RITE: INVESTIGATING AND MITIGATING RUBRIC DRIFT IN LLM-BASED GRADING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models are increasingly used to grade open-ended student responses, yet construct-irrelevant variation can change their judgments. We call this failure *rubric drift*: grading changes despite fixed questions, rubrics, and rubric-relevant evidence. We introduce **EduR^{2}Bench**, with 1,060 verified rubric-equivalence groups and 3,297 variants across four STEM subjects, and Criterion Flip Rate (CFR), a label-free drift diagnostic. Rubric drift is widespread across open-weight, frontier, and specialized graders and is distinct from grading accuracy. Our central finding is **supervision-induced invariance (SII)**: correctness supervision on original responses reduces CFR on unsupervised equivalent variants from 26.3% to 11.7%, accompanied by contraction of upper-middle-layer representation gaps. However, SII is limited by supervision coverage, label reliability, and transformation-family transfer. These boundaries motivate **Rubric-Invariant Training via Equivalence (RITE)**, which combines correctness grounding with equivalence supervision. Using labels only on original responses, RITE reduces CFR to 5.2% while preserving grading quality and retains substantial benefit on a held-out transformation family and unseen human-written rewrites. Together, our results characterize when rubric invariance emerges from supervision and how to strengthen it explicitly.

1 INTRODUCTION

Large language models (LLMs) are increasingly used to grade open-ended student responses by following natural-language rubrics and producing scores and feedback (Xie et al., 2025). Reliable grading, however, requires more than high average accuracy. Educational measurement treats construct-irrelevant variance as a threat to validity (Messick, 1995; Eignor, 2013): judgments should depend on evidence relevant to the intended construct, rather than on irrelevant differences in expression (Zhai et al., 2020; 2021). As illustrated in Fig. 1, LLM graders can change their judgments when content order, length, style, or irrelevant information changes, even though the question, rubric, and rubric-relevant evidence remain fixed. We call this failure *rubric drift*: systematic changes in grading judgments induced by construct-irrelevant variation beyond random sampling noise. Such drift can lead graders to assign different grades to answers demonstrating the same rubric-relevant knowledge. Prior work has documented biases and instability in LLM evaluation (Zheng et al., 2023; Wang et al., 2024; Ye et al., 2025; Stureborg et al., 2024; Haldar & Hockenmaier, 2025) and studied calibration, specialized evaluators, evidence grounding, and perturbation consistency (Hashemi et al., 2024; Kim et al., 2024; Hong et al., 2026; Ding et al., 2026; Hejabi et al., 2026). However, these studies do not directly ask: *when rubric-relevant evidence is unchanged, does the grader preserve its judgment?* Existing benchmarks rarely provide verified equivalence structures for answering this question. We therefore introduce **EduR^{2}Bench**, containing 1,060 rubric-equivalence groups and 3,297 verified construct-irrelevant variants across four STEM subjects, constructed from Mohler (Mohler et al., 2011) and SciEntsBank (Dzikovska et al., 2013). We further introduce *Criterion Flip Rate* (CFR), a label-free criterion-level diagnostic. Across open-weight, frontier, and specialized graders, rubric drift is widespread and is not captured by grading accuracy.

Surprisingly, explicit invariance supervision is not necessary for invariance to emerge. Supervising only original responses for correctness reduces CFR on equivalent variants from 26.3% to 11.7% on Computer Science, although these variants receive neither direct correctness supervision nor an explicit consistency objective. We call this phenomenon **supervision-induced invariance (SII)**.

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.

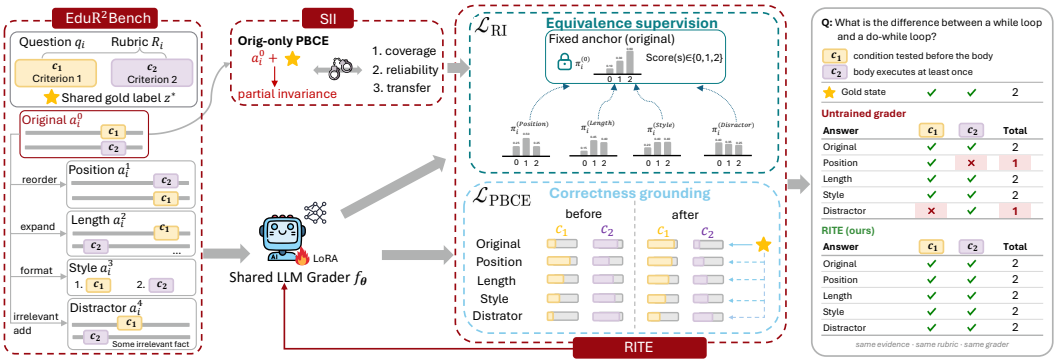

Figure 1: **Overview. Left:** EduR^{2}Bench groups original answer a_i^0 with construct-irrelevant variants from four transformation families, sharing the question, rubric, rubric-relevant evidence, and gold label z^*. **Middle:** supervising original answers (Orig-only PBCE) induces partial invariance on the unsupervised variants (SII), bounded by coverage, reliability, and transfer; RITE fine-tunes a shared grader f_θ with correctness grounding ($\mathcal{L}_{\mathrm{PBCE}}$) toward the gold state and label-free equivalence supervision ($\mathcal{L}_{\mathrm{RI}}$) that aligns each variant's score distribution $\pi_i^{(m)}$ to the stop-gradient anchor $\pi_i^{(0)}$. **Right:** the untrained grader ($T{=}0$) flips different criteria under Position and Distractor edits (rubric drift), whereas RITE returns identical judgments for the same evidence, rubric, and grader.

This effect has a representation-level signature: correctness supervision selectively contracts upper-middle-layer representation gaps associated with grading flips. Thus, supervision specifying *what judgment is correct* can induce partial generalization about *which input differences should not change it*. SII, however, is support-dependent rather than guaranteed. It strengthens with broader supervision coverage but degrades under label noise and transfer to certain unseen transformation families, particularly irrelevant content additions. Theoretically, correctness supervision identifies the grading target on supervised inputs but imposes no objective-level constraint on distinct unsupervised inputs under an unconstrained predictor class. SII is therefore learned generalization rather than objective-guaranteed invariance. This exposes the missing signal: correctness specifies *how an answer should be graded*, but not *which answers should be graded equivalently*.

Motivated by this distinction, we propose **Rubric-Invariant Training via Equivalence (RITE)**, combining correctness grounding with explicit equivalence supervision. Correctness provides an absolute grading target, while equivalence supervision aligns score distributions of verified rubric-equivalent responses without requiring variant correctness labels. Our grounding–invariance decomposition formalizes their complementary roles and bounds variant score-distribution error by grounding and equivalence terms. Using labels only on original responses, RITE reduces CFR from 11.7% to 5.2% on Computer Science while preserving grading quality, performs strongly across four subjects and two open-weight backbones, and retains substantial benefit on a held-out transformation family and unseen human-written rewrites. Our main contributions are:

- **Understanding Rubric Invariance.** We identify SII and characterize its representation-level signature and boundaries in supervision coverage, label reliability, and transformation-family transfer.
- **Measuring Rubric Drift.** We formalize rubric drift and introduce EduR2Bench with verified rubric-equivalence groups and CFR as a label-free criterion-level diagnostic.
- **Mitigating Rubric Drift.** Motivated by SII's boundaries, we propose RITE, combining correctness grounding with equivalence supervision to reduce drift while preserving grading quality and transferring beyond observed transformation families.

2 Methodology

In this section, we first formally state the rubric drift problem that we aim to address. We then introduce our proposed framework and theoretical analysis.

Problem Formulation. We first formulate rubric-based grading over the training set $\mathcal{D}_{\text{tr}}=\{(q_i, \mathcal{R}_i, \mathcal{A}_i, \mathbf{z}_i^*)\}_{i=1}^N$. Each instance contains a question q_i, a rubric $\mathcal{R}_i=\{c_{i,k}\}_{k=1}^{K_i}$ with K_i binary criteria, and an indexed answer collection $\mathcal{A}_i=\{a_i^{(m)}\}_{m=0}^{M_i-1}$ with $M_i \geq 2$. Here, $a_i^{(0)}$ is the original answer and $a_i^{(m)}$, $m \geq 1$, is a rubric-equivalent variant that may differ in construct-irrelevant properties such as order, length, style, or irrelevant information. All answers in $\mathcal{A}_i$ preserve the same rubric-relevant evidence and therefore share the ground-truth grading state $\mathbf{z}_i^*=(z_{i,1}^*, \ldots, z_{i,K_i}^*) \in \{0,1\}^{K_i}$ and total score $s_i^*=\sum_{k=1}^{K_i} z_{i,k}^*$.

Given $\mathbf{x}_i^{(m)}=(q_i, \mathcal{R}_i, a_i^{(m)})$, a grader produces criterion-level satisfaction probabilities

$$\mathbf{p}_i^{(m)}=f_\theta(\mathbf{x}_i^{(m)})=(p_{i,1}^{(m)}, \ldots, p_{i,K_i}^{(m)}) \in (0,1)^{K_i}.$$

The corresponding decisions are $\hat{z}_{i,k}^{(m)}=\mathbb{I}[p_{i,k}^{(m)} \geq 1/2]$ and the total score is $\hat{s}_i^{(m)}=\sum_{k=1}^{K_i} \hat{z}_{i,k}^{(m)}$.

Our **objective** is to learn an accurate and rubric-invariant grader. Grading accuracy is assessed through agreement with the ground-truth criterion labels and total score. We distinguish two levels of invariance. A grader is *criterion-level invariant* on group i if $\hat{z}_{i,k}^{(m)}=\hat{z}_{i,k}^{(0)}$ for all k and m, and *score-level invariant* if $\hat{s}_i^{(m)}=\hat{s}_i^{(0)}$ for all $m \geq 1$. The former implies the latter but not conversely: criterion vectors $(1,0)$ and $(0,1)$ yield the same total score while flipping both criteria. Score-level invariance governs the fairness of the grade; criterion-level invariance additionally governs the fidelity of criterion-level feedback. RITE optimizes score-level invariance of the induced score distributions, for which Sec. 2 gives an objective-level guarantee; invariance of the emitted decisions is assessed empirically at both levels, using the stricter criterion-level diagnostic (CFR) as the primary metric. We define *rubric drift* as a change in predicted score or criterion decision caused by construct-irrelevant variation under a fixed question, rubric, rubric-relevant evidence, and grader.

Let $\mathcal{S}_i \subseteq \{0, \ldots, M_i-1\}$ denote the indices receiving direct correctness supervision. In standard RITE, $0 \in \mathcal{S}_i$, so the original answer provides an absolute grading target. Because correctness supervision does not explicitly encode equivalence among answers, RITE combines it with equivalence supervision to promote score-level invariance. At inference time, each answer is graded independently, without access to other group members or ground-truth labels.

Overview of RITE. As shown in Fig. 1 (middle), RITE jointly optimizes correctness grounding and score-level rubric invariance within each rubric-equivalence group. $\mathcal{L}_{\text{PBCE}}$ supervises answers indexed by $\mathcal{S}_i$ toward the gold grading state, while $\mathcal{L}_{\text{RI}}$ aligns each variant's total-score distribution with the original's, without requiring correctness labels for variants. Thus, correctness supervision identifies the grading target, while equivalence supervision prevents construct-irrelevant variation from altering the score. We next define both losses and their joint objective.

Direct Correctness Supervision. Given an input $\mathbf{x}_i^{(m)}$, we model the satisfaction of criterion $c_{i,k}$ as a conditional Bernoulli variable with parameter $p_{i,k}^{(m)}$, where $\mathbf{p}_i^{(m)}=f_\theta(\mathbf{x}_i^{(m)})=(p_{i,1}^{(m)}, \ldots, p_{i,K_i}^{(m)}) \in (0,1)^{K_i}$. In practice, $p_{i,k}^{(m)}$ is obtained by applying a softmax to the grader's `hit` and `miss` logits at the decision position for criterion $c_{i,k}$. We use the conditionally factorized likelihood $p_\theta(\mathbf{z}_i^* \mid \mathbf{x}_i^{(m)})=\prod_{k=1}^{K_i}(p_{i,k}^{(m)})^{z_{i,k}^*}(1-p_{i,k}^{(m)})^{1-z_{i,k}^*}$, whose normalized negative log-likelihood gives the per-answer correctness loss as:

$$\ell_i^{(m)}=-\frac{1}{K_i}\sum_{k=1}^{K_i}\left[z_{i,k}^* \log p_{i,k}^{(m)}+(1-z_{i,k}^*)\log(1-p_{i,k}^{(m)})\right].$$

Given the pre-defined indices of answers receiving direct correctness supervision $\mathcal{S}_i$, we define the per-criterion binary cross-entropy objective as:

$$\mathcal{L}_{\text{PBCE}}=\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\mathcal{S}_i|}\sum_{m \in \mathcal{S}_i}\ell_i^{(m)}. \tag{1}$$

Because all answers in $\mathcal{A}_i$ share the same ground-truth grading state $\mathbf{z}_i^*$ and total score s_i^*, the same likelihood applies to any supervised member of the instance. Under original-only supervision, $\mathcal{S}_i=\{0\}$; under full supervision, $\mathcal{S}_i=\{0, \ldots, M_i-1\}$. Thus, $\mathcal{L}_{\text{PBCE}}$ provides criterion-level correctness grounding and anchors the grader to the correct total score within the supervision scope.

Explicit Equivalence Supervision. The conditionally factorized Bernoulli model induces a distribution over the total score. Specifically, the criterion-level satisfaction probabilities for answer $a_i^{(m)}$ define a Poisson–Binomial distribution, denoted by $\mathrm{PB}(\cdot)$, with probability mass vector

$$\boldsymbol{\pi}_i^{(m)} = \mathrm{PB}\left(p_{i,1}^{(m)}, \ldots, p_{i,K_i}^{(m)}\right) = (\pi_{i,0}^{(m)}, \ldots, \pi_{i,K_i}^{(m)}),$$

where $\pi_{i,r}^{(m)}$ denotes the probability of assigning total score $r \in \{0, \ldots, K_i\}$. This induced distribution provides a probabilistic representation of the grader's final scoring outcome.

Because all answers in $\mathcal{A}_i$ are rubric-equivalent and share the same ground-truth score, their induced score distributions should remain invariant to construct-irrelevant variation. We use the original answer $a_i^{(0)}$ as the anchor and align each variant distribution $\boldsymbol{\pi}_i^{(m)}$ with the anchor distribution $\boldsymbol{\pi}_i^{(0)}$. During optimization, the anchor distribution is detached from gradient computation and treated as a fixed alignment target. Using the symmetric and bounded Jensen–Shannon divergence D_{JS}, we define the rubric-invariance objective as

$$\mathcal{L}_{\mathrm{RI}} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{M_i - 1} \sum_{m=1}^{M_i - 1} D_{\mathrm{JS}}\left(\boldsymbol{\pi}_i^{(0)} \parallel \boldsymbol{\pi}_i^{(m)}\right). \tag{2}$$

Unlike direct correctness supervision, $\mathcal{L}_{\mathrm{RI}}$ leverages the verified equivalence relation among answers, requiring no direct correctness labels for the variants, and encourages rubric-equivalent answers to receive consistent score distributions. However, equivalence supervision alone cannot identify which invariant score is correct, motivating its combination with correctness grounding.

Joint Objective. We combine the criterion-level correctness likelihood with score-distribution alignment to jointly optimize grading accuracy and score-level rubric invariance:

$$\mathcal{L}_{\mathrm{RITE}} = \mathcal{L}_{\mathrm{PBCE}} + \lambda_{\mathrm{RI}} \mathcal{L}_{\mathrm{RI}}, \tag{3}$$

where $\lambda_{\mathrm{RI}} \geq 0$ controls the strength of equivalence supervision. The two terms provide complementary statistical signals: $\mathcal{L}_{\mathrm{PBCE}}$ grounds supervised answers to the correct grading state and total score, whereas $\mathcal{L}_{\mathrm{RI}}$ propagates score-level consistency from the grounded anchor to its rubric-equivalent variants. Under original-only supervision, $\mathcal{S}_i = \{0\}$, RITE therefore requires direct correctness labels only for the original answer while using the equivalence structure to regularize the unlabeled variants. During training, all answers in each sampled instance are processed jointly because $\mathcal{L}_{\mathrm{RI}}$ couples their induced score distributions. We align total-score distributions rather than criterion vectors because score-level alignment is a weaker constraint that proved more robust under sparse or noisy supervision: aligning each criterion's Bernoulli separately yields 6.1% CFR versus 4.8% for score-level alignment under original-only supervision on the validation set (Tab. 11). The pseudocode for the training procedure is provided in Appendix A.2.

Theoretical Analysis. We characterize the objective-level roles of correctness and equivalence supervision under an unconstrained predictor class, without assumptions on optimization or generalization. We work on the closure $[0,1]^{K_i}$ of the grader's probability space and adopt the extended-value convention for binary cross-entropy at the boundary. Let $\boldsymbol{\pi}_i^* = \delta_{s_i^*}$ denote the point mass on the ground-truth score, $d_{\mathrm{TV}}(\boldsymbol{\pi}, \boldsymbol{\pi}') = \frac{1}{2}\|\boldsymbol{\pi} - \boldsymbol{\pi}'\|_1$, and $\mathfrak{P}_i = \{\mathrm{PB}(\mathbf{p}) : \mathbf{p} \in [0,1]^{K_i}\}$ denote the family of realizable Poisson–Binomial score distributions. D_{JS} uses natural logarithms. Proofs are in Appendix A.3.

