S1–2) correctly credits two criteria, but the standard grader then rejects two correct phrases due to wording mismatches (S3), shifting its belief away from the gold mark and concluding 2/4. EDIT locally revises only that step (S3′), linking each disputed phrase to its rubric criterion, which pulls the belief trajectory back to the gold mark and restores the correct mark.

This makes rubric grading a demanding credit-assignment problem. Outcome-based reinforcement learning (RL) methods such as GRPO (Shao et al., 2024) assign a single final reward to the whole reasoning trajectory (Pignatelli et al., 2024), and they cannot identify which reasoning step caused an incorrect mark. More fine-grained credit assignment methods, such as token-reallocation methods (Jin et al., 2026) or intervention-style training (Yang et al., 2026), offer a possible remedy.

[*] Equal contribution.
[†] Corresponding author.



*Findings of the Association for Computational Linguistics: EMNLP 2026*, pages 25397–25413
October 24-29, 2026 ©2026 Association for Computational Linguistics

However, these methods were primarily designed for *self-contained* reasoning tasks like mathematics, where the answer can often be checked by a verifier and the error is usually local. They are less suited to *rule-application reasoning* tasks, where judging an answer requires checking whether each step correctly invokes and applies an external rule, rubric, or policy. Rubric grading has two additional requirements. First, it requires **external grounding**. Rather than just maintaining internal logical consistency, each judgement step must be anchored in both the mark scheme and the student's written evidence. Second, rule-based reasoning is an **uncertainty reduction** process. The model's belief about the final mark should progressively converge. Existing methods do not directly control this decision trajectory. Outcome-only RL ignores harmful belief shifts that occur in the middle of reasoning process, while token-level constraints can over-penalise benign exploratory phrasing.

To address these challenges, we introduce EDIT (Evidence-Diagnosed Intervention Training), a novel two-phase framework for training rubric-faithful LLM graders. As shown in Figure 1, rather than relying on prompt-based self-audits or rigid token-level rewards, EDIT uses the model's internal decision signals to locate grading failures, revise them locally, and calibrate how the model's belief over the final mark evolves during reasoning. Both phases read the same internal quantity, the model's posterior belief over the final mark. The first phase uses harmful transitions of this posterior to locate the step to repair, and the second constrains its trajectory during RL, so the supervision and the reward act in one shared coordinate system (§3.2).

To enforce strict external grounding, the first phase, EDIT-SFT, replaces prompt-based self-audits with **internal-state localisation**. At each reasoning step, it inspects the model's posterior belief over the final mark and computes mask-based grounding signals that measure whether the step depends on the student answer, the rubric, and the preceding reasoning steps. These signals identify where the reasoning first moves away from the correct mark or becomes weakly grounded. The selected step is then revised as a small atomic edit under a locality constraint. A rubric checklist is provided as privileged context, helping the reviser use the mark scheme without forcing the output to follow a rigid checklist format.

To manage the uncertainty reduction process, the second phase, EDIT-RL, introduces **belief-guided reward shaping** for calibration RL. Instead of reallocating sequence-level reward across tokens, EDIT-RL augments the standard mark-distance outcome reward with a penalty for large mid-trajectory belief drifts. Small fluctuations are allowed as benign exploration, but belief drifts that move too far from the gold mark are penalised in proportion to their excess.

Experimental results on two real-world, multi-subject student response grading benchmarks demonstrate that EDIT significantly outperforms strong SFT and RL baselines on both in-domain and out-of-domain splits. Furthermore, ablation studies confirm that our internal-state diagnostics are the primary source of these performance gains.

In conclusion, our contributions are three-fold:

- We introduce EDIT, a novel two-phase **training framework for rubric-faithful grading**, that adapts credit assignment to externally grounded rubric grading rather than closed-system reasoning.
- We replace unreliable prompt-based self-audits with robust posterior and grounding signals (**Internal State Diagnostics**), and use rubric-checklist context to support local atomic revisions without imposing a rigid output template.
- We propose a calibration RL mechanism, called **Belief-Guided Reward Shaping**, that penalises severe mid-trajectory belief drifts while tolerating benign exploration, improving grading accuracy and out-of-domain generalisation.

## 2 Methodology

EDIT (Evidence-Diagnosed Intervention Training) fine-tunes an LLM grader in two phases (Fig. 2). Phase 1 (EDIT-SFT, §2.2) builds a supervised training set by revising small, local parts of incorrect reasoning steps. The *atomic* steps to revise are chosen using the model's own internal belief signals and grounding signals. The revised content is guided by a *rubric-checklist*, which is extracted from the mark scheme and the student answer. Phase 2 (EDIT-RL, §2.3) then calibrates the resulting policy with group-relative outcome-RL. The reward is shaped by the posterior belief signal used in Phase 1, giving a marking-task-aligned formulation that we call *Belief-guided Reward Shaping*.

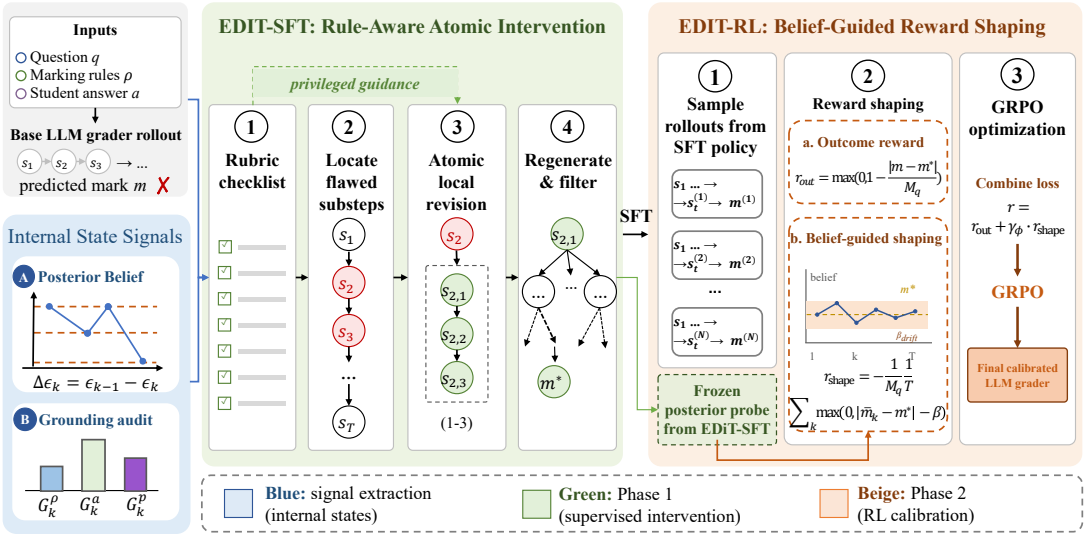


Figure 2: The EDIT pipeline. **Phase 1 (EDIT-SFT):** For incorrect rollouts, internal-state signals, a posterior belief probe and masked-support grounding audits, locate and rank flawed substeps. A per-response rubric checklist is then provided to the reviser as privileged input (dashed arrow) to produce atomic corrective edits under a locality constraint, without imposing a rigid output template. These corrected trajectories are aggregated for supervised fine-tuning (SFT). **Phase 2 (EDIT-RL):** The SFT policy is further calibrated using GRPO, augmented with a threshold-respecting, belief-guided reward (Eq. 6). This shaping signal is computed from a frozen posterior probe to penalise mid-trajectory excursions.

## 2.1 Internal State Signals

EDIT first extracts two step-level internal-state signals from the grading model (i.e., policy), both computed offline from existing rollouts. Consider a rollout be generated for the question $q$, marking rules $\rho$, and a student answer $a$. The rollout contains $T$ reasoning steps $s_{1:T}$. For each step boundary, we compute two signals.

**Posterior belief probe.** At step boundary $k \in \{0, \ldots, T\}$, we estimate the policy's belief about the final integer mark, $m \in \{0, \ldots, M_q\}$, where $M_q$ is the maximum available mark for question $q$. To do this, we append the final marking *scaffold* to the partial reasoning chain and read the next-token logits over the valid mark tokens:

$$\hat{p}_k(m) \propto \exp(\text{logit}_\theta(m|q, \rho, a, s_{1:k}, \textit{scaff.})) \quad (1)$$

From $\hat{p}_k$, we derive the expected mark $\hat{m}_k = \mathbb{E}_{m\sim\hat{p}_k}[m]$ and its absolute error $\epsilon_k = |\hat{m}_k - m^*|$, where $m^*$ is the gold mark. We then define the signed per-step change as:

$$\Delta\epsilon_k = \epsilon_{k-1} - \epsilon_k \quad (2)$$

A positive value means that step $s_k$ moved the policy's belief *closer* to the gold mark, while a negative value means that the step moved the belief *farther away*. Thus, the sign shows whether the reasoning step locally helped or hurt the model's marking belief.

**Mask-based grounding audit.** For each reasoning step $s_k$, we compute mask-based support scores that measure how much the likelihood of $s_k$ depends on three information sources: the marking rule $\rho$, the student answer $a$, and the reasoning prefix $s_{<k} = (s_1, \ldots, s_{k-1})$. Let $\ell_k(q, \rho, a, s_{<k}) = -\log p_\theta(s_k \mid q, \rho, a, s_{<k})$ denote the negative log-likelihood of step $s_k$ under the full context. We write

$$\ell_k^{\text{full}} = \ell_k(q, \rho, a, s_{<k}).$$

The support scores are then defined as

$$\begin{aligned} G_k^\rho &= \ell_k(q, \texttt{[MASK]}, a, s_{<k}) - \ell_k^{\text{full}}, \\ G_k^a &= \ell_k(q, \rho, \texttt{[MASK]}, s_{<k}) - \ell_k^{\text{full}}, \\ G_k^p &= \ell_k(q, \rho, a, \varnothing) - \ell_k^{\text{full}}. \end{aligned} \quad (3)$$

Larger values indicate that $s_k$ is more strongly supported by the corresponding masked input source. For example, if masking the student answer makes the step much harder to predict, then the step is likely grounded in the student answer. Because $G_k^p$ naturally increases for later steps (later steps tend to depend more on the earlier reasoning prefix), we use a position-residualised $\widetilde{G}_k^p$ as the reasoning-prefix-support score. Let $u_{i,k} = k/T_i$

denote the relative position of step $k$ in example $i$, and let $\mu_p(u) = \mathbb{E}_{(i,j)\sim\mathcal{D}}\left[G^p_{i,j} \,\middle|\, u_{i,j} = u\right]$ be the dataset-level mean prefix-support score at relative position $u$. We then define

$$\widetilde{G}^p_{i,k} = G^p_{i,k} - \mu_p(u_{i,k}). \tag{4}$$

This audit is an *input-ablation* probe. We replace one input section with a placeholder and measure how much harder it becomes to predict the already-generated step. We validate this audit against external per-step labels produced by a stronger annotator model, as reported in Appendix E. The grounding-audit scores correlate with these labels in the expected direction, and Appendix F further validates the causal utility of the selected steps.

### 2.2 Phase 1: EDIT-SFT

EDIT-SFT builds a supervised training pool by revising small parts of incorrect rollouts. Each rollout passes through four sub-stages: rubric-checklist generation, candidate step selection using internal signals, atomic revision, and outcome filtering. Successful revisions are then mixed with naturally correct rollouts and used as SFT targets.

**Rubric-checklist.** For each sample $x_i = \{q_i, \rho_i, a_i\}$, we generate a rubric-checklist by prompting the policy with the gold mark $m^*_i$ as privileged information (available only during training-pool construction, not at inference). The policy is asked to list every marking point verbatim and to judge whether each point is covered, supported by a quoted evidence span. This checklist is later used as context during intervention.

**Candidate steps selection.** In this phase, we locate steps in an incorrect rollout ($m \neq m^*$) that should be edited, using the internal-state signals of §2.1. Steps are first *shortlisted by the belief-shift criterion alone*: we rank steps by the most-negative $\Delta\epsilon_k$, discard positive or near-tie values, and keep the top-$k$ that most strongly moved the belief away from the gold mark. A step with a large negative belief shift but strong grounding, where the evidence was available yet the judgement still went wrong, is thus fully eligible and is our primary target. The grounding audit then *re-ranks within this shortlist*, promoting steps whose $G^\rho_k$, $G^a_k$, and $\widetilde{G}^p_k$ all fall below the dataset-wide 25th percentile ahead of the others before the final candidate selection. The audit never excludes a step. Algorithm 1 (Appendix C) gives the full procedure, Appendix F validates this rule against random and belief-shift-only alternatives, and implementation details are in Appendix B.

**Atomic revision under locality constraint.** For each selected candidate step $s_k$, we ask the policy to revise it into 1–3 new sub-steps that fix only the local reasoning error. The prompt includes the sample context $\{q, \rho, a\}$, the full reasoning attempt with $s_k$ highlighted, the internal-signal diagnostic, and the previously generated rubric checklist.

The revision must satisfy a locality constraint, ensuring that the intervention does not destroy the reasoning structure or coherence of the original rollouts. The constraint also prevents the revised step from simply copying dense information or oracle knowledge from the gold mark or rubric checklist. The details for the prompt and checklist-mimicry detector are shown in Appendix J.

**Pool composition and training.** For each revision, we sample $N$ continuations and retain the revision if any continuation reaches the correct mark, $m = m^*$. The surviving rollouts are used as SFT samples. To preserve the policy's performance on easy or already-correct cases, we also mix the originally successful rollouts in the training pool. The policy is then trained on this combined dataset.

### 2.3 Phase 2: EDIT-RL

Initialised from the EDIT-SFT policy, EDIT-RL applies distance-aware GRPO (Shao et al., 2024) augmented by a belief-guided reward shaping term. By re-using the same posterior probe, this shaping term calibrates marking accuracy while penalising harmful drifts in intermediate beliefs.

**Distance-aware outcome reward.** We use a dense reward based on mark distance:

$$r_{\text{out}} = \max\Big(0,\, 1 - \tfrac{|m-m^*|}{M_q}\Big) \tag{5}$$

with $r_{\text{out}} = 0$ on parse failure. This reward gives higher scores to rollout whose predicted mark $m$ is closer to the gold mark $m^*$, providing distance-sensitive signal for RL calibration.

**Belief-guided reward shaping.** After EDIT-SFT, the model has learned to avoid reasoning steps that cause large harmful drifts in its posterior marking belief. In EDIT-RL, we strengthen this behaviour by adding a per-rollout shaping signal that uses the posterior probe to check whether the

model's belief drifted too much from the gold mark at any intermediate step:

$$r_{\text{shape}} = -\frac{1}{M_q}\frac{1}{T}\sum_{k=1}^{T}\max\big(0, |\hat{m}_k - m^*| - \beta_{\text{drift}}\big), \tag{6}$$

where $\hat{m}_k$ is the expected posterior mark at step boundary $k$ (Eq. 1). The threshold $\beta_{\text{drift}}$ is an *exploration tolerance*, measured in raw marks, which accepts small belief fluctuations that may reflect useful exploration. However, when the belief moves beyond this tolerance, the rollout is penalised in proportion to the size of excess drift. The total reward is:

$$r = r_{\text{out}} + \gamma_\Phi \cdot r_{\text{shape}} \tag{7}$$

This reward is used directly in the distance-aware GRPO objective. As a result, a rollout can be penalised relative to others in the same group if it contains a large belief deviation, even when its final mark is close to $m^*$.

The posterior belief probe is run on the *frozen* EDIT-SFT checkpoint rather than the active policy. This maintains a stationary shaping signal, reduces computational overhead, and prevents reward hacking. We validate this choice by proving a small KL divergence between the trained policy with the initialised reference in Appendix D. Although this relaxes the strict policy-invariance of classic PBRS (Ng et al., 1999), it provides a stable, task-aligned constraint that significantly enhances out-of-distribution generalisation.

## 3 Experiment

### 3.1 Experiment Setup

**Datasets** We evaluate policy grading on two complementary benchmarks, covering five datasets in total. The first benchmark is **SAS**, drawn from SAS-Bench (Lai et al., 2025), a Chinese short-answer scoring benchmark. We use its three diverse subjects (History, Geography, and Physics). The second benchmark is **Private-Science**[1], a proprietary collection of student responses to GCSE-level science exam questions in two subjects (Biology and Physics), each marked by trained examiners against an official mark scheme. The two benchmarks are deliberately complementary. SAS comprises many questions with only a handful of responses each, while Private-Science contains fewer questions, but each is answered by hundreds of students. Full dataset statistics are given in Appendix A.

**Evaluation protocol** For Private-Science, we use three splits: a train set, an **in-distribution (ID)** test set of held-out responses to questions that *appear* in training, and an **out-of-distribution (OOD)** test set of responses to questions that are *never* seen in training. Separating ID from OOD tests whether a grader generalises to unseen questions rather than fitting question-specific surface patterns. SAS has no ID split. Each SAS question receives only two to five responses, so we evaluate the SAS only on the OOD questions, treating the SAS dataset as a subject-level fitting question.

Following standard practice in automated scoring (Li et al., 2025), we use quadratic weighted kappa (QWK) as the main metric, macro-averaged over questions. We also report exact-match accuracy, within-1 accuracy, and mean absolute error (MAE).