Theorem 1 (Grounding and equivalence are not interchangeable)**.** *For any $m \in \mathcal{S}_i$, $\ell_i^{(m)}$ is uniquely minimized over $[0,1]^{K_i}$ at $\mathbf{p}_i^{(m)} = \mathbf{z}_i^*$. By contrast, for any $\boldsymbol{\nu} \in \mathfrak{P}_i$, predictions satisfying $\boldsymbol{\pi}_i^{(0)} = \cdots = \boldsymbol{\pi}_i^{(M_i-1)} = \boldsymbol{\nu}$ yield zero equivalence loss for instance i, regardless of whether $\boldsymbol{\nu} = \boldsymbol{\pi}_i^*$. Thus, equivalence supervision alone does not identify the ground-truth score distribution.*

Correctness supervision identifies the absolute grading target, whereas equivalence supervision specifies which predictions should remain invariant. In standard RITE, $0 \in \mathcal{S}_i$, so the original answer provides a grounded anchor.

Theorem 2 (Grounding–invariance decomposition)**.** *For any $m \in \{1, \ldots, M_i - 1\}$,*

$$d_{\mathrm{TV}}\left(\boldsymbol{\pi}_i^{(m)}, \boldsymbol{\pi}_i^*\right) \leq \sqrt{2 D_{\mathrm{JS}}\left(\boldsymbol{\pi}_i^{(0)} \parallel \boldsymbol{\pi}_i^{(m)}\right)} + K_i \ell_i^{(0)}. \tag{4}$$

The proof combines the triangle inequality, a Jensen–Shannon-to-total-variation bound, and a Poisson–Binomial coupling argument. Eq. (4) decomposes variant error into an invariance term controlled by $\mathcal{L}_{\mathrm{RI}}$ and a grounding term controlled by $\mathcal{L}_{\mathrm{PBCE}}$ when $0 \in \mathcal{S}_i$. If both vanish, then $\boldsymbol{\pi}_i^{(m)} = \boldsymbol{\pi}_i^*$.

Proposition 1 (No guarantee beyond the supervision scope)**.** *Over an unconstrained predictor class, the prediction on any distinct unsupervised input $\mathbf{x}_i^{(m)}$ with $m \notin \mathcal{S}_i$ can be changed while all supervised predictions remain fixed, without changing $\mathcal{L}_{\mathrm{PBCE}}$. Thus, correctness supervision alone imposes no objective-level constraint outside its supervision scope.*

Consequently, invariance induced on unsupervised answers by correctness supervision is learned generalization through shared parameters, rather than a direct guarantee of the objective.

Proposition 2 (Score-level invariance does not identify criterion-level predictions)**.** *Because $\mathrm{PB}(\cdot)$ is invariant to permutations of its parameters, distinct criterion-level probability vectors can induce identical total-score distributions. Therefore, perfect score-distribution alignment does not uniquely identify a shared criterion-level prediction pattern.*

Prop. 2 limits RITE's guarantee to score-level rubric invariance. On directly supervised answers, $\mathcal{L}_{\mathrm{PBCE}}$ removes this ambiguity by grounding each criterion probability to its gold label. These results do not establish finite-sample generalization, optimization convergence, or invariance to unseen transformation families; the absence of a general transfer guarantee is formalized in Prop. 4 (Appendix A.3). Empirically, RITE lowers criterion-level CFR under original-only supervision, the regime it is designed for (Tab. 4), and retains most of its benefit on a held-out family (RQ4).

3 Experiments

We introduce EduR^{2}Bench and the evaluation protocol, then present results organized around four research questions; details beyond those given here are in Appendix A.4.

EduR^{2}Bench. EduR^{2}Bench is built from authentic short-answer responses in the Mohler dataset (Mohler et al., 2011) and SciEntsBank (Dzikovska et al., 2013), covering Computer Science (CS), Physics, Chemistry, and Earth Science. Each reference answer is decomposed into $K_i \in \{1, \ldots, 4\}$ independently verifiable binary criteria, drafted by `Fable 5` and reviewed by two authors. The unit of analysis is a rubric-equivalence group: an original answer plus variants from four construct-irrelevant transformation families (Position, Length, Style, Distractor). Position and Style variants are produced by deterministic rewriting scripts; Length and Distractor variants are generated by `Fable 5` under a fixed written protocol. Every variant passes an automatic verifier and then manual review for (i) correct application of the transformation, (ii) preservation of all rubric-relevant evidence, and (iii) an unchanged criterion-level gold state; candidates failing any check are regenerated. Gold states are initialized by `Fable 5` and manually reviewed: on a stratified sample of 500 criterion decisions, three annotators agree on 96.2% (Fleiss' κ=0.91), and `Fable 5` agrees with teaching-assistant labels (Sonkar et al., 2024) in 95.0% of decisions. The benchmark contains 1,060 groups and 3,297 variants from 176 questions, split at the question level to prevent leakage (680/124/256 groups for train/val/test; Tab. 6). Prompts, the verification checklist and statistics (Appendix Tab. 20), and a complete example group are given in Appendix A.4.

Measurement Protocol. We report grading quality and within-group stability separately. Quadratic Weighted Kappa (QWK, ↑) measures total-score agreement, computed within rubric-size buckets and then size-weighted; criterion-level Macro-F1 (C-F1, ↑) measures hit/miss classification quality; Score-Error Standard Deviation (Err. SD, ↓) measures the variability of total-score errors. Our primary label-free diagnostic is Criterion Flip Rate (CFR, ↓): among group–criterion pairs with predictions for at least two members, the fraction whose predicted hit/miss decision differs across members. Score Flip Rate (SFR, ↓) is the analogous fraction of groups whose predicted total score is not identical across members; CFR is the stricter counterpart, since unchanged total scores can hide criterion-level flips. Both measure consistency rather than error, so they are always reported alongside gold-referenced metrics. All graders use greedy decoding (T=0), so a flip is a change in the grade a student would actually receive rather than sampling noise; CFR is robust to the decision threshold ($\leq$3.7 points across 0.3–0.7) and to group-size normalization, which preserves the ordering of all methods (Appendix A.5, Tab. 16). For family-level analyses (Fig. 2c, Tabs. 2 and 15), *family-specific CFR* counts a flip only between the original answer and the variant of that family, attributing drift to a single transformation.

Representation Analysis Protocol. To relate behavioral drift to internal representations, we extract the hidden state $\mathbf{h}^{(m)}_{i,k,l}$ at layer l from the decision position for criterion $c_{i,k}$ and define the gap to the original answer as $g^{(m)}_{i,k,l}=1-\cos(\mathbf{h}^{(m)}_{i,k,l}, \mathbf{h}^{(0)}_{i,k,l})$, averaged over valid variants per group–criterion pair and compared between flip and stable cases across layers (peak band chosen on the validation set without flip labels; Appendix A.4). Larger gaps among flip cases are read as a representation-level signature of drift, not as evidence of causality.

Experimental Setup. We study three settings: supervision coverage (varying $\mathcal{S}_i$), reliability (criterion-level noise at rate ρ), and transfer (holding out one transformation family). Compute-heavy analyses use the CS subset; the main comparison spans all four subjects, with `Qwen3-8B` (Yang et al., 2025) as the primary backbone and `Llama-3.1-8B-Instruct` (Grattafiori et al., 2024) replicating the main findings. All trainable methods share backbone, split, LoRA setup, and training budget, with RITE fixing Jensen–Shannon divergence at $\lambda_{\mathrm{RI}}=1$ throughout. Hyperparameters are frozen after validation-set selection under matched search budgets (Appendix Tab. 8); we report mean±std over three seeds unless noted, with full settings in Appendix A.4.

Competing Methods. We compare RITE with three reference graders sharing its backbone—the untrained backbone with the shared checklist prompt (BASE; the Checklist row of Tab. 1b), direct supervision on original answers only (PBCE-ORIG), and on all members of each equivalence group (PBCE-ALL)—and with eight consistency and robustness baselines: UDA (Xie et al., 2020), R-Drop (Wu et al., 2021), FixMatch (Sohn et al., 2020), SMART (Jiang et al., 2020), Mean Teacher (Tarvainen & Valpola, 2017), PPCL (Qiang et al., 2024), GroupDRO (Sagawa et al., 2019), and Consistency Alignment (CA) (Zhao et al., 2024) (Appendix A.4).

Results. We ask whether LLM graders drift across rubric-equivalent responses (**RQ1**), whether correctness supervision alone induces rubric invariance (**RQ2**), where that induced invariance breaks down (**RQ3**), and whether RITE mitigates these failures while preserving grading quality (**RQ4**).

Table 1: a Grading quality (QWK) and rubric drift (CFR) for six off-the-shelf graders (open-weight, frontier, specialized) on the CS test set. All exhibit drift; higher QWK does not imply lower CFR. b Comparison on six inference-time strategies for `Qwen3-8B` on the CS test set.

(a)

Grader	QWK (↑)	CFR (%, ↓)
`Qwen3-8B`	0.325	26.3
`Llama-3.1-8B-Instruct`	0.400	37.5
`gpt-5.2`	0.641	15.0
`kimi-k2`	0.553	31.3
`claude-sonnet-5`	0.619	11.9
`Prometheus 2`	0.137	23.1

(b)

Inference strategy	QWK (↑)	CFR (%, ↓)
Direct judgment	0.385	25.0
Checklist	0.325	26.3
Chain-of-thought	0.289	18.8
Evidence-first	0.387	33.8
Invariance-aware prompt	0.316	23.1
Self-consistency	0.328	25.6

Rubric drift (RQ1). All six off-the-shelf graders exhibit rubric drift on the CS test set (Tab. 1a), including frontier models (`Kimi K2` 31.3% (Team et al., 2025); `gpt-5.2` 15.0%; `claude-sonnet-5` 11.9%, and 11.3–23.6% across the four subjects in Tab. 3) and the evaluation-specialized `Prometheus 2` (23.1%) (Kim et al., 2024). Drift is not an accuracy problem: `Llama-3.1-8B-Instruct` achieves higher QWK than `Qwen3-8B` (0.400 vs. 0.325) but also higher CFR (37.5% vs. 26.3%). Prompting does not fix it (Tab. 1b): chain-of-thought lowers CFR to 18.8% at the cost of QWK (0.289), an explicit invariance instruction still leaves 23.1%, and evidence-first prompting (Hong et al., 2026) raises it to 33.8%. Grading quality and rubric invariance are distinct dimensions of grader reliability, and neither scale nor prompting reliably improves both.

Supervision-Induced Rubric Invariance (RQ2). Because prompting does not remove drift, we turn to training and ask whether the weakest form of supervision, correctness labels on the original response alone, induces invariance on responses it never touches. On CS, PBCE-Orig reduces CFR on unseen test variants from 26.3% to 11.7% while raising QWK from 0.325 to 0.737, and CFR falls further as more group members are supervised before plateauing (Fig. 2a). Because these variants receive neither labels nor a consistency objective, we call this *supervision-induced invariance* (SII); it holds in all four subjects (Tab. 3, PBCE-Orig vs. Base). Fig. 3 shows the representational counterpart:

in the untrained grader, flipping criteria have a larger original–variant gap that peaks at layers 22–25, and original-only supervision contracts it (0.038→0.023) more for flipping than for stable criteria.

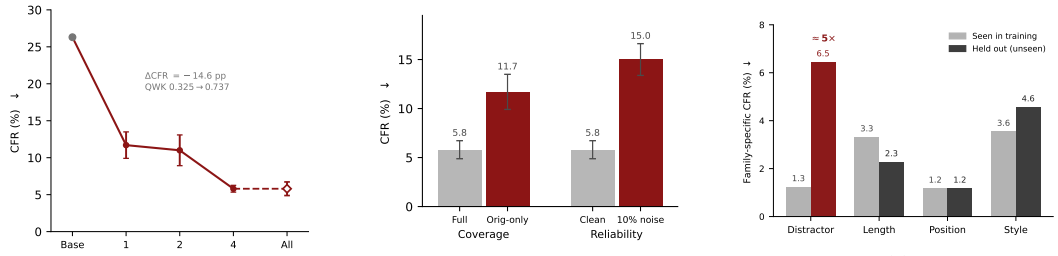

Figure 2: **SII and its boundaries** (CS, `Qwen3-8B`, 3 seeds, ±1 s.d.). a Test-set CFR vs. number of supervised group members (grey: untrained; diamond: fully supervised). b Coverage and label-noise boundaries relative to the fully supervised, clean-label reference (grey). c Leave-one-family-out PBCE-All training: only holding out Distractor raises its family-specific CFR (∼5×).

Boundaries of SII (RQ3). We vary coverage, label reliability, and family exposure with all else fixed (Figs. 2b and 2c). *Coverage:* relative to full supervision, original-only supervision doubles CFR (11.7% vs. 5.8%). *Reliability:* under full coverage, 10% label noise raises CFR from 5.8% to 15.0% and lowers QWK from 0.746 to 0.602. *Transfer:* holding out Length, Position, or Style costs nothing, whereas CFR on unseen Distractor variants is ∼5× higher than when the family is seen. In both failure modes the flip–stable representation gap re-widens, reversing the contraction in RQ2. SII is therefore local to the support of reliable supervision, as Prop. 1 anticipates: correctness labels say how a response should be graded, not which responses should be graded alike, and invariance beyond their scope is inherited through shared parameters rather than enforced by the objective. The three surface-form families perturb properties the pretrained grader is already partly invariant to, whereas ignoring added content is a distinct skill (Base is also least stable on Distractor; Appendix Tab. 15). RITE supplies the missing relational term; we test it first at this boundary, then against baselines across all subjects (RQ4).

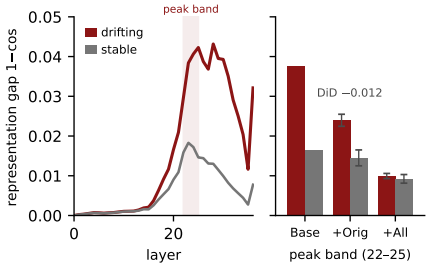

Figure 3: Representational signature of SII (CS). Left: layer-wise original–variant gap for drifting vs. stable criteria (untrained grader). Right: mean gap in the peak band (layers 22–25) as supervision widens; drifting criteria contract more (DiD −0.012, 95% CI $[-0.019, -0.005]$). ±1 s.d. (Orig-only: 3 seeds).

Transfer to a Held-out Family. Repeating the leave-one-family-out protocol with RITE (Tab. 2; all families in Appendix Tab. 17), held-out Distractor CFR drops from 6.5% (PBCE-All) to 3.1% (RITE-All) and from 10.6% (PBCE-Orig) to 3.8% (RITE-Orig), lower in 3/3 seeds, with QWK within 0.015 of each seen counterpart (Appendix Tab. 18). Relative to graders that saw Distractor in training (1.5%/1.9%), RITE recovers roughly two-thirds (All) and four-fifths (Orig) of the gap without a single distractor; the other three families show no gap. Transfer is thus substantial but incomplete, consistent with Prop. 4: equivalence supervision propagates the align-to-anchor bias across families, while covering the target family removes the residual. The same ordering holds on 251 human-written rewrites of the test originals that no grader saw in training, with RITE's CFR below the fully supervised grader's (Appendix Tab. 19).

Table 2: **Held-out Distractor** (CS, 3 seeds). Family-specific CFR (%) with Distractor seen vs. held out in training. †Uses no variants, hence unaffected.

Grader	seen	held out
PBCE-All	1.3 ± 0.6	6.5 ± 1.0
RITE-All	1.5 ± 0.4	$\mathbf{3.1 \pm 0.6}$
PBCE-Orig	10.6 ± 0.6	$10.6 \pm 0.6^{\dagger}$
RITE-Orig	1.9 ± 0.6	$\mathbf{3.8 \pm 0.6}$

RITE Mitigates Rubric Drift (RQ4). Tab. 3 compares RITE with eight consistency and robustness baselines on all four subjects under a shared backbone and training budget, with each baseline's

hyperparameters tuned on the validation set (Appendix Tab. 8). R-Drop and SMART, which regularize against dropout or adversarial perturbations rather than the verified equivalence structure, give little or no gain over PBCE-Orig; UDA, FixMatch, Mean Teacher, and PPCL, which use the variants as unlabeled inputs, reduce drift on most subjects but give up QWK on at least one; GroupDRO and CA trail RITE even with full-group labels. RITE is best or tied-best among trained graders on every metric in every subject, and the score-level flip rate follows the same ordering (Appendix Tab. 16). Earth Science has the highest CFR for every trained and frontier grader (CFR $\geq$ 19.8%), and it is where RITE's margin is smallest (13.4% vs. 16.2% for PBCE-All), indicating this drift is not specific to RITE. With `Llama-3.1-8B-Instruct`, CS CFR falls from 37.5% (Base) to 15.2% (PBCE-Orig) and 5.8% (RITE) with the best QWK (Appendix Tab. 21). Against the zero-shot frontier graders, RITE attains higher QWK on CS, Physics, and Chemistry ($+0.06$ to $+0.11$ over the better of the two) and is within 0.02 on Earth Science, has higher C-F1 and lower Err. SD on all four subjects, and 1.5–2.3$\times$ lower CFR than the better frontier grader on every subject: an 8B grader with equivalence supervision is more invariant than scale alone.