**Baselines** All methods use **Qwen3-8B** as the policy model, adapted with LoRA. Training details are reported in Appendix B. We keep the backbone fixed across methods so that performance differences mainly reflect the training objective rather than model capacity. We compare our method after each phase (EDIT-SFT and EDIT-RL) against four baselines:
(1) `Base`, the off-the-shelf Qwen3-8B grader.
(2) `GRPO` (Shao et al., 2024), RL on standard group-relative distance-aware outcome-reward.
(3) `DGPO` (Jin et al., 2026), RL with distribution-guided token-level advantage reallocation.
(4) `InT` (Yang et al., 2026), using SFT and RL stages based on self-proposed interventions.

EDIT and `InT` contain an SFT stage followed by an RL stage, whereas `GRPO` and `DGPO` apply RL directly to the base model.

### 3.2 Grading Performance

We first discuss the public SAS benchmark, where every evaluation question is unseen during training. As shown in Table 1, EDIT-RL achieves the best results on average across all metrics. In particular, it improves macro-QWK from 0.727 for `Base` and 0.731 for `InT` to 0.794. These results show that the method improves grading accuracy in a fully OOD setting.

The subject-level results show some variation. EDIT-RL gives the best QWK on History and

[1] A proprietary dataset shared with us under a data-sharing agreement. Licensing and privacy constraints prevent public release.

| Subject | Metric | Base | *RL-only* | | InT | | EDIT (Ours) | |
|---|---|---|---|---|---|---|---|---|
| | | | GRPO | DGPO | SFT | RL | EDIT-SFT | EDIT-RL |
| History | ACC↑ | 0.419 | 0.471 | 0.462 | 0.426 | 0.461 | 0.454 | **0.477** |
| | within-1↑ | 0.512 | 0.597 | 0.600 | 0.566 | **0.625** | 0.569 | 0.623 |
| | MAE↓ | 2.612 | 1.883 | 1.865 | 1.852 | **1.569** | 1.952 | 1.615 |
| | QWK↑ | 0.783 | 0.873 | 0.876 | 0.871 | 0.889 | 0.863 | **0.899** |
| Geography | ACC↑ | 0.414 | 0.400 | 0.367 | 0.367 | 0.333 | **0.433** | **0.433** |
| | within-1↑ | 0.552 | 0.533 | 0.600 | 0.600 | 0.533 | **0.633** | 0.600 |
| | MAE↓ | 1.483 | 1.533 | 1.533 | 1.533 | 1.667 | **1.267** | 1.300 |
| | QWK↑ | 0.706 | 0.676 | 0.685 | 0.745 | 0.675 | **0.779** | 0.756 |
| Physics | ACC↑ | 0.333 | 0.311 | 0.289 | 0.311 | 0.356 | **0.378** | **0.378** |
| | within-1↑ | 0.511 | 0.489 | 0.489 | 0.444 | 0.444 | **0.556** | 0.533 |
| | MAE↓ | 1.756 | 2.133 | 1.956 | 2.444 | 2.022 | **1.600** | 1.644 |
| | QWK↑ | 0.693 | 0.518 | 0.602 | 0.427 | 0.630 | 0.701 | **0.729** |
| **SAS-Avg** | ACC↑ | 0.389 | 0.394 | 0.372 | 0.368 | 0.383 | 0.421 | **0.429** |
| | within-1↑ | 0.525 | 0.540 | 0.563 | 0.535 | 0.534 | **0.586** | **0.586** |
| | MAE↓ | 1.950 | 1.850 | 1.784 | 1.943 | 1.753 | 1.606 | **1.520** |
| | QWK↑ | 0.727 | 0.689 | 0.721 | 0.681 | 0.731 | 0.781 | **0.794** |

Table 1: Main results on **SAS**. All SAS questions are unseen during training, so every split is out-of-distribution (no ID split). Metrics are over the combined out-of-distribution evaluation set, with the best result per row in **bold**.

Physics. `InT-RL` remains strongest on History for several metrics, while EDIT-SFT performs best on Geography, where all RL variants degrade the performance. However, EDIT-RL remains the only RL calibrated model that strongly outperforms `Base`. Appendix H traces most of this variation to final score-band and credit-synthesis errors. For example, `InT-SFT` never withholds full credit on zero-score Physics responses, whereas EDIT-SFT correctly does so on 64% of them. Nevertheless, our full method provides the most consistent overall improvement across SAS.

Because the smallest SAS subjects contain only 25–30 test responses, we further assess significance at the question level, pooling all 37 held-out questions with equal weight. Intervals use a cluster bootstrap over questions, comparisons against `Base` are paired, and Holm correction is applied across the six comparisons (Appendix G). As shown in Table 3, baseline gains near 0.03 do not survive correction, whereas EDIT's gains are two to three times larger and remain significant (both $p$=0.0006).

We next evaluate Private-Science, where both ID and OOD splits are available. As shown in Table 2, EDIT-SFT achieves the best QWK among all baselines on Private-Biology, improving macro-QWK by $+0.011$ on the ID-test split and $+0.013$ on the OOD-pool over the strongest baseline (`GRPO`). `InT-RL` outperforms across multiple metrics in Private-Physics ID setting, while degrading drastically in the OOD, suggesting an overfitting on the trained types of questions. In contrast, EDIT-SFT shows steady gains across both ID and OOD. This suggests that the improvement is not simply due to fitting question-specific patterns seen during training. We attribute the improvement from EDIT-SFT to three design choices. First, internal signals help locate reasoning steps that are likely to need revision. Second, atomic revisions keep the intervention local and avoid disrupting the full reasoning chain. Third, the rubric-checklist context gives the revisor task-specific guidance without forcing it to copy the checklist directly. These designs together preserve the model's original reasoning capabilities while making its marking reasoning chain more grounded and better aligned with the rubric.

EDIT-RL further improves performance over EDIT-SFT on both test pools. It improves macro-QWK by $+0.016$ and $+0.027$ on Biology's ID and OOD, respectively. It also improves ACC, within-1, and MAE, giving the best results across all metrics. EDIT-RL attains the highest OOD QWK on both subjects and the best value in seven of eight OOD subject-by-metric cells. This suggests that belief-guided reward shaping is especially useful when the model must grade unseen questions.

Table 5 further separates the effect of the SFT and the RL. Applying EDIT-RL directly to `Base` does not improve OOD performance over standard GRPO. In contrast, applying RL after EDIT-SFT gives stronger results, and the full combination EDIT-SFT+EDIT-RL is best on every metric. Figure 3 illustrates the QWK dynamics on the ID and

| Subject | Split | Metric | Base | *RL-only* | | InT | | EDIT (Ours) | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | GRPO | DGPO | SFT | RL | SFT | RL |
| Private-Biology | ID | ACC↑ | 0.456 | 0.491 | 0.470 | 0.437 | 0.448 | 0.485 | **0.507** |
| | | within-1↑ | 0.793 | 0.809 | 0.801 | 0.812 | 0.828 | 0.821 | **0.831** |
| | | MAE↓ | 0.821 | 0.753 | 0.790 | 0.838 | 0.803 | 0.756 | **0.717** |
| | | QWK↑ | 0.701 | 0.727 | 0.704 | 0.630 | 0.648 | 0.738 | **0.754** |
| | OOD | ACC↑ | 0.460 | 0.478 | 0.478 | 0.422 | 0.425 | 0.475 | **0.489** |
| | | within-1↑ | 0.785 | 0.825 | 0.817 | 0.837 | 0.836 | 0.816 | **0.840** |
| | | MAE↓ | 0.786 | 0.718 | 0.725 | 0.777 | 0.772 | 0.735 | **0.687** |
| | | QWK↑ | 0.524 | 0.535 | 0.533 | 0.494 | 0.508 | 0.548 | **0.575** |
| Private-Physics | ID | ACC↑ | 0.438 | 0.489 | 0.495 | 0.454 | **0.512** | 0.506 | **0.512** |
| | | within-1↑ | 0.762 | 0.780 | 0.795 | 0.804 | **0.822** | 0.782 | 0.815 |
| | | MAE↓ | 0.906 | 0.823 | 0.801 | 0.813 | **0.782** | 0.812 | 0.795 |
| | | QWK↑ | 0.477 | 0.480 | 0.535 | 0.454 | 0.551 | 0.527 | **0.555** |
| | OOD | ACC↑ | 0.429 | 0.431 | 0.435 | 0.374 | 0.433 | **0.452** | 0.445 |
| | | within-1↑ | 0.792 | 0.808 | 0.793 | 0.784 | 0.815 | 0.813 | **0.817** |
| | | MAE↓ | 0.829 | 0.805 | 0.820 | 0.950 | 0.840 | 0.784 | **0.772** |
| | | QWK↑ | 0.456 | 0.465 | 0.449 | 0.393 | 0.409 | 0.472 | **0.487** |