Table 3: Baseline comparison across four subjects of EduR^{2}Bench (QWK and CFR: mean $\pm$ s.d. over three seeds; C-F1 and Err. SD: three-seed means). Base is a single greedy run; `gpt-5.2` and `claude-sonnet-5` are zero-shot with the same checklist prompt. Trained methods receive correctness labels only for the original answer ($\mathcal{S}_i$=orig) and use the variants as unlabeled group members, except PBCE-All, GroupDRO, and CA (full-group labels, $\mathcal{S}_i$=all) and PBCE-Orig, R-Drop, and SMART (original-answer-only, never seeing variants). Best values per column in **bold**.

Method	CS				Physics				Chemistry				Earth Science			
	QWK↑	C-F1↑	Err. SD↓	CFR (%)↓	QWK↑	C-F1↑	Err. SD↓	CFR (%)↓	QWK↑	C-F1↑	Err. SD↓	CFR (%)↓	QWK↑	C-F1↑	Err. SD↓	CFR (%)↓
gpt-5.2	0.64	0.80	0.51	15.0	0.73	0.79	0.70	14.2	0.74	0.89	0.55	17.4	0.72	0.80	0.91	23.6
claude-sonnet-5	0.62	0.79	0.52	11.9	0.71	0.86	0.68	11.3	0.72	0.87	0.56	14.1	0.70	0.79	0.96	19.8
Base	0.33	0.62	0.69	26.3	0.48	0.73	0.91	25.6	0.33	0.64	0.79	32.1	0.64	0.80	0.92	26.0
PBCE-Orig	0.74±0.05	0.85	0.46	11.7±1.8	0.68±0.01	0.84	0.81	19.1±0.9	0.66±0.04	0.84	0.57	16.1±2.5	0.66±0.01	0.79	0.92	21.5±4.5
PBCE-All	0.75±0.04	0.85	0.46	5.8±1.6	0.69±0.03	0.84	0.79	12.4±1.2	0.71±0.03	0.86	0.53	12.8±0.9	0.63±0.03	0.79	0.96	16.2±2.8
UDA	0.69±0.03	0.83	0.49	8.3±1.4	0.71±0.02	0.84	0.76	11.4±0.7	0.74±0.04	0.87	0.50	9.5±1.0	0.59±0.02	0.77	0.99	22.3±1.4
R-Drop	0.64±0.05	0.80	0.52	15.8±4.5	0.63±0.06	0.80	0.89	19.2±1.8	0.68±0.03	0.83	0.56	14.4±2.5	0.62±0.02	0.77	0.95	24.9±0.6
FixMatch	0.68±0.02	0.83	0.48	7.9±0.7	0.68±0.05	0.82	0.78	11.8±0.7	0.67±0.02	0.83	0.57	12.9±1.1	0.59±0.02	0.77	0.95	22.1±1.5
SMART	0.65±0.03	0.81	0.51	15.2±1.4	0.67±0.02	0.82	0.84	18.4±1.2	0.67±0.02	0.83	0.55	18.9±3.0	0.63±0.04	0.79	0.90	21.4±0.8
Mean Teacher	0.70±0.02	0.84	0.47	8.3±0.7	0.65±0.02	0.82	0.84	14.2±2.0	0.76±0.02	0.89	0.48	8.6±2.0	0.58±0.07	0.76	0.95	20.6±2.0
PPCL	0.67±0.02	0.81	0.52	11.3±1.7	0.73±0.03	0.85	0.73	10.4±1.1	0.70±0.02	0.84	0.57	10.5±1.6	0.61±0.02	0.79	0.91	18.8±0.8
GroupDRO	0.68±0.02	0.82	0.49	9.4±1.3	0.68±0.05	0.82	0.80	11.0±1.1	0.67±0.02	0.82	0.58	11.6±1.7	0.59±0.02	0.78	0.94	19.4±6.5
CA	0.66±0.04	0.82	0.50	9.8±1.8	0.66±0.02	0.82	0.84	14.7±1.1	0.73±0.02	0.87	0.51	11.8±1.5	0.64±0.02	0.79	0.91	16.9±2.1
RITE (ours)	**0.75±0.03**	**0.88**	**0.41**	**5.2±0.7**	**0.79±0.02**	**0.88**	**0.66**	**7.6±0.8**	**0.82±0.01**	**0.92**	**0.43**	**6.2±1.1**	**0.70±0.02**	**0.83**	**0.86**	**13.4±1.5**

Supervision Complementarity. Tab. 4 separates the two signals. RI alone lowers CFR (26.3%→11.9%) but leaves QWK at the base level (0.316), the empirical face of Thm. 1: equivalence supervision enforces consistency without identifying the correct judgment. Under original-only labels, adding RI to PBCE-Orig cuts CFR from 11.7% to 5.2% and raises QWK (0.737→0.753), matching fully supervised PBCE-All without a single variant label. When every member is labeled, RI adds nothing and mildly over-regularizes (CFR 5.8%→7.7%, QWK 0.746→0.695; Appendix A.5), since full labels already encode the relational constraint. This is the pattern the grounding–invariance decomposition (Thm. 2) predicts: $\mathcal{L}_{\mathrm{RI}}$ controls the invariance term, which is largest when variants carry no labels, whereas $\mathcal{L}_{\mathrm{PBCE}}$ controls the grounding term, which full supervision minimizes on every member. Equivalence supervision therefore pays off exactly where correctness labels are sparse.

Table 4: **Supervision composition** (CS, `Qwen3-8B`; 3-seed means except Base). With original-only labels, adding RI matches fully supervised PBCE-All.

Correctness sup.	$\mathcal{L}_{\mathrm{RI}}$	QWK↑	CFR (%)↓
None (Base)	×	0.325	26.3
None (RI-only)	✓	0.316	11.9
Original only	×	0.737	11.7
Original only (**RITE**)	✓	**0.753**	**5.2**
Full	×	0.746	5.8
Full	✓	0.695	7.7

Sensitivity to the RI Weight. We sweep the RI weight $\lambda_{\mathrm{RI}} \in \{0, 0.1, 0.5, 1, 2, 4\}$ for `Qwen3-8B` on the EduR^{2}Bench CS validation set under original-only supervision (Tab. 5). The default λ_{RI}=1 gives the best QWK and CFR; smaller values under-regularize drift, and larger values trade grading quality for no further invariance gain.

Table 5: **RI-weight Sensitivity** (Qwen3-8B, CS validation, original-only supervision, 3 seeds).

λ_{RI}	0	0.1	0.5	1	2	4
QWK↑	0.739	0.723	0.753	**0.760**	0.742	0.697
CFR (%)↓	11.1	9.7	10.3	**4.8**	6.9	8.1

4 Related Work

LLM graders and judges. Rubric-based graders and specialized evaluators improve agreement with human judgments but score answers independently, without testing invariance across rubric-equivalent responses (Liu et al., 2023; Zhu et al., 2025; Kim et al., 2024; Hashemi et al., 2024). Judge-robustness studies expose position, format, and selection biases and measure gold-referenced degradation under perturbation (Wang et al., 2024; Park et al., 2024; Li et al., 2025; Deng et al., 2026), rather than within-group drift over verified label-preserving variants. RULERS improves a frozen grader at inference time via rubric locking, evidence grounding, and calibration (Hong et al., 2026); we compare against a prompt-only version of its evidence-anchoring stage (Tab. 1b), without its citation verification and score capping, and view the full pipeline as a training-free complement to RITE. Closest in spirit are knowledge-invariant perturbations (Li et al., 2024), paraphrase-consistency training (Elazar et al., 2021), and adversarially robust rubric evaluators (D'Souza et al., 2025); EduR^{2}Bench instead perturbs the graded answer, verifies that evidence and gold state are preserved, and diagnoses drift per criterion.

Measurement and consistency training. Educational measurement treats construct-irrelevant variance as a threat to validity; rubric drift is its input-conditional form in LLM grading, distinct from drift in human rater severity (Messick et al., 1989; Congdon & McQueen, 2000). Consistency and robust-learning methods align predictions under augmentation or stochasticity, optimize worst-group loss, or use consensus pseudo-labels (Xie et al., 2020; Sohn et al., 2020; Wu et al., 2021; Sagawa et al., 2019; Hejabi et al., 2026); RITE differs in coupling trusted criterion-level grounding ($\mathcal{L}_{\mathrm{PBCE}}$) with instance-level equivalence supervision ($\mathcal{L}_{\mathrm{RI}}$) over verified groups. A detailed discussion is in Appendix A.5.

5 Conclusion

We study *rubric drift* in LLM-based grading: construct-irrelevant changes can alter judgments even when the rubric-relevant evidence remains unchanged. Our central finding is *supervision-induced invariance (SII)*. Correctness supervision on original responses alone substantially reduces drift on unsupervised rubric-equivalent variants, revealing that learning *what judgment is correct* can also induce partial invariance to differences that should not affect that judgment. However, this effect is support-dependent: SII strengthens with broader supervision coverage but degrades under label noise and transfer to certain unseen transformation families. Its accompanying representation-level signatures further connect behavioral stability with changes in internal representations, without implying a causal mechanism. EduR^{2}Bench and CFR make this phenomenon measurable across verified rubric-equivalence groups and four STEM subjects. Our theoretical analysis complements the empirical findings by showing that correctness supervision alone provides no objective-level invariance guarantee beyond its supervision scope. This distinction motivates RITE, which combines criterion-level correctness grounding with label-free equivalence supervision. Across subjects and backbones, RITE reduces rubric drift while preserving grading quality and retains substantial benefit on a held-out transformation family and unseen human-written rewrites. Together, these results establish a distinction between two complementary forms of supervision: correctness grounding specifies *how an answer should be graded*, whereas equivalence supervision specifies *which differences should not change that grade*. More broadly, reliable LLM grading requires not only learning the correct decision rule, but also explicitly representing the invariances that the decision rule should respect.

Limitations & Future Work. First, our study focuses on short-answer grading across four STEM subjects with four controlled synthetic transformation families; a test-only set of human-written rewrites confirms the ordering (Appendix Tab. 19), but humanities, long-form responses, and essays remain open. Second, RITE transfers to a family absent from training but not completely, recovering roughly two-thirds to four-fifths of the held-out-Distractor gap; content additions remain the hardest family, and covering the target family in training removes the residual gap; how few verified groups suffice is open. Third, we fine-tune 8B open-weight graders; RITE requires parameter updates and our representation analysis needs hidden states, so adapting rubric invariance to black-box graders is future work. Local 8B graders are, however, attractive in practice: inference runs on a single GPU, student responses never leave the institution, and the grader can be fine-tuned, which API graders cannot.

Ethics Statement

EduR^{2}Bench is built from previously released, de-identified student responses (Mohler; SciEntsBank); no new student data or personally identifiable information were collected. Automated grading can affect students, and more invariant graders reduce, but do not eliminate, inconsistent scoring; we recommend deployment with human oversight and disclosure to students. The benchmark and code will be released for research use under the licenses of the source datasets.

Reproducibility Statement

All experiments use fixed seeds (42, 43, 44), question-level splits, and hyperparameters frozen on the validation set (Appendix A.4). The Appendix gives the training pseudocode (Algorithm 1), full proofs (Appendix A.3), the grading prompt, the variant-generation protocol, and the verification checklist, baseline search spaces and selected hyperparameters (Tab. 8), and the held-out-family results with seed-level win counts (Tab. 17). Datasets, code, checkpoints, and prediction dumps will be released upon publication.

AI Use Statement

`Fable 5` was used to draft criterion decompositions of reference answers, to generate candidate construct-irrelevant variants, and to produce initial criterion-level labels for EduR^{2}Bench; every criterion, variant, and label was subsequently reviewed by the authors as described in Sec. 3 and Appendix A.4. AI assistants were used for language polishing of the manuscript. All research questions, experimental designs, analyses, and claims are the authors' own.

References

Steven Burrows, Iryna Gurevych, and Benno Stein. The eras and trends of automatic short answer grading. *International journal of artificial intelligence in education*, 25(1):60–117, 2015.

Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15607–15631, 2023.

Peter J Congdon and Joy MeQueen. The stability of rater severity in large-scale assessment programs. *Journal of educational measurement*, 37(2):163–178, 2000.

Haotian Deng, Chris Farber, Jiyoon Lee, and David Tang. Rubric-conditioned llm grading: Alignment, uncertainty, and robustness. *arXiv preprint arXiv:2601.08843*, 2026.

Ruomeng Ding, Yifei Pang, He Sun, Yizhong Wang, Zhiwei Steven Wu, and Zhun Deng. Rubrics as an attack surface: Stealthy preference drift in llm judges. *arXiv preprint arXiv:2602.13576*, 2026.

Yuning Ding, Brian Riordan, Andrea Horbach, Aoife Cahill, and Torsten Zesch. Don't take "nswvt-nvakgxpm" for an answer–the surprising vulnerability of automatic content scoring systems to adversarial input. In *Proceedings of the 28th international conference on computational linguistics*, pp. 882–892, 2020.

Jennifer D'Souza, Hamed Babaei Giglou, and Quentin Münch. YESciEval: Robust LLM-as-a-judge for scientific question answering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13749–13783, 2025.

Myroslava O Dzikovska, Rodney Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa Trang Dang. Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pp. 263–274, 2013.

Daniel R Eignor. The standards for educational and psychological testing. 2013.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031, 2021.

Dominik Maria Endres and Johannes E Schindelin. A new metric for probability distributions. *IEEE Transactions on Information theory*, 49(7):1858–1860, 2003.

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.

Rajarshi Haldar and Julia Hockenmaier. Rating roulette: Self-inconsistency in llm-as-a-judge frameworks. *arXiv preprint arXiv:2510.27106*, 2025.

Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. Llm-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13806–13834, 2024.

Parsa Hejabi, Elnaz Rahmati, Alireza Salkhordeh Ziabari, and Morteza Dehghani. Flip-flop consistency: Unsupervised training for robustness to prompt perturbations in llms. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1571–1587, 2026.

Yihan Hong, Huaiyuan Yao, Bolin Shen, Wanpeng Xu, Hua Wei, and Yushun Dong. Rulers: Locked rubrics and evidence-anchored scoring for robust llm evaluation. *arXiv preprint arXiv:2601.08654*, 2026.

Andrea Horbach and Torsten Zesch. The influence of variance in learner answers on automatic content scoring. In *Frontiers in education*, volume 4, pp. 28. Frontiers Media SA, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. Smart: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2177–2190, 2020.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4334–4353, 2024.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. Benchmarking cognitive biases in large language models as evaluators. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 517–545, 2024.

Haitao Li, Junjie Chen, Qingyao Ai, Zhumin Chu, Yujia Zhou, Qian Dong, and Yiqun Liu. Calibraeval: Calibrating prediction distribution to mitigate selection bias in llms-as-judges. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16537–16552, 2025.

Jiatong Li, Renjun Hu, Kunzhe Huang, Yan Zhuang, Qi Liu, Mengxiao Zhu, Xing Shi, and Wei Lin. Perteval: Unveiling real knowledge capacity of LLMs with knowledge-invariant perturbations. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 2511–2522, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Samuel Messick. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American psychologist*, 50 (9):741, 1995.

Samuel Messick, Robert L Linn, et al. Educational measurement. *Validity*, pp. 13–103, 1989.

Michael Mohler, Razvan Bunescu, and Rada Mihalcea. Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pp. 752–762, 2011.

Junsoo Park, Seungyeon Jwa, Ren Meiying, Daeyoung Kim, and Sanghyuk Choi. Offsetbias: Leveraging debiased data for tuning evaluators. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1043–1067, 2024.

Yao Qiang, Subhrangshu Nandi, Ninareh Mehrabi, Greg Ver Steeg, Anoop Kumar, Anna Rumshisky, and Aram Galstyan. Prompt perturbation consistency learning for robust language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 1357–1370, 2024.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4902–4912, 2020.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *International Conference on Learning Representations*, volume 2024, pp. 25055–25083, 2024.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

Shashank Sonkar, Kangqi Ni, Lesa Tran Lu, Kristi Kincaid, John S Hutchinson, and Richard G Baraniuk. Automated long answer grading with ricechem dataset. In *International Conference on Artificial Intelligence in Education*, pp. 163–176. Springer, 2024.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. Large language models are inconsistent and biased evaluators. *arXiv preprint arXiv:2405.01724*, 2024.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

Kimi Team, Yifan Bai, Yiping Bao, Y Charles, Cheng Chen, Guanduo Chen, Haiting Chen, Huarong Chen, Jiahao Chen, Ningxin Chen, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, et al. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450, 2024.

Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, et al. R-drop: Regularized dropout for neural networks. *Advances in neural information processing systems*, 34: 10890–10905, 2021.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268, 2020.

Wenjing Xie, Juxin Niu, Chun Jason Xue, and Nan Guan. Grade like a human: Rethinking automated assessment with large language models. In *Proceedings of the International Conference on Research in Adaptive and Convergent Systems*, pp. 1–8, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, et al. Justice or prejudice? quantifying biases in llm-as-a-judge. In *International Conference on Learning Representations*, volume 2025, pp. 102351–102390, 2025.

Xiaoming Zhai, Kevin C Haudek, Molly AM Stuhlsatz, and Christopher Wilson. Evaluation of construct-irrelevant variance yielded by machine and human scoring of a science teacher pck constructed response assessment. *Studies in Educational Evaluation*, 67:100916, 2020.

Xiaoming Zhai, Kevin C Haudek, Christopher Wilson, and Molly Stuhlsatz. A framework of construct-irrelevant variance for contextualized constructed response assessment. In *Frontiers in Education*, volume 6, pp. 751283. Frontiers Media SA, 2021.

Yukun Zhao, Lingyong Yan, Weiwei Sun, Guoliang Xing, Shuaiqiang Wang, Chong Meng, Zhicong Cheng, Zhaochun Ren, and Dawei Yin. Improving the robustness of large language models via consistency alignment. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 8931–8941, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. Judgelm: Fine-tuned large language models are scalable judges. In *International Conference on Learning Representations*, volume 2025, pp. 51257–51296, 2025.

A APPENDIX

In the appendix, we present the pseudocode and full proofs, experimental setup, supplemental results, and a notation list.

A.1 NOTATION LIST

For reference, we summarize the main notation used for the data, RITE objective, theoretical analysis, evaluation metrics, and representation analysis in Tab. 22.

A.2 PSEUDOCODE OF RITE

We present the pseudocode of RITE and analyze its training procedure and computational complexity as follows.

Algorithm 1: RITE Training.

Require: Training set $\mathcal{D}_{\mathrm{tr}} = \{(q_i, \mathcal{R}_i, \mathcal{A}_i, \mathbf{z}_i^*)\}_{i=1}^N$; supervision scopes $\{\mathcal{S}_i\}_{i=1}^N$; grader $f_{\boldsymbol{\theta}}$; epochs E; RI weight λ_{RI}; learning rate η.

Initialize grader parameters $\boldsymbol{\theta}$;
for $e = 1$ to E **do**
  for a batch $\mathcal{B} \subset \mathcal{D}_{\mathrm{tr}}$ **do**
    $\mathcal{L}_{\mathrm{PBCE}} \leftarrow 0$; $\mathcal{L}_{\mathrm{RI}} \leftarrow 0$;
    for each $i \in \mathcal{B}$ **do**
      for $m = 0$ to $M_i - 1$ **do**
        $\{p_{i,k}^{(m)}\}_{k=1}^{K_i} \leftarrow f_{\boldsymbol{\theta}}(q_i, \mathcal{R}_i, a_i^{(m)})$;
        $\boldsymbol{\pi}_i^{(m)} \leftarrow \mathrm{PB}\big(\{p_{i,k}^{(m)}\}_{k=1}^{K_i}\big)$;
      end for
    end for
    Compute $\mathcal{L}_{\mathrm{RITE}}$ via Eq. (3);
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{RITE}}$;
  end for
end for
return Trained grader $f_{\boldsymbol{\theta}}$.

Training procedure. Algo. 1 summarizes RITE training. We first initialize the model parameters. For each instance i in a training batch, the grader processes all M_i answers and produces their criterion-level probabilities and induced Poisson–Binomial score distributions. We then compute the total loss by Eq. (3) and update the gradient. By the end of training, we obtain a trained grader.

Time and space complexity. Let C_{model} denote the average forward-and-backward cost of grading one answer. For instance i, processing all M_i answers costs $\mathcal{O}(M_i C_{\mathrm{model}})$. Constructing each Poisson–Binomial score distribution over $K_i + 1$ scores by dynamic programming costs $\mathcal{O}(K_i^2)$, giving $\mathcal{O}(M_i K_i^2)$. The correctness loss costs $\mathcal{O}(|\mathcal{S}_i| K_i)$, and the $M_i - 1$ anchor-to-variant Jensen–Shannon divergences cost $\mathcal{O}((M_i - 1)K_i)$. Thus, the total training-time complexity over E epochs is

$$\mathcal{O}\left(E \sum_{i=1}^{N} \left[M_i C_{\mathrm{model}} + M_i K_i^2 + |\mathcal{S}_i| K_i + (M_i - 1) K_i \right] \right).$$

For LLM graders, model computation typically dominates the auxiliary losses.

For a batch $\mathcal{B}$, let $M_{\boldsymbol{\theta}}$, M_{opt}, and $M_{\mathrm{act}}(\mathcal{B})$ denote the memory for parameters, optimizer states, and activations. A straightforward differentiable Poisson–Binomial implementation may retain $\mathcal{O}(\sum_{i \in \mathcal{B}} M_i K_i^2)$ intermediate states, giving peak memory

$$\mathcal{O}\left(M_{\boldsymbol{\theta}} + M_{\mathrm{opt}} + M_{\mathrm{act}}(\mathcal{B}) + \sum_{i \in \mathcal{B}} M_i K_i^2 \right).$$

Checkpointing or recomputation can reduce the auxiliary memory to $\mathcal{O}(\sum_{i\in\mathcal{B}} M_i K_i)$. The anchored topology uses only $M_i - 1$ divergences instead of $\binom{M_i}{2}$ all-pairs comparisons. At inference time, RITE grades each answer independently and has the same asymptotic cost as the underlying grader.

A.3 Theoretical Proofs

A.3.1 Auxiliary Results

Lemma 1 (Decision transfer). *Let $\boldsymbol{\pi}$ and $\boldsymbol{\pi}'$ be probability distributions on $\{0,\dots,K_i\}$ satisfying $D_{\mathrm{JS}}(\boldsymbol{\pi} \parallel \boldsymbol{\pi}') \le \varepsilon$. Then*

$$d_{\mathrm{TV}}(\boldsymbol{\pi}, \boldsymbol{\pi}') \le \sqrt{2\varepsilon}.$$

Moreover, for any statistic $g : \{0,\dots,K_i\} \to [a,b]$,

$$|\mathbb{E}_{\boldsymbol{\pi}}[g] - \mathbb{E}_{\boldsymbol{\pi}'}[g]| \le (b-a) d_{\mathrm{TV}}(\boldsymbol{\pi}, \boldsymbol{\pi}') \le (b-a)\sqrt{2\varepsilon}.$$

In particular, the probabilities of exceeding any score threshold differ by at most $\sqrt{2\varepsilon}$, and the expected total scores differ by at most $K_i\sqrt{2\varepsilon}$.

Proof. Let $\boldsymbol{\mu}=(\boldsymbol{\pi}+\boldsymbol{\pi}')/2$. By the definition of total variation,

$$d_{\mathrm{TV}}(\boldsymbol{\pi}, \boldsymbol{\mu}) = d_{\mathrm{TV}}(\boldsymbol{\pi}', \boldsymbol{\mu}) = \frac{1}{2} d_{\mathrm{TV}}(\boldsymbol{\pi}, \boldsymbol{\pi}').$$

Pinsker's inequality gives

$$\mathrm{KL}(\boldsymbol{\pi} \parallel \boldsymbol{\mu}) \ge 2 d_{\mathrm{TV}}^2(\boldsymbol{\pi}, \boldsymbol{\mu}), \qquad \mathrm{KL}(\boldsymbol{\pi}' \parallel \boldsymbol{\mu}) \ge 2 d_{\mathrm{TV}}^2(\boldsymbol{\pi}', \boldsymbol{\mu}).$$

Therefore,

$$D_{\mathrm{JS}}(\boldsymbol{\pi} \parallel \boldsymbol{\pi}') = \frac{1}{2}\mathrm{KL}(\boldsymbol{\pi} \parallel \boldsymbol{\mu}) + \frac{1}{2}\mathrm{KL}(\boldsymbol{\pi}' \parallel \boldsymbol{\mu}) \ge \frac{1}{2} d_{\mathrm{TV}}^2(\boldsymbol{\pi}, \boldsymbol{\pi}'),$$

which implies $d_{\mathrm{TV}}(\boldsymbol{\pi}, \boldsymbol{\pi}') \le \sqrt{2\varepsilon}$.

For $g : \{0,\dots,K_i\} \to [a,b]$, let $\widetilde{g}=g-(a+b)/2$. Since $\|\widetilde{g}\|_\infty \le (b-a)/2$, the dual characterization of total variation gives:

$$\begin{aligned} |\mathbb{E}_{\boldsymbol{\pi}}[g] - \mathbb{E}_{\boldsymbol{\pi}'}[g]| &= |\mathbb{E}_{\boldsymbol{\pi}}[\widetilde{g}] - \mathbb{E}_{\boldsymbol{\pi}'}[\widetilde{g}]| \\ &\le 2\|\widetilde{g}\|_\infty d_{\mathrm{TV}}(\boldsymbol{\pi}, \boldsymbol{\pi}') \\ &\le (b-a) d_{\mathrm{TV}}(\boldsymbol{\pi}, \boldsymbol{\pi}'). \end{aligned}$$

The threshold result follows by taking $g(r)=\mathbb{I}[r \ge t]$, whose range is 1. The expected-score result follows by taking $g(r)=r$, whose range is K_i. □

Lemma 2 (Poisson–Binomial coupling). *For any $\mathbf{p}, \mathbf{q} \in [0,1]^K$,*

$$d_{\mathrm{TV}}(\mathrm{PB}(\mathbf{p}), \mathrm{PB}(\mathbf{q})) \le \sum_{k=1}^{K} |p_k - q_k| = \|\mathbf{p} - \mathbf{q}\|_1.$$

Proof. Let $U_1,\dots,U_K$ be independent random variables uniformly distributed on $[0,1]$, and define

$$X_k=\mathbb{I}[U_k \le p_k], \qquad Y_k=\mathbb{I}[U_k \le q_k].$$

Then

$$X=\sum_{k=1}^{K} X_k \sim \mathrm{PB}(\mathbf{p}), \qquad Y=\sum_{k=1}^{K} Y_k \sim \mathrm{PB}(\mathbf{q}),$$

and $\Pr(X_k \ne Y_k)=|p_k - q_k|$. By the coupling inequality and the union bound,

$$\begin{aligned} d_{\mathrm{TV}}(\mathrm{PB}(\mathbf{p}), \mathrm{PB}(\mathbf{q})) &\le \Pr(X \ne Y) \\ &\le \sum_{k=1}^{K} \Pr(X_k \ne Y_k) \\ &= \|\mathbf{p} - \mathbf{q}\|_1. \end{aligned}$$

□

A.3.2 PROOFS OF MAIN-TEXT RESULTS

Proof of Theorem 1. Correctness part. For $z^*_{i,k} \in \{0,1\}$, the per-answer correctness loss can be written as

$$\ell_i^{(m)} = \frac{1}{K_i} \sum_{k=1}^{K_i} -\log\Big(1 - \Big|p_{i,k}^{(m)} - z^*_{i,k}\Big|\Big).$$

Under the extended-value convention, $-\log(1-t)$ is strictly increasing on $[0,1)$ and equals $+\infty$ at $t{=}1$. Each summand is therefore uniquely minimized at $|p_{i,k}^{(m)} - z^*_{i,k}|{=}0$. Hence, over $[0,1]^{K_i}$,

$$\mathbf{p}_i^{(m)}{=}\mathbf{z}^*_i$$

is the unique minimizer of $\ell_i^{(m)}$.

Equivalence part. Let $\boldsymbol{\nu} \in \mathfrak{P}_i$. By definition, there exists $\bar{\mathbf{p}} \in [0,1]^{K_i}$ such that

$$\boldsymbol{\nu}{=}\,\mathrm{PB}(\bar{\mathbf{p}}).$$

If the grader outputs $\bar{\mathbf{p}}$ for every answer in instance i, then

$$\boldsymbol{\pi}_i^{(0)} = \cdots = \boldsymbol{\pi}_i^{(M_i-1)} = \boldsymbol{\nu},$$

and every equivalence-loss term associated with that instance is zero.

To exhibit an incorrect invariant solution, take $\bar{\mathbf{p}}{=}(\frac{1}{2},\ldots,\frac{1}{2})$. Since $K_i \geq 1$, $\mathrm{PB}(\bar{\mathbf{p}})$ is non-degenerate and therefore differs from the ground-truth point mass $\boldsymbol{\pi}^*_i{=}\delta_{s^*_i}$. Thus, equivalence supervision alone cannot identify the ground-truth score distribution. □

Proof of Theorem 2. Since d_{TV} is a metric,

$$d_{\mathrm{TV}}\Big(\boldsymbol{\pi}_i^{(m)}, \boldsymbol{\pi}_i^*\Big) \leq d_{\mathrm{TV}}\Big(\boldsymbol{\pi}_i^{(m)}, \boldsymbol{\pi}_i^{(0)}\Big) + d_{\mathrm{TV}}\Big(\boldsymbol{\pi}_i^{(0)}, \boldsymbol{\pi}_i^*\Big).$$

By Lemma 1 and the symmetry of the Jensen–Shannon divergence,

$$d_{\mathrm{TV}}\Big(\boldsymbol{\pi}_i^{(m)}, \boldsymbol{\pi}_i^{(0)}\Big) \leq \sqrt{2D_{\mathrm{JS}}\Big(\boldsymbol{\pi}_i^{(0)} \parallel \boldsymbol{\pi}_i^{(m)}\Big)}.$$

For the grounding term,

$$\boldsymbol{\pi}^*_i = \delta_{s^*_i} = \mathrm{PB}(\mathbf{z}^*_i).$$

For each criterion, the inequality $-\log(1-t) \geq t$ for $t \in [0,1)$ gives

$$\Big|p_{i,k}^{(0)} - z^*_{i,k}\Big| \leq -\Big[z^*_{i,k} \log p_{i,k}^{(0)} + (1 - z^*_{i,k}) \log(1 - p_{i,k}^{(0)})\Big].$$

Summing over criteria yields

$$\|\mathbf{p}_i^{(0)} - \mathbf{z}^*_i\|_1 \leq K_i \ell_i^{(0)}.$$

Applying Lemma 2,

$$d_{\mathrm{TV}}\Big(\boldsymbol{\pi}_i^{(0)}, \boldsymbol{\pi}_i^*\Big) \leq \|\mathbf{p}_i^{(0)} - \mathbf{z}^*_i\|_1 \leq K_i \ell_i^{(0)}.$$

Combining the inequalities proves

$$d_{\mathrm{TV}}\Big(\boldsymbol{\pi}_i^{(m)}, \boldsymbol{\pi}_i^*\Big) \leq \sqrt{2D_{\mathrm{JS}}\Big(\boldsymbol{\pi}_i^{(0)} \parallel \boldsymbol{\pi}_i^{(m)}\Big)} + K_i \ell_i^{(0)}.$$

The bound for any statistic with range $b - a$ follows by applying the second part of Lemma 1. As a numerical sanity check, take $K_i{=}1$, $z^*{=}1$, $p^{(0)}{=}0.8$, $p^{(m)}{=}0.1$: the left-hand side is 0.9, while $\sqrt{2D_{\mathrm{JS}}}{=}\sqrt{2 \times 0.2754}{=}0.742$ and $K_i\ell^{(0)}{=}-\ln 0.8{=}0.223$, so the right-hand side is $0.965 \geq 0.9$. □

Proof of Proposition 1. The objective $\mathcal{L}_{\mathrm{PBCE}}$ depends only on predictions whose indices belong to $\mathcal{S}_i$. Consider an unsupervised input $\mathbf{x}_i^{(m)}$ with $m \notin \mathcal{S}_i$ that is distinct from all supervised inputs. Over an unconstrained predictor class, define a grader f' that agrees with f on every supervised input but assigns an arbitrary prediction to $\mathbf{x}_i^{(m)}$. Every term in $\mathcal{L}_{\mathrm{PBCE}}$ remains unchanged. Hence, the correctness objective imposes no constraint on predictions outside the supervision scope. □

Proof of Proposition 2. Assume $K_i \geq 2$. The mapping $\mathrm{PB}(\cdot)$ is invariant to permutations of its parameters. For any permutation σ of $\{1, \ldots, K_i\}$,

$$\mathrm{PB}(p_1, \ldots, p_{K_i}) = \mathrm{PB}\big(p_{\sigma(1)}, \ldots, p_{\sigma(K_i)}\big)\,.$$

Consequently, two distinct criterion-level probability vectors related by a nontrivial permutation induce the same total-score distribution. For example,

$$\mathrm{PB}(1, 0) = \mathrm{PB}(0, 1) = \delta_1.$$

Any divergence defined only on total-score distributions is therefore insensitive to such criterion-level permutations.

For directly supervised answers, Theorem 1 shows that $\ell_i^{(m)}$ is uniquely minimized at $\mathbf{p}_i^{(m)}{=}\mathbf{z}_i^*$, thereby identifying the gold criterion pattern. □

A.3.3 Consequences of the Decomposition

Corollary 1 (Propagation at the optimum)**.** *Suppose $\mathcal{S}_i{=}\{0\}$, the label $\mathbf{z}_i^*$ is correct, and all correctness and equivalence terms associated with instance i vanish. Then*

$$\boldsymbol{\pi}_i^{(m)} = \boldsymbol{\pi}_i^*$$

for every $m \in \{0, \ldots, M_i - 1\}$. Thus, every variant receives the correct total score without direct correctness supervision.