Table 2: Main results on **Private-Science**. ↑/↓ indicate higher/lower is better. Best fine-tuning result per row in **bold** (per-subject rows).

| Method | QWK [95% CI] | Δ vs. Base (Holm $p$) |
|---|---|---|
| `Base` | 0.760 [0.677, 0.830] | – |
| `GRPO` | 0.794 [0.707, 0.867] | +0.034 ($p$=0.391) |
| `DGPO` | 0.809 [0.732, 0.872] | +0.049 ($p$=0.021) |
| `InT-SFT` | 0.793 [0.707, 0.868] | +0.033 ($p$=0.391) |
| `InT-RL` | 0.822 [0.750, 0.883] | +0.062 ($p$=0.051) |
| EDIT-SFT | 0.831 [0.764, 0.887] | +0.072 ($p$=**0.0006**) |
| EDIT-RL | **0.856** [0.784, 0.913] | +0.096 ($p$=**0.0006**) |

Table 3: Question-level pooled QWK on SAS with cluster-bootstrap 95% CIs and Holm-corrected $p$-values against `Base` (Appendix G).

| Dataset | +GRPO | +EDIT-RL | Δ |
|---|---|---|---|
| Private-Biology | 0.559 ±0.003 | **0.580** ±0.004 | +0.021* |
| SAS-History | 0.873 ±0.008 | **0.885** ±0.004 | +0.012 |

Table 4: OOD QWK (mean ± SD over three seeds). Δ is the mean seed-wise paired difference. * marks a paired 95% CI excluding zero (CIs: $[+0.011, +0.030]$ and $[-0.007, +0.030]$).

| Method | ID | | OOD | |
|---|---|---|---|---|
| | QWK↑ | ACC↑ | QWK↑ | ACC↑ |
| `Base` | 0.701 | 0.456 | 0.524 | 0.460 |
| `Base` + GRPO | 0.727 | 0.491 | 0.535 | 0.478 |
| `Base` + EDIT-RL | 0.727 | 0.483 | 0.518 | 0.461 |
| EDIT-SFT + GRPO | 0.748 | 0.506 | 0.560 | 0.484 |
| EDIT-SFT + EDIT-RL | **0.754** | **0.507** | **0.575** | **0.489** |

Table 5: RL choices (single seed) applied either directly on `Base` or on the EDIT-SFT checkpoint, Private-Biology. The combination of EDIT-SFT SFT and EDIT-RL (belief-guided shaped GRPO) is best on every metric/split.

OOD with RL proceeds. Although the improvement matches with a GRPO on EDIT-SFT in ID, EDIT-RL shows steady improvement on the OOD setting. This suggests that belief-guided reward shaping is most effective when the model has first been trained on internally located, rubric-aware atomic revisions.

To isolate the effect of the belief-shaped reward from seed noise, Table 4 repeats this comparison over three seeds on two datasets. Both arms share the same EDIT-SFT checkpoint and identical budgets, decoding, and final checkpoint, differing only in the reward term. EDIT-RL outperforms GRPO in all three paired runs on both datasets, and the paired 95% CI excludes zero on Private-Biology.

### 3.3 Ablation for Rule-Aware Atomic Intervention

We ablate the main design choices of EDIT-SFT on SAS-History, the largest dataset in the SAS suite we adopted. The ablation focuses on three components: internal-signal localisation, rubric-checklist context, and atomic locality constraints. These components test whether performance gains come from selecting better edit locations, giving the revisor better rule context, and keeping the revision local.

As shown in Table 6, removing any module degrades performance. Excluding the rubric-checklist context leads to the largest drop in QWK and MAE while leaving ACC largely unchanged, suggesting

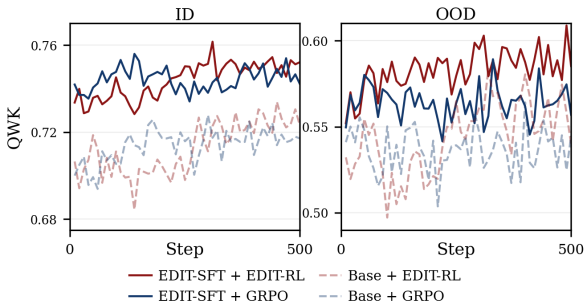


Figure 3: Training dynamics of RL choices (GRPO, EDIT-RL) with base choices (Base, EDIT-SFT) in QWK on ID and OOD test set. Curves are computed post hoc from saved checkpoints (Appendix B).

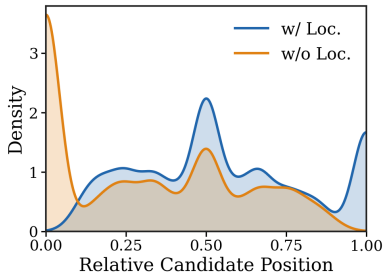


Figure 4: Comparison between the relative positions of the candidate step selected with or without the internal-state driven localiser from EDIT-SFT.

increased variance in grading behaviour. Removing atomic locality constraints during revision results in a slight relative degradation. This is likely because SAS scoring is naturally point-based and therefore already imposes contextual constraints. EDIT-SFT without internal-signal localisation exhibits the largest drop across all metrics, highlighting the critical role of our step-candidate selection for revision. Appendix F further shows counterfactually that EDIT-selected steps recover the gold mark more often than position-matched random or belief-shift-only selections. Figure 4 further shows that the base model tends to select earlier steps, whereas our localisation module selects middle or final steps, where reasoning synthesis typically occurs.

| **Ablation** | QWK↑ | ACC↑ | Within-1↑ | MAE↓ |
|---|---|---|---|---|
| EDIT-SFT | **0.863** | **0.454** | **0.569** | **1.952** |
| w/o Localiser | 0.820 | 0.408 | 0.515 | 2.300 |
| w/o Checklist | 0.817 | 0.454 | 0.554 | 2.346 |
| w/o Atomic | 0.842 | 0.446 | 0.562 | 2.046 |

Table 6: Ablation of three core designs of EDIT-SFT on SAS-History, reported in all four metrics.

### 3.4 Rule-Faithfulness under Interventions

We also test whether a grader *truly follows the rubric*, rather than relying on its internalised scoring scale. To do this, we apply deterministic edits to the mark scheme where the gold-score change can be computed exactly. We then measure the Rule-Sensitivity Ratio (RSR), defined as the ratio between the model's score change and the gold score change. (RSR=1 means that the model follows the rule edit exactly; <1 under-reacts, >1 over-reacts). The edits include level-of-response band shifts (LoR) and points-total rescaling (PTS). The detailed intervention strategies are reported in Appendix I.1. One complication is that a predicted score of 0 is unchanged by every edit, so we decompose RSR into two parts: 1. Floor Rate (FR), the fraction of predictions equal to zero, and 2. In-Rubric RSR (IB-RSR), responsiveness only among non-zero predictions. A further complication is that *signed* IB-RSR means can hide errors, because over-reaction and under-reaction cancel when averaged across conditions. We therefore report AE-RSR, the mean of $|\text{IB-RSR} - 1|$ across edit conditions (lower is better).

Table 7 averages results across the LoR and PTS intervention strategies. Under LoR band shifts, every grader is near-faithful. EDIT-SFT attains the lowest AE-RSR (0.070) and EDIT-RL the lowest floor rate (6.60%). The separation appears under PTS rescaling, where Base is the weakest system (AE-RSR 0.733) and EDIT-SFT the strongest (0.500). Aggregated over all six conditions, EDIT-SFT reaches 0.285 against 0.412 for Base. InT attains PTS responsiveness comparable to EDIT's, but partly by collapsing predictions to the floor (28.57% vs. 19.90% for the SFT variants), which removes responsive rows from the in-rubric statistic.

Overall, EDIT-SFT is the most rule-responsive system, and EDIT-RL trades a small amount of responsiveness for grading accuracy. Meanwhile, PTS rescaling remains unsolved for every evaluated system (AE-RSR ≥0.5), pointing to absolute-count anchoring as an open challenge for rubric grading. Our intervention protocol exposes exactly this kind of failure, which accuracy metrics alone cannot detect: a grader can score well while remaining insensitive to the rule it is meant to apply.

| Method | LoR | | PTS | | Overall |
|---|---|---|---|---|---|
| | AE↓ | FR (%) | AE↓ | FR (%) | AE↓ |
| Base | 0.090 | 10.07 | 0.733 | 22.43 | 0.412 |
| GRPO | 0.093 | 7.17 | 0.710 | 18.80 | 0.402 |
| DGPO | 0.090 | 7.20 | 0.697 | 19.33 | 0.393 |
| InT-SFT | 0.093 | 10.57 | 0.530 | 28.57 | 0.312 |
| InT-RL | 0.143 | 9.70 | 0.507 | 26.80 | 0.325 |
| EDIT-SFT | **0.070** | 7.00 | **0.500** | 19.90 | **0.285** |
| EDIT-RL | 0.083 | 6.60 | 0.510 | 19.87 | 0.297 |

Table 7: Rule faithfulness on Private-Biology. AE = AE-RSR (lower is better). FR = floor rate (%). Overall averages AE-RSR over all six conditions, with per-condition signed results in Appendix I.2.