Proof. The vanishing anchor correctness loss implies $\ell_i^{(0)}{=}0$. By Theorem 1,

$$\mathbf{p}_i^{(0)} = \mathbf{z}_i^*, \qquad \boldsymbol{\pi}_i^{(0)} = \mathrm{PB}(\mathbf{z}_i^*) = \boldsymbol{\pi}_i^*.$$

Each equivalence term also vanishes. Since $D_{\mathrm{JS}}(\boldsymbol{\pi}, \boldsymbol{\pi}'){=}0$ if and only if $\boldsymbol{\pi}{=}\boldsymbol{\pi}'$,

$$\boldsymbol{\pi}_i^{(m)} = \boldsymbol{\pi}_i^{(0)} = \boldsymbol{\pi}_i^*$$

for every variant. □

Corollary 2 (Supervised coupling)**.** *For any $m, m' \in \mathcal{S}_i$,*

$$d_{\mathrm{TV}}\Big(\boldsymbol{\pi}_i^{(m)}, \boldsymbol{\pi}_i^{(m')}\Big) \leq K_i\Big(\ell_i^{(m)} + \ell_i^{(m')}\Big)\,.$$

Thus, the score-distribution discrepancy between directly supervised answers is controlled by their correctness losses, even without an explicit equivalence objective.

Proof. By the triangle inequality through $\boldsymbol{\pi}_i^*$,

$$d_{\mathrm{TV}}\Big(\boldsymbol{\pi}_i^{(m)}, \boldsymbol{\pi}_i^{(m')}\Big) \leq d_{\mathrm{TV}}\Big(\boldsymbol{\pi}_i^{(m)}, \boldsymbol{\pi}_i^*\Big) + d_{\mathrm{TV}}\Big(\boldsymbol{\pi}_i^{(m')}, \boldsymbol{\pi}_i^*\Big)\,.$$

Applying Lemma 2 and the BCE bound used in the proof of Theorem 2 gives

$$d_{\mathrm{TV}}\Big(\boldsymbol{\pi}_i^{(m)}, \boldsymbol{\pi}_i^*\Big) \leq \|\mathbf{p}_i^{(m)} - \mathbf{z}_i^*\|_1 \leq K_i \ell_i^{(m)},$$

and analogously for m'. Combining the two bounds proves the result. □

Corollary 3 (Corrupted anchor)**.** *Suppose the anchor is supervised toward a corrupted label $\tilde{\mathbf{z}}_i \in \{0, 1\}^{K_i}$. Then, for every m,*

$$d_{\mathrm{TV}}\Big(\boldsymbol{\pi}_i^{(m)}, \boldsymbol{\pi}_i^*\Big) \leq \sqrt{2 D_{\mathrm{JS}}\Big(\boldsymbol{\pi}_i^{(0)} \,\|\, \boldsymbol{\pi}_i^{(m)}\Big)} + \eta_i,$$

where

$$\eta_i = \|\mathbf{p}_i^{(0)} - \tilde{\mathbf{z}}_i\|_1 + d_{\mathrm{H}}(\tilde{\mathbf{z}}_i, \mathbf{z}_i^*)\,.$$

Under independent criterion flips with rate ρ,

$$\mathbb{E}[d_{\mathrm{H}}(\tilde{\mathbf{z}}_i, \mathbf{z}_i^*)] = \rho K_i.$$

If the corrupted anchor correctness term and all associated equivalence terms vanish, then

$$\boldsymbol{\pi}_i^{(m)} = \delta_{\tilde{s}_i}, \qquad \tilde{s}_i = \sum_{k=1}^{K_i} \tilde{z}_{i,k},$$

and therefore

$$\hat{s}_i^{(m)} = \tilde{s}_i, \qquad \left|\hat{s}_i^{(m)} - s_i^*\right| = |\tilde{s}_i - s_i^*|\,.$$

Thus, at the corrupted joint optimum, perfect equivalence alignment propagates rather than corrects the anchor's score error.

Proof. The triangle inequality gives

$$\|\mathbf{p}_i^{(0)} - \mathbf{z}_i^*\|_1 \le \|\mathbf{p}_i^{(0)} - \tilde{\mathbf{z}}_i\|_1 + \|\tilde{\mathbf{z}}_i - \mathbf{z}_i^*\|_1.$$

Because both label vectors are binary,

$$\|\tilde{\mathbf{z}}_i - \mathbf{z}_i^*\|_1 = d_{\mathrm{H}}(\tilde{\mathbf{z}}_i, \mathbf{z}_i^*)\,.$$

Substituting this bound for the grounding term in the proof of Theorem 2 establishes the first claim.

Under independent flips, write

$$d_{\mathrm{H}}(\tilde{\mathbf{z}}_i, \mathbf{z}_i^*) = \sum_{k=1}^{K_i} \mathbb{I}\big[\tilde{z}_{i,k} \neq z_{i,k}^*\big]\,.$$

Taking expectations and using $\Pr(\tilde{z}_{i,k} \neq z_{i,k}^*)=\rho$ gives

$$\mathbb{E}[d_{\mathrm{H}}(\tilde{\mathbf{z}}_i, \mathbf{z}_i^*)] = \rho K_i.$$

At the corrupted joint optimum, the vanishing correctness term implies

$$\mathbf{p}_i^{(0)} = \tilde{\mathbf{z}}_i, \qquad \boldsymbol{\pi}_i^{(0)} = \mathrm{PB}(\tilde{\mathbf{z}}_i) = \delta_{\tilde{s}_i}.$$

The vanishing equivalence terms imply

$$\boldsymbol{\pi}_i^{(m)} = \boldsymbol{\pi}_i^{(0)} = \delta_{\tilde{s}_i}.$$

A Poisson–Binomial distribution is a point mass only when its Bernoulli parameters are binary, so the corresponding predicted total score is $\hat{s}_i^{(m)}=\tilde{s}_i$. The stated score error follows. □

Corollary 4 (Star topology certifies mutual consistency)**.** *Because $\sqrt{D_{\mathrm{JS}}}$ is a metric (Endres & Schindelin, 2003), if*

$$D_{\mathrm{JS}}\Big(\boldsymbol{\pi}_i^{(0)} \parallel \boldsymbol{\pi}_i^{(m)}\Big) \le \varepsilon$$

for every $m \in \{1, \ldots, M_i - 1\}$, then

$$D_{\mathrm{JS}}\Big(\boldsymbol{\pi}_i^{(m)} \parallel \boldsymbol{\pi}_i^{(m')}\Big) \le 4\varepsilon$$

for every $m, m' \in \{0, \ldots, M_i - 1\}$. Thus, the anchored star topology controls all pairwise score-distribution discrepancies using $M_i - 1$ alignment terms rather than $\binom{M_i}{2}$.

Proof. By the triangle inequality for $\sqrt{D_{\mathrm{JS}}}$ through the anchor $\boldsymbol{\pi}_i^{(0)}$,

$$\begin{aligned}
&\sqrt{D_{\mathrm{JS}}\Big(\boldsymbol{\pi}_i^{(m)} \parallel \boldsymbol{\pi}_i^{(m')}\Big)} \\
&\le \sqrt{D_{\mathrm{JS}}\Big(\boldsymbol{\pi}_i^{(m)} \parallel \boldsymbol{\pi}_i^{(0)}\Big)} + \sqrt{D_{\mathrm{JS}}\Big(\boldsymbol{\pi}_i^{(0)} \parallel \boldsymbol{\pi}_i^{(m')}\Big)} \\
&\le 2\sqrt{\varepsilon}.
\end{aligned}$$

Squaring both sides proves the claim. □

A.3.4 Design and Transfer Limits

Proposition 3 (Insufficiency of weaker targets)**.** (i) *There exist score distributions with identical expected scores but different probabilities of exceeding a score threshold. Thus, expected-score alignment alone does not control score-based decisions.*

(ii) *For distributions on* $\{0, \dots, K_i\}$*,*

$$d_{\mathrm{TV}}(\boldsymbol{\pi}, \boldsymbol{\pi}') \le \frac{1}{2}\sqrt{K_i + 1}\,\|\boldsymbol{\pi} - \boldsymbol{\pi}'\|_2.$$

Moreover, the maximum value of the unnormalized W_1 *distance grows linearly with* K_i*, whereas* $D_{\mathrm{JS}} \in [0, \ln 2]$ *independently of* K_i*.*

Proof. For part (i), consider the distributions on $\{0, 1, 2\}$

$$\boldsymbol{\pi} = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_2, \qquad \boldsymbol{\pi}' = \delta_1.$$

Both have expected score 1. However, for a random score T and threshold $t{=}2$,

$$\Pr_{\boldsymbol{\pi}}(T \ge 2) = \frac{1}{2}, \qquad \Pr_{\boldsymbol{\pi}'}(T \ge 2) = 0.$$

Hence, equal expected scores do not imply equal score-based decisions.

For part (ii), Cauchy–Schwarz on a support of size $K_i + 1$ gives

$$\|\boldsymbol{\pi} - \boldsymbol{\pi}'\|_1 \le \sqrt{K_i + 1}\,\|\boldsymbol{\pi} - \boldsymbol{\pi}'\|_2.$$

Since $d_{\mathrm{TV}}(\boldsymbol{\pi}, \boldsymbol{\pi}'){=}\frac{1}{2}\|\boldsymbol{\pi} - \boldsymbol{\pi}'\|_1$,

$$d_{\mathrm{TV}}(\boldsymbol{\pi}, \boldsymbol{\pi}') \le \frac{1}{2}\sqrt{K_i + 1}\,\|\boldsymbol{\pi} - \boldsymbol{\pi}'\|_2.$$

For W_1 on the ordinal support $\{0, \dots, K_i\}$,

$$W_1(\boldsymbol{\pi}, \boldsymbol{\pi}') = \sum_{r=0}^{K_i - 1} |F_{\boldsymbol{\pi}}(r) - F_{\boldsymbol{\pi}'}(r)| \le K_i.$$

The upper bound is attained by δ_0 and δ_{K_i}, so the raw scale of W_1 grows linearly with K_i. Our implementation therefore normalizes W_1 by K_i. Finally,

$$0 \le D_{\mathrm{JS}}(\boldsymbol{\pi} \parallel \boldsymbol{\pi}') \le \ln 2$$

for distributions on any finite support. □

Proposition 4 (No free transfer)**.** *Let* $\mathcal{T}'$ *be a transformation family whose variants do not appear in the training set. For any grader* f *attaining zero training objective, there exists a grader* f' *that also attains zero training objective, agrees with* f *on every training input, and assigns arbitrary predictions to inputs transformed by* $\mathcal{T}'$*. Without structural assumptions on the predictor class, the training objective therefore provides no general invariance guarantee for unseen transformation families.*

Proof. Over an unconstrained predictor class, construct f' by setting $f'(\mathbf{x}){=}f(\mathbf{x})$ for every training input and assigning arbitrary predictions to inputs produced by $\mathcal{T}'$. Because no $\mathcal{T}'$-variant appears in the training set, every term of $\mathcal{L}_{\mathrm{PBCE}}$ and $\mathcal{L}_{\mathrm{RI}}$ remains unchanged. Hence,

$$\mathcal{L}_{\mathrm{RITE}}(f') = \mathcal{L}_{\mathrm{RITE}}(f) = 0,$$

while the score distributions assigned to unseen $\mathcal{T}'$-variants may differ arbitrarily from those assigned by f. □

A.4 Experimental Setup

EduR^{2}Bench. Existing short-answer grading benchmarks mainly evaluate whether a grader gives the correct judgment (Burrows et al., 2015; Horbach & Zesch, 2019), but do not provide the controlled equivalence structure needed to measure rubric invariance. We therefore introduce **EduR^{2}Bench**, a criterion-level benchmark built from authentic student responses across four STEM subjects: computer science, physics, chemistry, and Earth science. Unlike robustness benchmarks based on generic text perturbations (Ribeiro et al., 2020; Ding et al., 2020), EduR^{2}Bench uses verified label-preserving variants and measures drift at the criterion level. Computer-science responses come from Mohler (Mohler et al., 2011); physics, chemistry, and Earth science from SciEntsBank (Dzikovska et al., 2013). For each question q_i, we represent the rubric as K_i binary criteria and associate each original response with a criterion-level rubric state. The basic unit of the benchmark is a *rubric-equivalence instance*, which pairs an original student response with applicable controlled variants from four families of rubric-irrelevant variation: **Position**, which reorders logically independent content; **Length**, which verbosely expands the wording; **Style**, which reformats prose into a bulleted list; and **Distractor**, which adds topically related but rubric-irrelevant information. Each transformation is designed to preserve rubric-relevant evidence and is applied only when suitable, so the number of valid variants differs across instances. Each instance has a shared gold rubric state $\mathbf{z}^*$ of criterion-level hit/miss labels. We obtain initial labels with `Fable 5` and manually review them. As an independent external check, we compare `Fable 5` annotations against teaching-assistant labels from RiceChem, obtaining 95% criterion-level agreement. We further verify every variant for correct transformation, preservation of rubric-relevant evidence, and an unchanged criterion-level rubric state; failed variants are regenerated or discarded. EduR^{2}Bench contains 1,060 rubric-equivalence instances and 3,297 construct-irrelevant variants from 176 questions, with rubrics containing 1–4 binary criteria. We split the data at the question level to prevent leakage, yielding 680 training, 124 validation, and 256 test instances. Tab. 6 reports the per-subject statistics.

Table 6: EduR^{2}Bench statistics by subject: rubric-equivalence groups, construct-irrelevant variants (excluding the original responses), and question-level train/val/test split. SEB denotes SciEntsBank.

Subject	Source	Groups	Variants	Train	Val	Test
Computer Science	Mohler	508	1,580	344	52	112
Physics	SEB	228	699	140	24	64
Chemistry	SEB	196	607	120	32	44
Earth Science	SEB	128	411	76	16	36
Total		**1,060**	**3,297**	**680**	**124**	**256**

Transformation families in EduR^{2}Bench. As shown in Tab. 7, each rubric-equivalence instance contains an original student answer and applicable variants from four controlled transformation families. **Position** reorders logically independent sentences or clauses without changing their meaning; it is applied only when the answer has at least two such units and no subordinating conjunction. **Length** verbosely expands the wording (1.5–3$\times$ the original word count; every original sentence kept verbatim and in order) while preserving the rubric-relevant evidence. **Style** reformats the answer from prose into a bulleted list without changing its wording. **Distractor** adds topically related but rubric-irrelevant information without changing the evidence used for scoring. A transformation is applied only when it is valid for the original answer, so not every instance contains all four variants.

Measurement Protocol. We evaluate LLM graders along two complementary dimensions: *grading quality* and *rubric invariance*. Grading quality measures agreement with the gold rubric state $\mathbf{z}^*$. Rubric invariance measures whether judgments remain stable across rubric-equivalent responses. The two are distinct: a grader may be consistently wrong, or broadly accurate yet unstable under rubric-irrelevant variation (Elazar et al., 2021; Chiang & Lee, 2023). We use three gold-referenced metrics for grading quality. **Quadratic Weighted Kappa (QWK, $\uparrow$)** measures agreement at the total-score level. Questions with different rubric sizes K have different score ranges. We therefore compute QWK separately for each rubric-size bucket and take a size-weighted average. **Criterion-level Macro-F1 (C-F1, $\uparrow$)** averages the F1 scores of the hit and miss classes over all criterion-level decisions. **Score-Error Standard Deviation (Err. SD, $\downarrow$)** is the standard deviation of the per-response

Table 7: A rubric-equivalence instance from the CS training split of EduR^{2}Bench. Each variant changes one construct-irrelevant property while preserving the rubric-relevant evidence and shared gold rubric state.

Question: What is the main advantage of linked lists over arrays?
Rubric: C1: "The linked lists can be of variable length." **Gold rubric state:** $\mathbf{z}^* = (1)$

Type	Answer	Controlled change
Original	You can expand when you need it. You do not have to predetermine the size.	Original benchmark answer.
Position	You do not have to predetermine the size. You can expand when you need it.	Reorders two logically independent statements.
Length	You can expand when you need it. *That is, growth happens on demand.* You do not have to predetermine the size. *In other words, no size needs deciding up front.*	Adds explanatory wording without changing the rubric-relevant evidence.
Style	• You can expand when you need it. • You do not have to predetermine the size.	Changes prose into a bullet-note register.
Distractor	You can expand when you need it. *ARPANET switched to the TCP/IP protocol suite on January 1, 1983.* You do not have to predetermine the size.	Adds topically related but rubric-irrelevant information.

All five answers preserve the evidence that a linked list can grow without a predetermined size and therefore share the same gold rubric state. Inserted material is shown in italics.

error between predicted and gold total scores. Our primary rubric-invariance metric is **Criterion Flip Rate (CFR, ↓)**. For each rubric-equivalence group and criterion, we record a flip when the grader's hit/miss judgment changes across responses in the group. A group–criterion pair is valid if predictions are available for at least two responses. CFR is the proportion of valid pairs that contain a flip. CFR does not require access to $\mathbf{z}^*$ because it only compares predictions within each group. Its validity instead relies on the verified equivalence of responses within that group. We measure drift at the criterion level because unchanged total scores do not guarantee unchanged criterion-level judgments. We use greedy decoding (T=0) to rule out random sampling as a source of observed flips. We report QWK, C-F1, and Err. SD alongside CFR rather than treating grading accuracy as a proxy for rubric invariance. Family-specific CFR restricts each group–criterion pair to the original answer and the variant of one family; a pair is valid for that family only if the family's variant exists and both predictions are available. It counts a subset of the disagreements counted by CFR, but over the groups that contain the family's variant, so the two rates have different denominators and neither bounds the other in general.