## 4 Related Work

**Credit Assignment for Reinforcement Learning** Credit assignment refers to the problem of attributing a delayed outcome to specific intermediate decisions, which remains a core challenge in LLM reasoning. Standard sequence-level objectives like GRPO (Shao et al., 2024) broadcast a single scalar reward uniformly across a trajectory. To refine this coarse signal, recent work generally falls into three paradigms. Return redistribution methods, such as DGPO (Jin et al., 2026), reallocate trajectory-level returns to individual steps. Process reward models (PRMs), like Math-Shepherd (Wang et al., 2024), train explicit step-level verifiers to provide fine-grained supervision. PRMs derive the process signal from external step-level labels, whereas EDIT derives it from the model's own decision state under final-mark supervision only. Such labels are also rubric-specific, defined only relative to one mark scheme, so we regard PRM-style supervision as a complementary direction that requires per-rubric annotation. Alternatively, hindsight interventions, such as InT (Yang et al., 2026), prompt the policy to self-audit and rewrite erroneous steps. However, a growing body of work shows that LLM-generated chains-of-thought frequently diverge from actual computation, often necessitating external rewards to enforce faithfulness (Turpin et al., 2023; Lanham et al., 2023). While prior methods rely on external verifiers or textual self-critique, our framework directly localises critical steps using the model's well-calibrated internal signals and guides the policy with belief-guided reward shaping.

**LLM-based Automated Student Assessment** Automated essay scoring (AES) has increasingly transitioned from training dedicated classifiers to utilizing pretrained LLMs as evaluators or distilling their rationales into smaller models (Li and Ng, 2024; Li et al., 2023). For short-answer and science grading, recent benchmarks and systems primarily rely on prompt engineering or verbal reflection of frozen models (Lai et al., 2025; Li et al., 2025). Other approaches explore aligning evaluation rationales via preference optimization on thought trees to improve explainability (Li et al., 2024). While most existing systems treat grading as a zero-shot prompting task or rely on external solvers to follow rules, our work frames grading as an end-to-end trainable credit-assignment problem. By supplying the rubric checklist as a privileged input during the rewriting phase, our approach decouples analytical reasoning from rigid structural formats to ensure faithful adherence to external grading criteria.

## 5 Conclusion

In this paper, we propose EDIT, a novel two-phase framework addressing the unique credit-assignment challenges of rule-faithful LLM grading. To enforce strict external grounding, EDIT replaces unreliable prompt-based self-audits with internal-state localisation and privileged-context rewriting. Furthermore, our belief-guided reward shaping guides the uncertainty reduction process by penalising severe belief excursions while tolerating benign exploration. Experiments on two real-world grading datasets demonstrate that EDIT significantly outperforms state-of-the-art baselines across in-domain and out-of-domain splits. Future work will explore generalising the framework to broader domains, streamlining the training pipeline, and dynamically decomposing holistic rubrics.

## Limitations

We identify four key limitations of the EDIT framework.

First, our evaluation scenarios are currently restricted in scope. Although we frame student answer scoring as a complex rule-based reasoning task, we evaluate EDIT on two datasets primarily consisting of short-answer and fill-in-the-blank questions. We have not yet tested our framework on long-form essay scoring tasks, such as those in the Automated Student Assessment Prize (ASAP) benchmark, nor have we explored its applicability in other high-stakes rule-based reasoning domains like legal judgment or medical diagnosis. In addition, the Private-Science OOD splits contain only two to three questions per subject, so our question-

level significance analysis is limited to SAS. An important direction for future work is to verify and extend the generalisability of our method across broader domains and longer, more unstructured text formats.

Second, the proposed training pipeline involves considerable complexity. While our probing-based internal-state localisation reduces inference overhead compared to prior prompt-elicited self-audits, the overall framework still relies on a heavy multi-stage process: rigorous SFT data construction followed by a calibration RL phase (GRPO). This multi-step orchestration increases both computational complexity and engineering effort. In addition, while the two RL arms in our controlled comparison are budget-matched, the pipeline-level comparison against baselines is not: EDIT incurs extra offline compute for posterior-belief probing, grounding audits, and revision sampling (Appendix B). Future research should investigate more streamlined, end-to-end training paradigms that simplify the pipeline while preserving the precise credit-assignment capabilities of evidence-diagnosed interventions.

Third, EDIT relies on the availability and granularity of explicit rule sets. Specifically, EDIT-SFT utilises a per-criterion rubric checklist as privileged input to constrain atomic rewrites. However, in many real-world settings, rubrics can be vague, holistic, or implicitly defined by human graders. The effectiveness of our internal-state diagnostics and rule-faithful rewriting may degrade when applied to such unstructured or noisy criteria. Future work could explore mechanisms to dynamically decompose holistic rubrics into structured checklists, or adapt the framework to handle fuzzy and implicit rule constraints.

Finally, like other outcome-filtered self-improvement methods in the STaR and ReST family (Zelikman et al., 2022; Gulcehre et al., 2023), EDIT-SFT retains only revisions whose continuations reach the gold mark: 62.8% of incorrect rollouts on Private-Biology and 76.4% on SAS-History yield no corrected example (Appendix F). This costs sampling compute rather than data validity, since every retained trajectory is verified against the gold mark, but the outcome filter does bias the retained pool toward cases that are repairable by a local edit.

## Ethics Statement

This research has been approved by our institutional ethics committee. Our dataset comprises LLM-augmented open-source question banks and proprietary data from a collaborative institution (which has also passed their internal rigorous ethical review). All data has been strictly de-identified to remove Personally Identifiable Information (PII). Given that the open-source subset relies on LLM generation, it may inherit sociopolitical or linguistic biases from pre-training corpora. We strongly recommend manual review before using this data for downstream training to avoid penalising diverse linguistic expressions or propagating cognitive distortions.

Furthermore, our framework is intended solely for researching LLM reasoning in rule-constrained environments. Automated grading involves high-stakes decisions with profound psychological impacts on students' academic trajectories. We strictly advise against deploying these models in real-world educational settings without human-in-the-loop to ensure fairness, accountability, and pedagogical empathy.

## Acknowledgements

This work was supported in part by the UK Engineering and Physical Sciences Research Council (EPSRC) through the Prosperity Partnership scheme (grant no. UKRI566) and a Turing AI Fellowship (grant no. EP/V020579/1, EP/V020579/2).

## References

Christian Grévisse. 2024. LLM-based automatic short answer grading in undergraduate medical education. *BMC Medical Education*, 24(1):1060.

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, Wolfgang Macherey, Arnaud Doucet, Orhan Firat, and Nando de Freitas. 2023. Reinforced self-training (ReST) for language modeling. *arXiv preprint arXiv:2308.08998*.

Hongbo Jin, Rongpeng Zhu, Zhongjing Du, Xu Jiang, Jingqi Tian, Qiaoman Zhang, and Jiayu Ding. 2026. DGPO: Distribution-guided policy optimization for fine-grained credit assignment. *arXiv preprint arXiv:2605.03327*.

Peichao Lai, Kexuan Zhang, Yi Lin, Linyihan Zhang, Feiyang Ye, Jinhao Yan, Yanwei Xu, Conghui He, Yilei Wang, Wentao Zhang, and Bin Cui. 2025. SAS-Bench: A fine-grained benchmark for evaluating

short answer scoring with large language models. *arXiv preprint arXiv:2505.07247*.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, and 1 others. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*.

Jiazheng Li, Lin Gui, Yuxiang Zhou, David West, Cesare Aloisi, and Yulan He. 2023. Distilling ChatGPT for explainable automated student answer assessment. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6007–6026, Singapore. Association for Computational Linguistics.

Jiazheng Li, Hainiu Xu, Zhaoyue Sun, Yuxiang Zhou, David West, Cesare Aloisi, and Yulan He. 2024. Calibrating LLMs with preference optimization on thought trees for generating rationale in science question scoring. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5452–5479, Miami, Florida, USA. Association for Computational Linguistics.

Jiazheng Li, Yuxiang Zhou, Junru Lu, Gladys Tyen, Lin Gui, Cesare Aloisi, and Yulan He. 2025. Two heads are better than one: Dual-model verbal reflection at inference-time. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3119–3140, Suzhou, China. Association for Computational Linguistics.

Shengjie Li and Vincent Ng. 2024. Automated essay scoring: A reflection on the state of the art. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 17876–17888, Miami, Florida, USA. Association for Computational Linguistics.