Representation Analysis Protocol. So far, we have characterized rubric drift as an output-level phenomenon: grading judgments change across rubric-equivalent responses. We further ask whether this behavioral instability is accompanied by changes in the model's internal representations. To study this, we define a criterion-level representation gap and relate it to behavioral flips. For each criterion $c_{i,k}$ and response $a_i^{(m)}$, we extract the hidden state $\mathbf{h}_{i,k,l}^{(m)}$ at layer l from the position used to predict the corresponding hit/miss judgment. We use the original response (m=0) as the reference and define the representation gap as

$$g_{i,k,l}^{(m)} = 1 - \cos\left(\mathbf{h}_{i,k,l}^{(m)}, \mathbf{h}_{i,k,l}^{(0)}\right).$$

This measures how much a rubric-irrelevant variation changes the representation of the same criterion. A larger gap indicates a larger change in representation and therefore weaker representation-level invariance. For each group–criterion pair, we average the gap over its valid variants so that the analysis uses the same unit as CFR. We compute the representation gap at every layer to obtain a layer-wise profile. Using a fixed window size, we define the consecutive layers with the largest average gap as the peak band. The peak band is selected from the overall gap profile without using flip/stable labels. Specific layers are reported in the results section. To connect internal representations with behavioral rubric drift, we classify each group–criterion pair as flip or stable according to CFR and compare their

gaps in the peak band. A systematically larger gap for flip cases indicates that behavioral flips are associated with larger internal representation changes. We treat this pattern as a representation-level signature of rubric drift, not as evidence of a causal mechanism.

Implementation Settings. Our main experiments and analyses cover all four subjects in EduR^{2}Bench and use the question-level splits in Tab. 6. All experiments use verified clean labels, except for the label-noise setting. We study three controlled settings, each varying one aspect of supervision while keeping the others fixed. The coverage setting changes the scope of direct correctness supervision, denoted by $\mathcal{S}_i$. The reliability setting injects criterion-level label noise at rate ρ. The transfer setting holds out one transformation family during training and evaluates the model on test variants from that unseen family. Unless otherwise stated, computationally intensive ablations and sensitivity analyses are conducted on the computer science subset, which is the largest. Within each comparison, all trainable methods use the same backbone, data splits, and training budget. We fine-tune the models with LoRA (Hu et al., 2022) using rank 16 and α=32. Training runs for 5 epochs with AdamW (Loshchilov & Hutter, 2017) and gradient clipping at 1.0. The learning rate for `Qwen3-8B` is 1×10^{-4}. For methods that use instance structure, batches preserve complete equivalence instances. RITE uses Jensen–Shannon divergence by default, with λ_{RI}=1 fixed across all experimental conditions. All tunable hyperparameters and data-dependent analysis choices are selected on the validation set and frozen before testing. The test set is used only for final evaluation. Each configuration is run with three random seeds (42, 43, and 44), and we report the mean and standard deviation. All graders use greedy decoding with T=0. Models fine-tuned in our experiments share the same output format, with generation capped at 128 tokens.

Hyperparameters for baselines. All trainable methods share the same backbone, data splits, LoRA configuration (r=16, α=32), AdamW optimizer, learning rate of 1×10^{-4}, gradient clipping at 1.0, and a total budget of five training epochs; these shared settings are fixed for all methods and are not re-tuned per method. Methods that use group structure preserve complete rubric-equivalence instances within each batch. Base receives no task-specific training. PBCE-Orig and PBCE-All apply direct criterion-level correctness supervision to the original answer only and to all members of each equivalence instance, respectively. Unless otherwise stated, trained baselines use original-only direct supervision; GroupDRO and CA receive full-group labels.

Method-specific hyperparameters are not taken from the original papers but are tuned on the EduR^{2}Bench CS validation set under the same protocol as RITE: the same backbone, supervision setting, and seeds, and a search budget at least as large as RITE's own λ_{RI} sweep (Tab. 5). Every consistency or regularization weight is swept over the same grid $\{0.1, 0.5, 1, 2, 4\}$ as λ_{RI}, and methods with additional knobs (temperature, confidence threshold, perturbation radius, ascent steps, EMA decay, step size) receive additional grid points. Configurations are selected by the criterion used for λ_{RI}, namely the lowest validation CFR with QWK breaking ties, frozen before test evaluation, and reused unchanged for all four subjects and both backbones, exactly as λ_{RI}=1 is for RITE. Tab. 8 lists the search space and the selected value for each baseline. The selected weights cluster at 0.5–1, the same range as λ_{RI}, and validation CFR is flat within this range for UDA ($\lambda_u \in [0.5, 2]$) and for the FixMatch threshold ($\tau \in \{0.8, 0.95\}$), so the comparison in Tab. 3 is not sensitive to the exact choice.

With these settings, the baselines are instantiated as follows. UDA aligns each variant's criterion distribution with the sharpened stop-gradient distribution of the original answer, using temperature T=0.4, confidence threshold τ=0.8, and unsupervised weight λ_u=1. R-Drop applies two dropout passes to the same input and uses a symmetric KL regularization weight of α=0.5. FixMatch converts the original answer's predictions into hard pseudo-labels for variants, using confidence threshold τ=0.8 and unsupervised weight λ_u=1. SMART perturbs the input embeddings within an ϵ=10^{-3} ball using one inner adversarial step, a symmetric-KL smoothness weight of λ_s=0.5, and the Bregman proximal term. Mean Teacher maintains an exponential-moving-average teacher with decay 0.99 and uses its detached criterion-level predictions on variants as training targets with weight λ=1. PPCL regularizes the divergence between the criterion-level losses of clean and perturbed answers with weight λ_u=1. GroupDRO performs worst-group optimization over five predefined groups (the original answer and four transformation families) with step size η=0.01 and receives full-group labels. CA uses a task-adapted implementation with full-group supervision that combines variant-augmented supervised fine-tuning with self-reward consistency alignment using λ_u=0.5, with the most confident variant serving as the instance's target. F^{2}C receives no direct correctness labels.

It forms a hard pseudo-label through majority voting within each equivalence instance and combines consensus cross-entropy with representation alignment, pulling lower-confidence and non-majority predictions toward the high-confidence consensus instance.

Table 8: **Baseline hyperparameter search.** Method-specific hyperparameters are tuned on the EduR^{2}Bench CS validation set (`Qwen3-8B`; original-only supervision unless noted) with a search budget at least as large as RITE's λ_{RI} sweep (Tab. 5), then frozen for all subjects and backbones. Shared settings (LoRA r=16, α=32, learning rate 1×10^{-4}, 5 epochs) are identical for all methods. ‡Uses full-group labels.

Method	Search space	Selected
UDA	$\lambda_u \in \{0.1, 0.5, 1, 2, 4\}$; $T \in \{0.4, 0.7, 1.0\}$; $\tau \in \{0.6, 0.8, 0.95\}$	λ_u=1; T=0.4; τ=0.8
R-Drop	$\alpha \in \{0.1, 0.5, 1, 2, 4\}$	α=0.5
FixMatch	$\lambda_u \in \{0.1, 0.5, 1, 2, 4\}$; $\tau \in \{0.7, 0.8, 0.95\}$	λ_u=1; τ=0.8
SMART	$\lambda_s \in \{0.1, 0.5, 1, 2, 4\}$; $\epsilon \in \{10^{-3}, 10^{-2}\}$; ascent steps $\in \{1, 2\}$	λ_s=0.5; ϵ=10^{-3}; 1 step
Mean Teacher	$\lambda \in \{0.1, 0.5, 1, 2, 4\}$; EMA decay $\in \{0.99, 0.999\}$	λ=1; decay 0.99
PPCL	$\lambda_u \in \{0.1, 0.5, 1, 2, 4\}$	λ_u=1
GroupDRO‡	$\eta \in \{0.001, 0.01, 0.1\}$	η=0.01
CA‡	$\lambda_u \in \{0.1, 0.5, 1, 2, 4\}$	λ_u=0.5

Inference-time strategies (Tab. 1b). All strategies use the untrained `Qwen3-8B` with T=0 unless noted. **Direct judgment** asks for a hit/miss decision per criterion with no intermediate output. **Checklist** presents the rubric as a numbered checklist and requests one decision per item; this is the prompt used by all fine-tuned graders and by Base. **Chain-of-thought** requests a short rationale before each decision. **Evidence-first** requests, for each criterion, the answer span that supports it before the decision, i.e., a prompt-only version of the evidence-anchoring step of RULERS (Hong et al., 2026), without its structured citation verification and evidence-conditioned score capping. **Invariance-aware prompt** adds an explicit instruction that order, length, style, and irrelevant additions must not affect the decision. **Self-consistency** samples k=5 decisions at T=0.7 and takes the per-criterion majority; it is the only strategy that uses sampling. The checklist prompt is given verbatim in the box below; the other strategies keep the same template and change only the instruction line as described above.

Grading prompt (checklist). Used by Base, all fine-tuned graders, and the frontier graders.

System.

```
You are a strict grader.  Apply the rubric exactly.  Judge each
criterion as hit or miss based only on the student answer, and give
the total.
```

User (template, filled per group member).

```
Question:
{question}
Rubric (criteria):
C1:  {criterion 1 description}
C2:  {criterion 2 description}
...
Student answer:
{answer text (original or variant)}
For each criterion output 'hit' or 'miss', then the total (number of
hits).
Output format:
C1:  hit/miss
C2:  hit/miss
Total:  <integer>
```

Each group member is graded in a separate call with no access to other members. The decision position for criterion k is the token that follows `Ck:`; under greedy decoding it is the argmax over the `hit`/`miss` candidates, and the hit probability $p_{i,k}$ used by $\mathcal{L}_{\mathrm{PBCE}}$ is read at the same position. The

reported total score is the number of predicted hits, so criterion decisions and total are consistent by construction; the emitted `Total` line is not used for scoring.

Variant generation. Position and Style variants are produced by deterministic scripts: Position re-orders sentences or clauses and is applied only when the answer has at least two such units and no subordinating conjunction; Style converts prose into a bulleted list. Length variants, Distractor sentences, and the initial criterion labels of each original answer are produced by `Fable 5` from the written protocol below (SciEntsBank version; the CS version is identical except for a 1.5–3.0× length target and CS-specific distractor topics); the Distractor sentence is then inserted into the original answer by script. Criteria for the SciEntsBank subjects were drafted by `Fable 5` from the reference answers and reviewed before annotation; the protocol forbids changing them.

Generation protocol given to `Fable 5` (one batch of groups per call).

```
INPUT: a batch JSONL, one group per line, fields: group_id, subject
    , question,
criteria (list of {cid, desc}), original (the student's original
    answer).

For EACH group produce THREE things (criteria are GIVEN -- do NOT
    change them):

## 1. fable_z -- judge the ORIGINAL answer against each given
    criterion
For each cid: does the ORIGINAL answer state/contain that content?
    1 = hit, 0 = miss.
- Judge semantically; obvious typos count as hits when intent is
    clear.
- Credit only content actually present; vague hand-waving that
    merely gestures at the
  topic is a MISS. Extra wrong content elsewhere does NOT cancel a
    hit.
- z_reason: one short sentence per cid quoting the evidence (or
    noting its absence).

## 2. length_variant -- meaning-preserving VERBOSE EXPANSION of the
     original
- Every original sentence must appear VERBATIM, character-for-
    character, IN ORDER --
  DO NOT fix typos/spelling/grammar; copy each original sentence
    exactly, then
  insert/append filler between or around them.
- Additions may ONLY restate content already present ("in other
    words...", "that is...",
  "so ...", generic meta). STRICTLY FORBIDDEN: introducing any NEW
    rubric-relevant
  content, especially content matching a criterion you judged 0 (
    that flips its label).
- No contradictions. Keep the original's register/format: if it has
     no newlines/bullets,
  add none.
- Target 1.6-2.5x the original word count (hard cap 3x); if
    original <= 8 words, add 6-25 words.

## 3. distractor_sentence -- ONE irrelevant, non-contradictory
    sentence
- {subject}-related plausible student filler, IRRELEVANT to THIS
    question's rubric:
  it must NOT contain content words overlapping ANY criterion desc,
     must NOT contradict
  the answer, must NOT answer the question.
- 5-40 words, single sentence, no newline. Make it distinct per
    group.
```

```
## OUTPUT
Write ONLY a JSON file (path given in your instructions):
{"items": [{"group_id": "...", "fable_z": {"C1": 1, ...}, "z_reason
    ": {"C1": "..."},
            "length_variant": "...", "distractor_sentence": "..."},
     ...]}
Include EVERY group of the batch exactly once. Self-check each item
     against the rules
before writing. Do not modify any other file.
```

Verification checklist. Every generated item passes an automatic verifier before manual review; any flag sends the item back for regeneration or, for CS, for written adjudication. The checks are listed below; content words are stemmed, stop-word-filtered tokens of length > 2.

Verification checklist applied to every variant.

Automatic checks on `Fable 5` items (Length, Distractor, labels).

L-A	Length, content preserved: every content word of the original occurs in the variant.
L-B	Length, sentences verbatim: every original sentence with ≥ 5 words is a verbatim substring of the variant.
L-C	Length, ratio: word ratio in $[1.4, 3.5]$, or the original has ≤ 8 words and 6–30 words are added.
L-D	Length, label hazard: no added word belongs to the vocabulary of a criterion judged *miss* (vocabulary of already-hit criteria exempt).
L-E	Length, format: no newline introduced if the original had none.
D-A	Distractor, shape: a single sentence of 5–40 words, no newline.
D-B	Distractor, irrelevance: no content-word overlap with any criterion description of the group.
D-C	Distractor, diversity: no exact duplicate distractor sentence across groups.
Z-A	Labels, completeness: the label set covers exactly the group's criteria with values in $\{0, 1\}$.

Automatic checks on script-generated variants. Position and Style: identical multiset of content words as the original. Distractor insertion: every original sentence remains verbatim in the variant.

Manual review (every variant and label). (i) The intended single factor was changed and nothing else; (ii) rubric-relevant evidence is preserved (no criterion content added or dropped; original sentences intact where the family requires it); (iii) the criterion-level gold state is unchanged (the variant inherits $\mathbf{z}^*$; Length additions never mention content of a criterion judged *miss*).

Records. CS: 24 automatic flags, each waived with a written reason (preposition collisions, idioms, glued typos); all 508 groups pass afterwards. SciEntsBank and RiceChem annotation batches (RiceChem is used only for the teaching-assistant agreement check and is not part of the benchmark): 47 of 684 items (6.9%) regenerated; no flag remains after the regeneration round.

Instruction given to the writers of the test-only human rewrites (Tab. 19).

```
You will see a short-answer question and a student's answer.  Rewrite
the answer as a different student who understands exactly the same
things would write it, in your own words.  Keep the same points.  Do
not add, remove, or correct anything related to the answer to the
question:  if the original is wrong or incomplete, your rewrite must
be wrong or incomplete in the same way; do not fix it, complete it,
or add hedges.  Do not consult any reference solution.  Do not copy
any sentence verbatim and do not merely swap a few words; change the
wording and sentence structure as another student naturally would.
```

Hardware and software. Our experiments were performed on the following servers.

- Operating System: Ubuntu 22.04.4 LTS
- CPU: AMD EPYC 7K62 48-Core Processor
- GPU: NVIDIA A100-SXM4 40GB GPUs
- Software: Python 3.10.20, PyTorch 2.7.1, CUDA 12.6.

A.5 SUPPLEMENTAL RESULTS

F^{2}C and Label-Free Consistency Methods. F^{2}C achieves unusually low CFR on several subjects, which, when viewed in isolation, could appear to indicate strong rubric invariance. However, as shown in Tab. 9, its QWK simultaneously approaches zero on Physics, Chemistry, and Earth Science. This discrepancy suggests that the low flip rate does not arise from reliable grading, but from a degenerate prediction regime.

Table 9: F^{2}C results across the four subjects. Although CFR is extremely low on several subjects, QWK simultaneously approaches zero, indicating that the apparent consistency is not accompanied by reliable grading.

Subject	QWK↑	C-F1↑	Err. SD↓	CFR (%)↓
CS	0.144 ± 0.041	0.596	0.561	5.8 ± 0.7
Physics	0.002 ± 0.003	0.371	1.055	0.6 ± 0.9
Chemistry	-0.009 ± 0.016	0.332	0.939	0.0 ± 0.0
Earth Science	0.000 ± 0.000	0.331	1.285	0.0 ± 0.0

To understand this behavior, we inspect the prediction dumps directly. The result is particularly clear on Chemistry. As shown in Tab. 10, two independent training seeds converge to opposite constant predictors: seed 42 predicts *miss* for all 353 criterion judgments, whereas seed 43 predicts *hit* for all 353 judgments. A third seed exhibits near-complete single-class collapse, predicting *miss* for 347 of 353 judgments. In contrast, the Base model and RITE retain non-degenerate prediction distributions.