Andrew Y. Ng, Daishi Harada, and Stuart J. Russell. 1999. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning (ICML)*, pages 278–287.

Eduardo Pignatelli, Johan Ferret, Matthieu Geist, Thomas Mesnard, Hado van Hasselt, and Laura Toni. 2024. A survey of temporal credit assignment in deep reinforcement learning. *Transactions on Machine Learning Research*. Survey Certification.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Yu Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36. ArXiv:2305.04388.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439, Bangkok, Thailand. Association for Computational Linguistics.

Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S Yu, and Qingsong Wen. 2026. Large language models for education: A survey and outlook. *IEEE Signal Processing Magazine*, 42(6):51–63.

Matthew Y. R. Yang, Hao Bai, Ian Wu, Gene Yang, Amrith Setlur, and Aviral Kumar. 2026. InT: Self-proposed interventions enable credit assignment in LLM reasoning. *arXiv preprint arXiv:2601.14209*.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. STaR: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35.

## A Dataset Statistics

Table 8 reports, for each dataset, the number of responses and distinct questions in every split, together with the range of per-question maximum marks. The three splits play different roles across the two benchmarks. For **Private-Science**, each question is answered by hundreds of students: the *Test* split holds out responses to questions that are *seen* during training (in-distribution, ID), whereas the *Holdout* split consists of entirely *unseen* questions (out-of-distribution, OOD). For **SAS**, each question has only a handful of responses, so questions are partitioned disjointly across all three splits; *both* the Test and Holdout splits are therefore out-of-distribution, and SAS has no ID split.

**Data availability.** The Private-Science responses were fully anonymised by the data provider before being shared with us. We report experimental results on this dataset, but it cannot be publicly released owing to licensing and privacy constraints. The SAS benchmark is publicly available (Lai et al., 2025).

## B Implementation Details

**Training** All methods adapt Qwen3-8B using a shared LoRA recipe: rank $r$=64, $\alpha$=128, applied to all attention and MLP projection matrices ($q, k, v, o$, gate, up, down). We train for 2 epochs at learning rate $2 \times 10^{-4}$ (cosine schedule, 10% warmup). For the SAS dataset, we set the effective batch size to 8 due to the small amount of data, and

| Benchmark | Dataset | Train (resp / q) | ID (resp / q) | OOD (resp / q) | Marks |
|---|---|---|---|---|---|
| Private-Science (proprietary) | Biology | 10,159 / 12 | 1,264 / 12 | 2,087 / 2 | 3–6 |
| | Physics | 1,573 / 7 | 395 / 7 | 655 / 3 | 3–6 |
| SAS (Lai et al., 2025) | History | 510 / 102 | – | 130 / 26 | 11–26 |
| | Geography | 110 / 22 | – | 30 / 6 | 10 |
| | Physics | 95 / 19 | – | 25 / 5 | 5–20 |

Table 8: Dataset statistics. "resp / q" is the number of responses and the number of distinct questions in each split. For Private-Science, the ID split is in-distribution (held-out student responses to seen questions), and the OOD split is out-of-distribution (unseen questions). For SAS, there is no ID split.

to 32 on Private-Science. The same recipe is used for all subjects and methods, so comparisons are not confounded by hyperparameter differences.

**Decoding and evaluation** At evaluation, we obtain 1 completion per marking using greedy sampling. This ensures consistency across all datasets and methods. QWK is computed per question over the integer score range $[0, M_q]$, where $M_q$ is the question's maximum mark, using quadratic weights, and then macro-averaged across questions.

**RL configuration and checkpoint selection** No test metric or test data enters hyperparameter or checkpoint selection at any point, and every reported result uses the final-step checkpoint rather than a selected peak. Training horizons are fixed in advance at 200 RL steps for SAS and 500 for Private-Science. All RL methods share one optimisation recipe (Table 9): learning rate $5 \times 10^{-6}$ with a cosine schedule and 3% warmup, KL coefficient $\beta_{\mathrm{KL}}$=0.02 towards the initialisation, 8 rollouts per prompt and 4 prompts per optimiser step (32 completions), sampling temperature 1.0, and a completion budget of 1,024 tokens. A fresh LoRA adapter with the SFT configuration ($r$=64, $\alpha$=128, all projection matrices) is trained on top of the merged initialisation, which also serves as the KL reference. The test-set curves in Figure 3 were generated post hoc after training had finished and were never used for monitoring or selection. The three-seed comparison in Table 4 repeats the full training run with seeds 42, 43, and 44.

EDIT-RL introduces two additional hyperparameters, the drift tolerance $\beta_{\text{drift}}$ and the shaping weight $\gamma_\Phi$. Both are calibrated per dataset on a random 10% subset of training rollouts using signal statistics only, never a performance metric. We sweep $\beta_{\text{drift}}$ and keep the value that maximises the separation between the shaping-penalty firing rates on incorrect and correct rollouts, and we set $\gamma_\Phi$ to the largest value that remains *flip-safe*: the accumulated penalty on a correct rollout stays below the outcome-reward gap to an incorrect one, so the shaping term can never invert the outcome ranking within a group. The posterior probe is read at temperature 0.7. The selected values are listed in Table 9.

| *Shared by all RL methods* | |
|---|---|
| Learning rate | $5 \times 10^{-6}$ |
| LR schedule | cosine, 3% warmup |
| KL coefficient $\beta_{\mathrm{KL}}$ | 0.02 |
| Rollouts per prompt | 8 |
| Prompts per step | 4 (32 completions) |
| Sampling temperature | 1.0 |
| Completion budget | 1,024 tokens |
| LoRA | as in SFT ($r$=64, $\alpha$=128) |
| Steps, Private-Science | 500 |
| Steps, SAS | 200 |
| EDIT-RL *only:* $\beta_{\text{drift}}$ (marks) / $\gamma_\Phi$ | |
| Private-Biology | 0.75 / 0.3 |
| Private-Physics | 0.5 / 0.15 |
| SAS-History | 0.75 / 0.05 |
| SAS-Geography | 0.75 / 0.10 |
| SAS-Physics | 0.75 / 0.10 |

Table 9: RL hyperparameters. The upper block is shared by GRPO, DGPO, InT-RL, and EDIT-RL. The lower block gives the per-dataset belief-shaping parameters of EDIT-RL.

**Compute budget** All RL runs use a single GPU (H200 141 GB) with the rollout engine co-located. SFT uses two GPUs. We compare the two phases separately on Private-Biology, the largest dataset. *Phase 1 (data construction).* EDIT-SFT and InT start from the same pool of 20 sampled rollouts per training response and the same 55,278 incorrect rollouts. InT spends $\approx$ 8 GPU-hours proposing interventions and sampling guided continuations. EDIT-SFT spends $\approx$ 9 GPU-hours: $\approx$ 2.5 for posterior-belief probing, $\approx$ 4 for candidate selection and the grounding audit, and $\approx$ 2.5 for sampling and validating atomic revisions. The probe reads the mark posterior at every step boundary of

a rollout (5.2 boundaries on average). We obtain all of them in a single forward pass by appending one answer scaffold per boundary and using a block attention mask with matching position indices, so that each scaffold attends only to its own prefix. The resulting SFT pools hold 56,595 samples for EDIT-SFT (33,958 golden rollouts and 22,637 revisions) and 29,822 for InT. A retained rollout can contribute more than one revised candidate step, so the number of revisions slightly exceeds the number of repaired rollouts implied by the retention rate of Appendix F. The two pools contain comparable numbers of generated revisions, and the difference is the golden anchor set, which is drawn from the shared rollout pool at no generation cost. SFT cost is proportional to pool size and identical per sample. *Phase 2 (RL).* EDIT-RL and GRPO from the same EDIT-SFT checkpoint differ only by the frozen-probe read on each rollout group, which adds about 2 s to a 12 s optimiser step: $\approx 1.9$ versus 1.7 GPU-hours for the 500-step run. Base-initialised GRPO and DGPO incur only this RL cost. On the smaller datasets every phase takes at most a few GPU-hours for either method.

## C Candidate Step Selection Procedure

Algorithm 1 formalises the two-stage selection rule of §2.2. The belief-shift criterion alone forms the shortlist. The grounding audit then re-ranks within that shortlist, promoting weakly grounded steps, and never excludes a step. In particular, a step with a large negative belief shift and strong grounding remains fully eligible.