Table 10: Prediction distributions on the Chemistry test set (353 criterion judgments). Two F^{2}C seeds converge to opposite constant predictors, while the third exhibits near-complete single-class collapse. Base and RITE retain non-degenerate prediction distributions.

Model	Prediction distribution	Majority class
F^{2}C (s42)	353 miss / 0 hit	100.0%
F^{2}C (s43)	0 miss / 353 hit	100.0%
F^{2}C (s44)	347 miss / 6 hit	98.3%
Base	266 miss / 87 hit	75.4%
RITE	174 miss / 179 hit	50.7%

The opposite directions of collapse are particularly informative. If the behavior were driven only by a dominant class prior, independent runs would be expected to favor the same class. Instead, two seeds converge to opposite constant solutions, while the third remains overwhelmingly concentrated on one class, despite optimizing the same agreement objective. This pattern is consistent with an underconstrained optimization problem in which agreement determines equality but not correctness.

This behavior follows directly from the structure of an agreement-only objective. Let x and x' denote two rubric-equivalent inputs, and let $p_\theta(\cdot \mid x)$ denote the grader's prediction distribution. An agreement objective encourages $D(p_\theta(\cdot \mid x), p_\theta(\cdot \mid x')) \to 0$, where D denotes a discrepancy measure between the two prediction distributions. However, in the absence of direct correctness supervision, this constraint specifies only that the two predictions should agree; it does not determine what their common prediction should be. In particular, a constant predictor $p_\theta(\cdot \mid x) = c, \forall x$, trivially satisfies $D(c, c) = 0$.

For binary criterion judgments, always predicting *hit* and always predicting *miss* therefore constitute two symmetric degenerate zero-loss solutions of the agreement term. Two of the three Chemistry runs empirically converge exactly to these opposite constant solutions, while the remaining run approaches the same degenerate regime without becoming perfectly constant.

The resulting metric pattern is self-consistent. A constant grader never changes its prediction within an equivalence group and therefore obtains CFR $= 0$. A near-constant grader similarly produces very few within-group flips and retains an extremely low CFR. At the same time, because such predictions carry little discriminative information about the gold grading labels, QWK approaches zero. The non-zero C-F1 values are consistent with strongly single-class prediction: the rarely or never predicted

class receives little or no F1, while the dominant class retains a non-zero F1 determined by the underlying class distribution. Similarly, when predicted scores exhibit little variation, variation in the gold scores is transferred directly into the error distribution, explaining why a near-zero CFR can coexist with a large Err. SD.

Importantly, this failure mode is not specific to F^{2}C. Our agreement-only adaptations, as well as an $\mathcal{L}_{\mathrm{RI}}$-only variant of our own objective, exhibit the same qualitative behavior. This localizes the failure to the absence of correctness grounding rather than to a particular implementation of F^{2}C. Agreement supervision constrains which predictions should remain equal, but does not identify what those predictions should be.

Table 11: **RI design across supervision regimes.** `Qwen3-8B` results on the EduR^{2}Bench CS validation set with T=0, averaged over three seeds. No design dominates every regime, while JS-total provides the best average balance and is used as the default.

RI design	Full-clean CFR ↓	ρ=10% CFR ↓	ρ=20% CFR ↓	Original-only CFR ↓	Avg. CFR ↓	Avg. QWK ↑
JS (total) ⋆	6.3	8.5	12.8	**4.8**	**8.1**	**0.728**
JS (criterion)	6.7	9.6	13.6	6.1	9.0	0.714
MSE (total)	**5.9**	10.8	14.9	6.6	9.6	0.708
W1-criterion	8.0	**7.8**	**12.4**	7.2	8.9	0.698

Table 12: **Loss-structure ablations.** `Qwen3-8B` results on the EduR^{2}Bench CS validation set with T=0, averaged over three seeds. 'SG' denotes 'Stop-gradient'. 'F-' denotes 'full-clean', whereas 'ρ' denotes 'ρ=10%'. The default design balances correctness grounding, rubric invariance, and alignment cost.

Design	Variant	F-QWK ↑	F-CFR(%) ↓	ρ-QWK ↑	ρ-CFR(%) ↓
SG	w/ SG ⋆	0.748	6.3	**0.706**	8.5
	w/o SG	**0.756**	**5.8**	0.486	**7.2**
Topology	anchor ⋆	0.748	6.3	**0.706**	**8.5**
	pairwise	**0.751**	**6.0**	0.692	8.9
Score loss	w/o $\mathcal{L}_S$⋆	0.748	**6.3**	–	–
	w/ $\mathcal{L}_S$	**0.764**	6.9	–	–

Equivalence verification. We verify each variant along three dimensions: the intended transformation is correctly applied, the rubric-relevant evidence is preserved, and the criterion-level gold rubric state remains unchanged. Variants that fail any check are regenerated or removed. Tab. 7 presents a complete example from the CS training split.

Loss Design Ablations. We study the divergence function, alignment granularity, and other loss designs in Tabs. 11 and 12. No design performs best in every regime. MSE-total achieves the lowest CFR under full-clean supervision, while W1-criterion performs best under 10% and 20% label noise. Comparing alignment granularity, JS-criterion remains competitive under full-clean supervision but consistently underperforms JS-total under original-only and noisy supervision. This suggests that score-level alignment provides a more robust equivalence signal when correctness supervision is sparse or unreliable. JS-total performs best under original-only supervision and provides the best average balance across regimes. We therefore use JS-total as the default design. The remaining ablations further show that lower CFR alone does not imply a better grader. Without stop-gradient, CFR is reduced under 10% label noise, but QWK drops substantially from 0.706 to 0.486, indicating weakened correctness grounding. Anchor-based and pairwise alignment perform comparably, while anchor-based alignment is more robust under label noise, reduces the group alignment cost from $O(M^2)$ to $O(M)$, and preserves the original response as a fixed reference. Adding the score loss L_S improves QWK but increases CFR, providing no joint gain in grading quality and rubric invariance. Based on these results, we retain score-level JS divergence, stop-gradient, original anchoring, and the simpler objective without L_S.

Loss-Scale Fairness Ablation. Because criterion-level and score-level alignment objectives may operate on different numerical scales, we examine whether the weaker performance of JS-criterion can be explained by the choice of λ_{RI}. We vary the RI weight over $\{0.3, 0.5, 1, 2, 4\}$ under original-only correctness supervision while keeping all other settings fixed. As shown in Tab. 13, JS-criterion achieves its best balance at $\lambda_{\mathrm{RI}} = 1$. Reducing the weight to 0.5 or 0.3 weakens both grading quality and invariance, while larger weights of 2 and 4 also degrade performance, indicating that overly strong criterion-level alignment is not beneficial. Importantly, even at its best setting, JS-criterion reaches a CFR of 6.1%, which remains higher than the 4.8% CFR of JS-total under the same validation setting. These results indicate that the weaker performance of criterion-level alignment is not an artifact of loss scaling or an unfair choice of λ_{RI}.

Table 13: **Loss-Scale Fairness Ablation.** `Qwen3-8B` results on the EduR^{2}Bench CS validation set under original-only correctness supervision with T=0. $\star$ marks the best setting for JS-criterion.

λ_{RI}	QWK ↑	CFR (%) ↓
0.3	0.682	9.0
0.5	0.721	7.1
1.0 $\star$	**0.744**	**6.1**
2.0	0.694	7.3
4.0	0.661	8.6

Table 14: **Hyperparameter sensitivity**. Hyperparameter sensitivity on the EduR^{2}Bench CS validation set under original-only correctness supervision. QWK and C-F1 are higher-is-better; Err. SD and CFR are lower-is-better. Results are averaged over three seeds. $\star$ marks the default setting.

Hyperparam.	Value	QWK ↑	C-F1 ↑	Err. SD ↓	CFR (%) ↓
Epochs	3	0.746	0.866	0.439	6.3
	5 $\star$	**0.760**	**0.879**	**0.414**	**4.8**
	7	0.751	0.872	0.425	5.0
Learning rate	5×10^{-5}	0.728	0.854	0.461	7.6
	1×10^{-4} $\star$	**0.760**	**0.879**	**0.414**	4.8
	2×10^{-4}	0.709	0.843	0.487	**4.5**
LoRA rank	8	0.744	0.865	0.441	5.7
	16 $\star$	**0.760**	**0.879**	**0.414**	**4.8**
	32	0.722	0.850	0.469	5.1

Hyperparameter Sensitivity. We vary the number of training epochs, learning rate, and LoRA rank on the validation set, as shown in Tab. 14. Increasing training from three to five epochs reduces CFR from 6.3% to 4.8% while also improving QWK, C-F1, and Err. SD. Extending training to seven epochs provides no further invariance gain and slightly degrades grading quality. A small learning rate does not learn sufficient rubric invariance. A larger learning rate further reduces CFR to 4.5%, but clearly harms grading quality, making 1×10^{-4} the best overall balance. For LoRA rank, $r = 16$ provides the best balance between grading quality and rubric invariance. Increasing the rank to 32 degrades both grading quality and CFR. We therefore use five epochs, a learning rate of 1×10^{-4}, and LoRA rank 16 in the main experiments.

Table 15: **CFR stratified by rubric size and transformation family** (`Qwen3-8B`, CS test, 3-seed means, %). Rubric-size columns report CFR over groups with K criteria. Family columns report family-specific CFR, which counts a flip only between the original answer and that family's variant (the definition of Fig. 2c and Tab. 17); valid pairs: Position 28, Length 160, Style 131, Distractor 160.

	Rubric size (CFR)			Family-specific CFR			
Model	K=1	K=2	K≥3	Position	Length	Style	Distractor
Base	23.8	26.9	30.0	10.7	10.6	13.7	19.4
PBCE-Orig	10.3	11.5	14.0	4.8	3.1	5.6	10.6
RITE	**4.4**	**5.1**	**6.7**	**1.2**	**2.5**	**3.3**	**1.9**

Stratified Rubric-Drift Analysis. We examine whether the aggregate CFR is dominated by particular rubric sizes or transformation families (Tab. 15). Drift grows with rubric size for every grader,

and both training regimes reduce it in every stratum. Across families, Distractor is by far the hardest for the untrained grader (19.4% versus 10.6–13.7% for Length and Style) and for original-only correctness supervision (10.6% versus 3.1–5.6% for the surface-form families), consistent with the transfer boundary in RQ3. RITE reduces family-specific CFR on all four families and removes the Distractor excess (1.9%), so its aggregate improvement is not driven by a single rubric size or family. Position has only 28 valid pairs and its values should be read with caution.

Table 16: **Criterion-level, pairwise, and score-level drift on the four test sets (%).** CFR is the any-pair criterion flip rate of Tab. 3. PFR (pairwise flip rate) is the fraction of member pairs, over all valid group–criterion pairs, whose hit/miss decisions differ; it is normalized by the number of pairs and therefore insensitive to group size. SFR (score flip rate) is the fraction of valid groups whose predicted total scores are not identical across members. For every trained grader the fraction of groups in which at least one criterion flips equals SFR, i.e., no group changes a criterion judgment while keeping its total score identical across all members (compensating flips between individual member pairs inside a group that also changes its total elsewhere are not excluded); for the untrained Base the group-level flip fraction exceeds SFR by 3.6 / 3.1 / 4.5 / 5.6 points on CS / Physics / Chemistry / Earth Science. Trained graders: mean ± SD over 3 seeds; Base: single greedy run. Best in **bold**.

	CS			Physics			Chemistry			Earth Science		
Grader	CFR	PFR	SFR	CFR	PFR	SFR	CFR	PFR	SFR	CFR	PFR	SFR
Base	26.3	13.4	26.8	25.6	14.6	39.1	32.1	15.8	36.4	26.0	13.4	36.1
PBCE-Orig	11.7±1.8	6.4±1.4	15.8±1.9	19.1±0.9	10.1±0.7	35.4±2.4	16.1±2.5	8.2±1.3	28.0±3.5	21.5±4.5	10.9±2.5	34.3±5.8
PBCE-All	5.8±1.6	3.0±0.7	8.3±2.3	12.4±1.2	6.5±0.7	25.5±2.4	12.8±0.9	6.4±0.6	22.0±1.3	16.2±2.8	8.1±1.4	27.8±5.6
RITE	**5.2±0.7**	**2.9±0.5**	**6.9±1.4**	**7.6±0.8**	**3.9±0.5**	**17.2±1.6**	**6.2±1.1**	**3.0±0.4**	**9.9±1.3**	**13.4±1.5**	**6.5±0.8**	**22.2±2.8**

Size-normalized and score-level drift. Tab. 16 reports, next to CFR, a pairwise-normalized rate and the score-level rate on all four subjects. PFR is 0.48–0.57 of CFR for every grader and subject (a flipping group–criterion pair typically contains a single dissenting member, which disagrees with about half of the other pairs) and preserves the ordering Base > PBCE-Orig > PBCE-All > RITE everywhere. SFR is a per-group quantity and is larger than the per-criterion CFR on every subject here, increasingly so in the subjects with longer rubrics (this is an empirical observation, not an identity: a group flips if any of its criteria flips); RITE has the lowest SFR in every subject and lies below the fully supervised PBCE-All (CS 6.9% vs. 8.3%, Physics 17.2% vs. 25.5%, Chemistry 9.9% vs. 22.0%, Earth Science 22.2% vs. 27.8%), which is the score-level invariance that the bound of Sec. 2 addresses. Groups whose criterion judgments change while every member's total score stays the same occur only for the untrained grader (3.1–5.6% of groups) and vanish under any correctness supervision, so for trained graders criterion-level and score-level drift move together and CFR remains the stricter, per-criterion diagnostic. On CS, the seed-paired differences make the complementarity result of Tab. 4 precise: adding $\mathcal{L}_{\mathrm{RI}}$ under original-only supervision lowers every measure in every seed (CFR -6.5 ± 1.6, PFR -3.5 ± 1.1, SFR -8.9 ± 1.8 points; QWK $+0.016$), whereas under full supervision it raises them slightly (RITE-All: CFR 7.7%, PFR 4.3%, SFR 11.0%; paired changes $+1.9\pm0.6$, $+1.3\pm0.2$, $+2.7\pm0.9$ points; QWK -0.051). Because all methods are evaluated on identical groups, any group-size effect on CFR is common to every grader, and PFR removes it. Finally, for the untrained grader on CS, moving the decision threshold on the hit probability from the greedy value $\tau=0.5$ to 0.3 or 0.7 changes CFR by at most 3.7 points (26.3% vs. 29.4% and 30.0%; 28.1% at 0.4 and 0.6) and C-F1 by 0.01, so the reported drift is not an artifact of the 0.5 cut.

RITE under a held-out transformation family. Tab. 17 reports the full leave-one-family-out results for both supervision regimes, and Tab. 18 the overall test metrics of the held-out graders. PBCE-Orig supervises only original answers, so removing a family's variants leaves its training set, and hence all its numbers, unchanged; the relevant original-only comparison is therefore RITE-Orig held out versus RITE-Orig seen, with PBCE-Orig as the no-equivalence reference. Position has only 28 valid pairs and is reported for completeness.

Human-written rewrites (test only). Tab. 19 evaluates the graders of Tab. 3, unchanged, on a test-only set of human-written rewrites of the original test answers. Two writers who had not worked on the benchmark each rewrote half of the 256 test originals from the question and answer alone, under the instruction reproduced in Appendix A.4, and a third person verified every rewrite against the rubric with the three-point checklist: 219 (85.5%) passed at first review, 37 were returned once

Table 17: **Leave-one-family-out training on CS with RITE.** Family-specific CFR (%) on the test variants of each family, for a grader that saw the family in training (*seen*) and one trained with that family removed (*held out*); mean ± SD over 3 seeds. Δ is the paired held-out reduction of RITE relative to PBCE (positive = RITE lower), with the number of seeds in which RITE is lower. PBCE-Orig supervises only original answers and never uses variants, so removing a family leaves it unchanged (†). Pairs: Distractor 160, Length 160, Style 131, Position 28 (Position is too small to interpret). Fig. 2c plots the PBCE-All rows of this table.

		PBCE		RITE		Held-out Δ	
Scope	Family	seen	held out	seen	held out	pp	wins
All	Distractor	1.25±0.63	6.46±0.95	1.46±0.36	**3.13±0.63**	+3.33±0.72	3/3
	Length	3.33±1.57	2.29±0.36	3.75±0.63	2.92±0.72	−0.63±1.08	0/3
	Position	1.19±2.06	1.19±2.06	2.38±2.06	2.38±2.06	−1.19±2.06	0/3
	Style	3.56±1.59	4.58±0.76	5.34±0.76	5.34±0.76	−0.76±0.00	0/3
Orig	Distractor	10.63±0.63	10.63±0.63†	1.88±0.63	**3.75±0.63**	+6.88±0.63	3/3
	Length	3.13±0.63	3.13±0.63†	2.50±0.63	2.71±0.36	+0.42±0.36	2/3
	Position	4.76±2.06	4.76±2.06†	1.19±2.06	1.19±2.06	+3.57±0.00	3/3
	Style	5.60±1.17	5.60±1.17†	3.31±0.44	3.82±0.76	+1.78±0.44	3/3

Table 18: **Overall test metrics of the held-out graders (CS, 3 seeds).** Holding out a family changes QWK by at most 0.015 relative to the corresponding seen grader (PBCE-All 0.746, RITE-All 0.695, PBCE-Orig 0.737, RITE-Orig 0.753). Overall CFR is the any-variant CFR of Tab. 3.

Method	Held out	Overall CFR (%) ↓	QWK ↑	C-F1 ↑	Err. SD ↓
PBCE-All	Distractor	9.79±2.19	0.736±0.034	0.858	0.444
PBCE-All	Length	5.83±2.01	0.738±0.032	0.859	0.440
PBCE-All	Position	5.83±2.60	0.743±0.039	0.860	0.438
PBCE-All	Style	6.88±2.17	0.738±0.028	0.859	0.441
RITE-All	Distractor	8.75±1.65	0.686±0.023	0.823	0.488
RITE-All	Length	7.50±1.08	0.688±0.014	0.829	0.488
RITE-All	Position	7.92±0.95	0.682±0.016	0.825	0.490
RITE-All	Style	8.33±1.91	0.691±0.026	0.825	0.481
RITE-Orig	Distractor	8.33±2.01	0.756±0.026	0.884	0.404
RITE-Orig	Length	6.25±0.63	0.749±0.033	0.882	0.407
RITE-Orig	Position	5.42±1.30	0.744±0.018	0.881	0.410
RITE-Orig	Style	6.46±2.01	0.753±0.015	0.882	0.411

(16 had corrected an error of the original, 11 dropped and 4 added rubric-relevant evidence, 6 were near-copies), and 5 were discarded after revision, leaving 251; on a double-verified subset of 128, two verifiers agreed on validity in 95.3% of cases (κ=0.84). The rewrites replace more of the original wording than any synthetic family (content-word Jaccard 0.47 vs. 0.55–1.00), with a word-level edit distance comparable to the Length family (0.48 vs. 0.52) at nearly unchanged length (word ratio 1.06). The three questions the family was designed to answer resolve the same way in all four subjects. The untrained grader drifts on natural rewrites (12.9–16.2% family-specific CFR); original-only correctness supervision reduces this by a quarter to a half (5.9–9.8%) although it never saw a rewrite, so supervision-induced invariance is not an artifact of the synthetic families; and RITE is lowest in every subject (3.3–7.4%), below the fully supervised PBCE-All, with the best or tied-best QWK and the best C-F1 among trained graders on the rewrites, so its benefit transfers outside the benchmark's transformation space. The frontier graders also drift on natural rewrites (5.6–11.8%); RITE's CFR is 1.3–2.4× lower. The writer effect is small (Base 13.4% vs. 13.9% on the two writers' rewrites). As a faithfulness check for the generated Distractor family, a writer also produced one rubric-irrelevant sentence under the generation protocol's constraints for 80 CS test groups (114 valid pairs); on the same groups, family-specific CFR on the human-written distractors is within 0.9 points of the generated ones for every grader (Base 18.4% vs. 19.3%, PBCE-Orig 9.9% vs. 10.5%, PBCE-All 1.5% vs. 1.2%, RITE 2.3% vs. 1.8%).

Table 19: **External check under natural variation: test-only human rewrites.** Every original answer in the four test splits was rewritten once by one of two writers who had not worked on the benchmark and who saw only the question and the original answer (no rubric, gold state, reference answer, or synthetic variant); rewrites preserve the original's errors and omissions and were verified by a third person against the rubric with the three-point checklist of Appendix A.4 (251 of 256 retained). Unlike the four synthetic families, a rewrite changes several surface properties at once, so its drift cannot be attributed to a single factor: it is an external check under natural variation, not a fifth attributable family, and no grader is trained on it. We report family-specific CFR (%, ↓; a flip is counted only between the original and its rewrite), QWK (↑) and C-F1 (↑) on the rewrites; the first two columns repeat the CS family-specific CFR of the single-factor Length and Style families (Tab. 15) as references. Trained graders: 3-seed means (± s.d. on CFR); Base and frontier graders: single greedy run (T=0) under the same checklist prompt. Best among trained graders in **bold**.

	Synthetic ref. (CS)		CS			Physics			Chemistry			Earth Science		
Grader	Length	Style	CFR	QWK	C-F1	CFR	QWK	C-F1	CFR	QWK	C-F1	CFR	QWK	C-F1
`gpt-5.2`	6.3	6.9	7.2	0.62	0.79	7.0	0.71	0.78	8.6	0.72	0.88	11.8	0.70	0.79
`claude-sonnet-5`	5.0	5.3	5.9	0.60	0.78	5.6	0.69	0.85	7.0	0.70	0.86	9.9	0.68	0.78
Base	10.6	13.7	13.1	0.31	0.61	12.9	0.46	0.72	16.2	0.31	0.63	13.0	0.62	0.79
PBCE-Orig	3.1	5.6	5.9±0.7	0.72	0.84	8.7±1.0	0.66	0.83	7.5±1.2	0.64	0.83	9.8±1.6	0.64	0.78
PBCE-All	3.3	3.6	4.4±0.8	**0.73**	0.84	6.9±0.9	0.67	0.83	7.0±0.8	0.69	0.85	8.8±1.3	0.61	0.78
RITE	**2.5**	**3.3**	**3.3±0.7**	**0.73**	**0.87**	**4.3±0.7**	**0.77**	**0.87**	**3.6±0.6**	**0.80**	**0.91**	**7.4±1.1**	**0.68**	**0.82**

Table 20: **Human validation of EduR^{2}Bench**. All variants are manually verified, and inter-annotator agreement is measured on a stratified subset of criterion-level decisions. An item is one group's `Fable 5` output (Length variant, Distractor sentence, and labels); the regeneration count covers the SciEntsBank and RiceChem annotation batches together, and RiceChem groups serve only the teaching-assistant agreement check.

Validation statistic	**Result**
Variants manually verified	3,297
Automatic verifier flags, CS (all resolved by written adjudication)	24 / 508 items
Items regenerated, SciEntsBank and RiceChem batches	47 / 684 (6.9%)
Flags remaining after the regeneration round	0
Criterion decisions in agreement study	500
Three-annotator agreement	481 / 500 (96.2%)
Decisions requiring adjudication	19 / 500 (3.8%)
Fleiss' κ	0.91
RiceChem TA agreement	95.0%

Human Review and Gold-Label Reliability. For data without existing human criterion annotations, we use `Fable 5` to obtain initial criterion-level hit/miss labels and manually review them before assigning the final gold rubric state. We also manually verify all 3,297 generated variants for transformation correctness, preservation of rubric-relevant evidence, and an unchanged criterion-level rubric state. Verification is performed by two authors, independently of the generation model, using a fixed checklist; variants that fail any check are regenerated or discarded. We further measure annotation agreement on a stratified sample of 500 criterion-level decisions from the four subjects. Three annotators label these decisions independently. As shown in Tab. 20, they agree on 481 of the 500 decisions (96.2%), while the remaining 19 decisions (3.8%) are resolved through adjudication. Fleiss' κ is 0.91. As an external check, Fable 5 agrees with the existing teaching-assistant annotations in RiceChem (Sonkar et al., 2024) on 95.0% of criterion-level decisions.

Cross-Backbone Replication. We replicate the main findings with `Llama-3.1-8B-Instruct` to examine whether they are specific to `Qwen3-8B`. We use the same data splits, evaluation protocol, and LoRA configuration as in the main experiments, and fix the learning rate to 5×10^{-5} based on validation performance of the supervised PBCE baseline, without method-specific retuning. As shown in Tab. 21, original-only correctness supervision substantially reduces CFR relative to the untrained model across all four subjects, confirming that supervision-induced invariance also emerges

Table 21: **Cross-backbone replication with `Llama-3.1-8B-Instruct`.** Results on the four EduR^{2}Bench subjects. Trainable methods are averaged over three seeds; Base denotes a single greedy evaluation of the untrained backbone. RITE achieves the best grading quality and rubric invariance across all four subjects.

	Computer Science				Physics				Chemistry				Earth Science			
Method	QWK ↑	C-F1 ↑	Err. SD ↓	CFR (%) ↓	QWK ↑	C-F1 ↑	Err. SD ↓	CFR (%) ↓	QWK ↑	C-F1 ↑	Err. SD ↓	CFR (%) ↓	QWK ↑	C-F1 ↑	Err. SD ↓	CFR (%) ↓
Base	0.40	0.65	0.62	37.5	0.43	0.68	0.96	31.9	0.24	0.55	0.79	38.1	0.60	0.74	0.86	33.0
PBCE-Orig	0.67±0.03	0.82	0.50	15.2±1.8	0.72±0.03	0.83	0.71	17.0±2.1	0.65±0.04	0.80	0.61	17.8±2.0	0.66±0.04	0.79	0.82	17.1±2.2
PBCE-All	0.70±0.02	0.84	0.47	9.2±1.3	0.74±0.02	0.84	0.68	11.4±1.6	0.69±0.03	0.83	0.57	12.1±1.5	0.69±0.03	0.81	0.79	13.4±1.7
RITE	**0.76±0.02**	**0.88**	**0.41**	**5.8±0.7**	**0.79±0.02**	**0.88**	**0.62**	**7.4±1.0**	**0.75±0.02**	**0.87**	**0.51**	**8.3±1.1**	**0.73±0.02**	**0.85**	**0.74**	**10.2±1.3**

with `Llama-3.1-8B-Instruct`. Increasing correctness-supervision coverage further improves both grading quality and rubric invariance. RITE consistently achieves the best overall performance across all four subjects, obtaining the highest QWK and C-F1 together with the lowest Err. SD and CFR. These results show that the gains of RITE are not specific to the Qwen backbone and that its advantage in jointly improving grading quality and rubric invariance generalizes across backbone families.

Consistency of LLM evaluators. A parallel line of work asks whether LLM evaluators are stable enough to stand in for human raters (Chiang & Lee, 2023). Such evaluators exhibit systematic preference biases that do not disappear with model scale (Koo et al., 2024), and LLM predictions more generally shift under meaning-preserving changes to input formatting, with substantial performance differences across formats that a task designer would treat as equivalent (Sclar et al., 2024). Consistency has therefore been treated as a model property in its own right, both as a measurement target under paraphrase and as a training objective (Elazar et al., 2021), and invariance tests are a standard component of behavioral evaluation suites for NLP models (Ribeiro et al., 2020). Our setting differs in what is held fixed and what is varied: the perturbation is applied to the graded student answer rather than to the prompt or instruction, the equivalence relation is verified to preserve both the rubric-relevant evidence and the gold rubric state, and the diagnostic is defined at the level of individual criteria. This distinction also motivates RITE's design: correctness labels specify the grading target, while the verified equivalence structure specifies which differences should leave that target unchanged.

A.6 Other Statements

Scientific artifacts used. We use publicly available educational datasets and language models for research purposes. EduR^{2}Bench is constructed from the Mohler short-answer grading dataset and SciEntsBank, covering computer science, physics, chemistry, and Earth science. Construct-irrelevant variants and initial criterion labels are generated with `Fable 5` and then manually reviewed. We fine-tune the open-weight backbones `Qwen3-8B` and `Llama-3.1-8B-Instruct`. Other open-weight and API-based models are used only for inference-time drift evaluation and receive no task-specific training. All datasets, models, and software artifacts are used in accordance with their intended research purposes and applicable licenses or terms of use. The source datasets contain previously released student responses. We collect no new private student data or personally identifiable information.

Reproducibility and artifact release. Upon publication, we will release datasets, source code, and experimental protocol on GitHub.

Computational budgets. The experiments were conducted on 4 NVIDIA A100 GPUs. All conducted experiments required at least 2,400 GPU-hours.

Ethical and broader impacts. Broadly speaking, this work may contribute to more reliable and equitable AI-assisted grading by reducing sensitivity to construct-irrelevant differences in student responses. More invariant graders may reduce inconsistent scoring across equivalent answers, although automated grading systems should still be deployed with appropriate human oversight.

Table 22: Main notation used in this study.

Symbol	Description
$\mathcal{D}_{\mathrm{tr}}$	Training set $\{(q_i, \mathcal{R}_i, \mathcal{A}_i, \mathbf{z}_i^*)\}_{i=1}^N$.
N	Number of training instances.
i	Instance index.
q_i	Question in instance i.
$\mathcal{R}_i$	Rubric $\{c_{i,k}\}_{k=1}^{K_i}$ associated with q_i.
$c_{i,k}$	The k-th binary rubric criterion in instance i.
K_i	Number of rubric criteria in instance i.
$\mathcal{A}_i$	Indexed answer collection $\{a_i^{(m)}\}_{m=0}^{M_i-1}$.
$a_i^{(m)}$	The m-th answer variant, where $a_i^{(0)}$ denotes the original answer.
m	Answer index in $\{0, \ldots, M_i - 1\}$.
M_i	Number of answers in $\mathcal{A}_i$, including the original answer.
$\mathcal{S}_i$	Indices receiving direct correctness supervision, $\mathcal{S}_i \subseteq \{0, \ldots, M_i - 1\}$; standard RITE assumes $0 \in \mathcal{S}_i$.
$\mathbf{z}_i^*$	Ground-truth grading-state vector $(z_{i,1}^*, \ldots, z_{i,K_i}^*) \in \{0,1\}^{K_i}$.
$z_{i,k}^*$	Ground-truth label for criterion $c_{i,k}$: 0 for `miss` and 1 for `hit`.
s_i^*	Ground-truth total score $\sum_{k=1}^{K_i} z_{i,k}^*$.
$\widetilde{\mathbf{z}}_i$	Corrupted grading-state vector used in the label-noise analysis.
$\widetilde{s}_i$	Corrupted total score $\sum_{k=1}^{K_i} \widetilde{z}_{i,k}$.
ρ	Criterion-level label-noise rate.
$\mathbf{x}_i^{(m)}$	Grader input triple $(q_i, \mathcal{R}_i, a_i^{(m)})$.
f_θ	Trainable LLM grader.
θ	Trainable grader parameters.
$p_{i,k}^{(m)}$	Predicted satisfaction probability for criterion $c_{i,k}$ on answer $a_i^{(m)}$.
$\mathbf{p}_i^{(m)}$	Criterion-probability vector $(p_{i,1}^{(m)}, \ldots, p_{i,K_i}^{(m)})$.
$\widehat{z}_{i,k}^{(m)}$	Binary criterion prediction $\mathbb{I}[p_{i,k}^{(m)} \geq 1/2]$.
$\widehat{s}_i^{(m)}$	Predicted total score $\sum_{k=1}^{K_i} \widehat{z}_{i,k}^{(m)}$.
$\mathrm{PB}(\cdot)$	Poisson–Binomial mapping from criterion probabilities to a distribution over total scores.
$\boldsymbol{\pi}_i^{(m)}$	Induced total-score probability vector $\mathrm{PB}(\mathbf{p}_i^{(m)}) = (\pi_{i,0}^{(m)}, \ldots, \pi_{i,K_i}^{(m)})$.
$\pi_{i,r}^{(m)}$	Probability that answer $a_i^{(m)}$ receives total score $r \in \{0, \ldots, K_i\}$.
$\boldsymbol{\pi}_i^*$	Ground-truth score distribution $\delta_{s_i^*}$.
$\mathfrak{P}_i$	Family of realizable score distributions $\{\mathrm{PB}(\mathbf{p}) : \mathbf{p} \in [0,1]^{K_i}\}$.
$\boldsymbol{\nu}$	Generic realizable score distribution in $\mathfrak{P}_i$.
δ_s	Point-mass distribution concentrated at total score s.
D_{JS}	Jensen–Shannon divergence between total-score distributions.
d_{TV}	Total variation distance, $d_{\mathrm{TV}}(\boldsymbol{\pi}, \boldsymbol{\pi}') = \frac{1}{2}\|\boldsymbol{\pi} - \boldsymbol{\pi}'\|_1$.
d_{H}	Hamming distance between binary grading-state vectors.
$\ell_i^{(m)}$	Mean per-criterion binary cross-entropy for answer $a_i^{(m)}$.
$\mathcal{L}_{\mathrm{PBCE}}$	Direct criterion-level correctness objective.
$\mathcal{L}_{\mathrm{RI}}$	Score-distribution rubric-invariance objective.
$\mathcal{L}_{\mathrm{RITE}}$	Joint objective $\mathcal{L}_{\mathrm{PBCE}} + \lambda_{\mathrm{RI}}\mathcal{L}_{\mathrm{RI}}$.
λ_{RI}	Weight assigned to $\mathcal{L}_{\mathrm{RI}}$.
$\mathcal{B}$	Mini-batch.
E	Number of training epochs.
η	Learning rate.
QWK	Rubric-size-bucketed and size-weighted quadratic weighted kappa for total-score agreement.
C-F1	Macro-F1 over criterion-level `hit`/`miss` predictions.
Err. SD	Standard deviation of predicted-minus-ground-truth total-score errors.
CFR	Criterion Flip Rate, the proportion of valid group–criterion pairs with at least one prediction flip across equivalent answers.
l	Transformer-layer index.
$\mathbf{h}_{i,k,l}^{(m)}$	Hidden state at layer l and the decision position for criterion $c_{i,k}$ on answer $a_i^{(m)}$.
$g_{i,k,l}^{(m)}$	Representation gap $1 - \cos(\mathbf{h}_{i,k,l}^{(m)}, \mathbf{h}_{i,k,l}^{(0)})$.