**Algorithm 1** Candidate step selection in EDIT-SFT

**Require:** incorrect rollout $s_{1:T}$ ($m \neq m^*$); belief shifts $\{\Delta\epsilon_j\}$; grounding scores $\{G^\rho_j, G^a_j, \widetilde{G}^p_j\}$; shortlist size $k$; candidate budget $n$; near-tie margin $\tau$; dataset-wide 25th percentiles $p_{25}(\cdot)$

$\mathcal{C} \leftarrow \{\, j : \Delta\epsilon_j < -\tau \,\}$ ▷ belief-shift criterion alone

$\mathcal{C} \leftarrow$ top-$k$ of $\mathcal{C}$ by most-negative $\Delta\epsilon_j$ ▷ shortlist

**for** $j \in \mathcal{C}$ **do**

  $w_j \leftarrow \big[G^\rho_j < p_{25}(G^\rho)\big] \wedge \big[G^a_j < p_{25}(G^a)\big] \wedge \big[\widetilde{G}^p_j < p_{25}(\widetilde{G}^p)\big]$

**end for**

stably reorder $\mathcal{C}$ so steps with $w_j{=}1$ precede those with $w_j{=}0$; ties keep the $\Delta\epsilon$ order ▷ re-rank only; no step removed

**return** first $\min(n, |\mathcal{C}|)$ steps of $\mathcal{C}$

## D Validity of the Frozen-Probe Reference

EDIT-RL reads the per-step mark belief $\hat{m}_k$ from a *frozen* copy of the EDIT-SFT policy and penalises its drift from gold. We freeze the reference for two reasons. (i) **Non-gameability:** RL could lower the shaping penalty by shifting its own belief *readout* rather than by grading more faithfully. A fixed, exogenous reference removes this shortcut. (ii) **Calibration:** the EDIT-SFT policy is already a competent grader, so its step-boundary mark posterior is a meaningful target for "where belief should sit," unlike the raw policy. This design is sound only if the trained policy's belief remains well-described by the frozen readout, so we verify this directly.

We probe the mark posterior $\hat{p}_k(m)$ (Eq. 1) at every step boundary of 200 Private-Biology test rollouts under the frozen EDIT-SFT and under each SFT-initialised RL policy, and measure their divergence (Table 10). The EDIT-RL policy stays very close to the frozen reference (mean KL=0.026, JSD=0.0055) and is flat across chain position, confirming that the frozen probe remains a faithful proxy for the policy's belief. Plain GRPO, the same SFT initialisation and KL-to-init term, but *without* belief shaping, drifts about twice as far (KL=0.064), and increasingly so deeper in the chain. Belief shaping thus anchors the *intermediate* trajectory to the calibrated reference, not merely the final mark.

| | | | JSD by position | | |
|---|---|---|---|---|---|
| Policy | KL | JSD | early | mid | late |
| EDIT-RL | **0.026** | **0.0055** | 0.0049 | 0.0055 | 0.0061 |
| GRPO | 0.064 | 0.0119 | 0.0099 | 0.0133 | 0.0132 |

Table 10: Step-boundary belief divergence between each SFT-initialised RL policy and the *frozen* EDIT-SFT probe reference. KL and JSD are means over boundaries. The last three columns give JSD averaged within early/mid/late chain terciles. EDIT-RL stays $\sim$2$\times$ closer to the reference and is flat across position, whereas GRPO drifts further and grows deeper in the chain.

## E Masked-Support Audit Validation

We validate the masked-support audit against external step labels produced by a stronger annotator model (not human annotations), which tag each grading step as a *grounding* step (a pointwise judgement read off the answer against the

rubric) or a *synthesis* step (one that composes prior conclusions). These model-generated labels validate the grounding signal itself. The causal question of whether selected steps are error-causing is addressed separately in Appendix F. Figure 5 shows two clean dissociations. Grounding steps occupy the high Answer-Grounding region (panel a), whereas synthesis steps collapse to $G^a \approx 0$; conversely, synthesis steps carry markedly higher Prefix Grounding (panel b, mode at $G^p \approx 1.5$ vs. a near-zero mode for grounding steps). Both dissociations are highly significant (Table 11): Answer Grounding separates the two classes with a large effect ($r$=0.71, $p < 10^{-300}$) and Prefix Grounding likewise ($r$=0.58, $p < 10^{-300}$), and both survive collapsing to per-rollout means. Rubric Grounding, in contrast, is statistically significant but practically negligible ($r$=0.08): both step types lean equally on the rubric, so the discriminating axes are *which evidence* a step reads (the answer, for grounding steps) and *whether* it integrates prior reasoning (the prefix, for synthesis steps). This confirms that $G^a$ and $G^p$ behave as intended and are well separated from zero.

| Signal | Grounding (median) | Synthesis (median) | $r$ | $p$ |
|---|---|---|---|---|
| $G^a$ | **0.72** | 0.08 | 0.71 | $<10^{-300}$ |
| $G^\rho$ | 0.89 | 0.83 | 0.08 | $5.6\times10^{-17}$ |
| $G^p$ | 0.66 | **1.64** | 0.58 | $<10^{-300}$ |

Table 11: Significance of the masked-grounding dissociations (Figure 5). $p$ is a one-sided Mann–Whitney $U$ test in the hypothesised direction (Grounding > Synthesis for $G^a$, $G^\rho$; Synthesis > Grounding for PG); $r$ is the rank-biserial effect size. $G^a$ and $G^p$ dissociate with large effects. $G^\rho$ is significant but negligible ($r$=0.08). All three remain significant when collapsed to per-rollout means ( $G^a$, $G^p$ $p\approx0$; $G^\rho$ $p$=$5.8\times10^{-36}$), ruling out within-rollout dependence as the driver.

## F Counterfactual Validation of Step Selection

We test whether the steps selected by EDIT are actually error-causing, in the counterfactual sense that repairing the selected step recovers the correct final mark. For each incorrect rollout, we compare three selection rules under an otherwise identical pipeline. **Random** draws a random step matched to EDIT's relative-position distribution. **Belief shift only** applies the shortlist criterion without grounding re-ranking. The full EDIT rule applies both. Every arm uses the same rewriter, the same $N$=4 sampled continuations, and the same gold-mark filter, so differences are attributable to the choice of step alone. **Success@4** is the fraction of rewrites in which at least one continuation reaches the gold mark. **Retention** is the fraction of incorrect rollouts that yield a usable SFT target.

| Dataset | Selection rule | S@4 | Ret. |
|---|---|---|---|
| Private-Biology | Random | 30.0 | 33.1 |
| | Belief shift only | 32.7 | 36.6 |
| | EDIT | **33.5** | **37.2** |
| SAS-History | Random | 16.6 | 18.5 |
| | Belief shift only | 17.1 | 19.6 |
| | EDIT | **22.0** | **23.6** |

Table 12: Counterfactual repair comparison of step-selection rules. S@4 = Success@4 (%). Ret. = Retention (%).

As shown in Table 12, EDIT-selected steps recover the gold mark more often than position-matched random steps on both datasets. On SAS-History, EDIT raises the repair rate from 16.6% to 22.0%, a relative gain of 1.33×, and yields 1.28× more usable SFT examples. Private-Biology moves in the same direction with a smaller margin. The advantage also survives training: removing the localiser costs 4.3 QWK points on SAS-History (Table 6).

Outcome filtering is nonetheless costly: 62.8% of incorrect rollouts on Private-Biology and 76.4% on SAS-History yield no corrected example. This costs sampling compute rather than data validity, because every retained trajectory is verified against the gold mark. The resulting selection bias toward repairable cases is discussed in the Limitations section.

## G Question-Level Significance Protocol

The pooled analysis of Table 3 treats the question as the unit of independence. All 37 held-out SAS questions are pooled with equal weight, so the pooled QWK differs from the subject-macro averages of Table 1, which first average within subjects. Confidence intervals are obtained with a cluster bootstrap that resamples entire question clusters, never splitting responses to the same question across resamples. Comparisons against `Base` are paired: each method's per-question scores are differenced against `Base` on the same questions, so the tests condition on between-question variance even where marginal intervals overlap. Holm

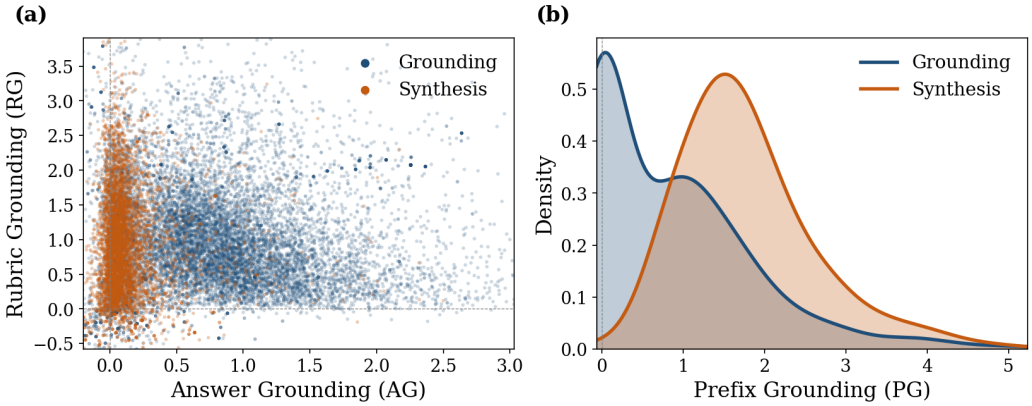


Figure 5: Grounding audit distribution with different external step label (grounding, synthesis).

correction is applied across the six comparisons against `Base`. The pooled analysis complements the per-subject averages of Table 1.

## H Subject-Level Error Analysis

The smallest SAS subjects contain only 25–30 test responses, so we draw no statistical conclusions from per-subject splits. The remaining variance is systematic. Across all 73 prediction disagreements between EDIT-SFT and `InT-SFT`, 66% are final score-band or credit-synthesis errors rather than missed evidence or rubric mapping, and the subjects stress exactly that mechanism because they differ sharply in supervision density and score range. In particular, 44% of SAS-Physics evaluation responses carry a gold score of zero, so the grader must repeatedly withhold all credit.

| System | History | Geography | Physics |
|---|---|---|---|
| `Base` | 67 | 43 | 55 |
| `InT-SFT` | **78** | 43 | 0 |
| EDIT-SFT | 72 | **57** | **64** |

Table 13: Percentage of zero-score SAS responses that each system correctly assigns zero.

Table 13 reports the percentage of zero-score responses each system correctly assigns zero. `InT-SFT` never withholds full credit on Physics, and its Physics QWK falls to 0.427, while EDIT-SFT retains 0.701 (Table 1). This follows from the training targets: `InT` inserts a corrective redirection, which biases the continuation toward finding credit, whereas EDIT rewrites the erroneous rubric judgement itself, so the model also learns when credit must be withheld.

## I Rule-Faithfulness Intervention

### I.1 Intervention Strategies

We probe rule-faithfulness using six deterministic, gold-computable edits to the mark scheme, grouped into two families that correspond to the two Private-Biology scheme types. For each edit we recompute the reference (gold) mark under the edited scheme, regrade the *same* student response with the edited scheme supplied in-prompt, and compare the model's score change to the gold change via the Rule-Sensitivity Ratio $\mathrm{RSR} = \overline{|\Delta\mathrm{pred}|}/\overline{|\Delta\mathrm{gold}|}$ (§3.4).

**Level-of-response (LoR) band shifts.** For LoR questions, the scheme maps each response-quality level to a mark band. We apply: **A1 (uniform** $+2$**)**: adding a constant 2 marks to every level ($\overline{|\Delta\mathrm{gold}|}$=1.85); **A2 (expand** $\times 2$**)**: doubling each level's mark, stretching inter-level spacing ($\overline{|\Delta\mathrm{gold}|}$=3.00); **B1 (differential)**: a non-uniform per-level shift (e.g. $+1/+2/+2$) from the lowest to the highest band.

**Points-total (PTS) rescales.** For PTS questions, the scheme awards a fixed mark per creditable point. We multiply every point's value by a constant factor: $\times 0.5$ ($\overline{|\Delta\mathrm{gold}|}$=0.35), $\times 1.5$ (0.90), and $\times 2.0$ (1.25).

**Floor confound.** A prediction of 0 is a fixed point of every edit (it cannot move), so a grader can appear faithful merely by predicting 0. We therefore decompose RSR into a *floor rate* (fraction of perturbed predictions equal to 0) and an *in-rubric RSR* (RSR computed only over rows with a non-zero perturbed prediction), and report both.

| | Base | | GRPO | | DGPO | | InT-SFT | | InT-RL | | EDIT-SFT | | EDIT-RL | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Rule edit** | FR | $R_{ir}$ | FR | $R_{ir}$ | FR | $R_{ir}$ | FR | $R_{ir}$ | FR | $R_{ir}$ | FR | $R_{ir}$ | FR | $R_{ir}$ |
| *LoR* | | | | | | | | | | | | | | |
| LoR-A1 (+2) | 8.9 | 1.00 | 5.8 | 0.81 | 6.1 | 0.83 | 10.2 | 0.82 | 9.7 | 0.78 | 6.5 | 0.87 | 6.4 | 0.86 |
| LoR-A2 ($\times$2) | 12.3 | 1.13 | 8.9 | 0.96 | 8.8 | 0.96 | 10.6 | 0.99 | 10.3 | 0.92 | 7.8 | 0.96 | 6.9 | 0.94 |
| LoR-B1 (diff) | 9.0 | 1.14 | 6.8 | 0.95 | 6.7 | 0.94 | 10.9 | 0.91 | 9.1 | 0.87 | 6.7 | 0.96 | 6.5 | 0.95 |
| *PTS* | | | | | | | | | | | | | | |
| PTS ($\times$0.5) | 14.2 | 1.47 | 13.3 | 1.35 | 13.5 | 1.35 | 31.9 | 1.06 | 29.6 | 1.06 | 16.8 | 1.14 | 17.2 | 1.11 |
| PTS ($\times$1.5) | 25.6 | 0.16 | 21.4 | 0.13 | 21.7 | 0.13 | 26.0 | 0.27 | 25.6 | 0.29 | 21.9 | 0.35 | 20.7 | 0.33 |
| PTS ($\times$2.0) | 27.5 | 0.11 | 21.7 | 0.09 | 22.8 | 0.13 | 27.8 | 0.20 | 25.2 | 0.25 | 21.0 | 0.29 | 21.7 | 0.25 |

Table 14: Full per-strategy rule-faithfulness on Private-Biology. Each model column reports **FR** (floor rate, % predicted 0) and $\mathbf{R}_{ir}$ (in-rubric RSR; 1.0=faithful).

### I.2 Full Intervention Results

Table 14 gives the full breakdown. **LoR edits: all graders are near-faithful.** In-rubric RSR stays in $[0.78, 1.14]$ and floor rates are low (6–12%); the model tracks band shifts as intended. The EDIT family is the most faithful trained family here (in-rubric RSR 0.86–0.96, closest to 1) and floors the least (6.4–7.8%), whereas `Base` mildly over-reacts (1.13–1.14 on A2/B1). **PTS rescales: unfaithful for everyone.** Multiplicative point rescaling collapses in-rubric RSR to 0.1–0.35 on the up-rescales ($\times$1.5, $\times$2.0) and inflates floor rates to 21–32%, evidence of absolute-count anchoring that no method removes. `InT` floors hardest (26–32%); EDIT-SFT retains the highest PTS in-rubric RSR among the trained arms (0.25–0.35 on the up-rescales). The decomposition also exposes a confound in the headline RSR: `InT`'s competitive aggregate RSR on PTS is partly a flooring artefact (its high floor rate removes responsive rows), which the floor-rate / in-rubric split makes explicit. Signed averages can likewise flatter `Base`, whose over-reactions on LoR and under-reactions on PTS partly cancel. The AE-RSR aggregation in Table 7 removes this cancellation.

## J Prompt

We document the core prompt behind EDIT's Phase-1 atomic rewriting under locality constraint (shown in Figure 6). For space we show only its structural skeleton: the section headers, the per-instance privileged inputs (rendered as `{·}` placeholders), and the output contract; the verbose per-section guidance is elided (`[...]`).

```
SYSTEM: Refine a wrong grading attempt by fixing exactly ONE substep.
        The rest of the chain regenerates after the rewrite.

# Grading task
  {PROBLEM}

# Oracle reference
  correct mark = {GOLD} / {TOTAL}
  [oracle only; never state the mark digit or its spelled-out word]

# The full grading attempt
  {NUMBERED_ATTEMPT}

# Target for rewriting: Substep {i}
  "{TARGET_STEP_TEXT}"

# Diagnostic hints for Substep {i}
  {DIAGNOSTIC_HINT_BULLETS}
  [from the internal-state locator (C1); a checklist to CONSIDER, not satisfy]

# Per-criterion rubric analysis (REFERENCE)
  {RCI_ANALYSIS}
  [privileged input (C2); informs reasoning, never copied into the output]

# Atomicity + locality
  [serve the SAME local role as the target step; emit multiple atomic
   substeps only when the prefix has already synthesised criteria]

# Style
  [one claim per step; match the chain's voice; no "(1)(2)" markers,
   no covered/partial verdict labels, no enumeration]

# Output format
  <rewrite location="{i}" span="1|2|3">
  step 1: <first step text>
  step 2: <optional>
  </rewrite>
```

Figure 6: EDIT Phase-1 *Rule-Aware Atomic Rewriter* prompt (structural skeleton; per-section guidance elided as `[...]`). Braced tokens `{·}` are filled per instance. The diagnostic hints come from the internal-state locator (**C1**) and the per-criterion rubric analysis is the privileged input (**C2**); neither the gold mark nor the rubric block's structure may appear in the rewrite.