From Process Evidence to Policy Improvement: Auditing Reward Fidelity in All-Wrong GRPO Groups

Authors and affiliations: [TBD]

2026-09-12

Draft status: This manuscript describes a proposed research design. All empirical results are explicitly marked [TBD], without presupposing their direction. It is a complete-paper draft for discussing the argument and methodology, not a completed empirical study.

Abstract

Restoring a training signal in an all-wrong response group is not the same as improving the policy. In Group Relative Policy Optimization (GRPO), binary outcome rewards provide no within-group advantage when every response is incorrect, yet adding process scores can introduce preferences without an informative outcome anchor. We propose a framework for auditing reward fidelity in these groups: extract deterministically checkable evidence from task-provided structure, define reference alignment in the advantage and gradient spaces of the current policy, and test its relevance through held-out outcomes after actual parameter updates. We further describe an advantage-injection mechanism that separates process scores from their eligibility for use, allowing downweighting and abstention. The candidate training mechanism requires neither an additional learned evaluator nor a teacher model, and does not generate auxiliary continuations at each training iteration; bounded, independent continuations from the same policy are reserved for offline auditing. The experimental design examines the prevalence of analyzable natural all-wrong groups, incremental prediction under matched static quality, and whether gating provides value beyond coverage, random selection, and update magnitude. **[TBD: principal findings, effect sizes, confidence intervals, and scope.]** The central question is not how many nonzero advantages can be recovered, but which process evidence warrants influencing the current policy.

1 Introduction

When all eight responses to a question are wrong, are they equally uninformative for learning? Under binary outcome rewards, GRPO assigns them the same task advantage: zero. This follows from within-group comparison, not from an assessment of their reasoning quality. One response might select appropriate supporting facts but make an incorrect final inference, whereas another might depart from the question at the outset. A binary answer check cannot express that distinction. GRPO constructs a baseline from responses to the same question rather than a separate critic, exposing a signal gap when outcomes are degenerate. Shao et al. (2024)

Several approaches address this gap. Model confidence can distinguish negative samples, process scorers can supply local feedback for unsuccessful trajectories, and task-quality scores can separate responses with identical outcomes. These approaches show that a zero outcome advantage does not necessarily imply the absence of useful auxiliary information. However, restoring a preference

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.

answers whether an update can be produced, not whether that update should be applied. Feng et al. (2025), Agrawal et al. (2026)

We study the latter question. Once process rewards distinguish responses in an all-wrong group, their preferences determine which generation behaviors are reinforced. A response may score highly by including more reference facts without assembling an explanation that supports its conclusion. Conversely, a valid alternative explanation may receive a low score because it is absent from the annotations. Greater reward variation can therefore reflect either more useful information or stronger misleading preferences. Research on WorldTree explanation evaluation shows that a limited set of gold explanations does not exhaust valid explanations, constraining how overlap-based scores should be interpreted. Jansen et al. (2021)

Static scoring accuracy alone cannot fully characterize these preferences. Reinforcement learning uses relative advantages on samples generated by the current policy, rather than average scoring performance on a separate test set. The same number of scoring errors can induce different updates when those errors affect different responses, groups, or regions of the policy distribution. Conversely, disagreement with a reference scorer need not reduce final performance: the effect of reward error must be understood in the context of the policy and optimization procedure. Shang et al. (2026)

We therefore separate three levels of process-reward evaluation: **whether the evidence is checkable, whether its induced preference aligns with an outcome-based reference under the current policy, and whether the actual update improves task performance.** The first concerns what the program verifies; the second concerns how within-group differences enter the gradient; the third concerns whether the updated policy performs better under fresh sampling and on unseen questions. These levels are not interchangeable. Passing local checks does not establish directional correctness, and directional alignment does not guarantee robust gains at a finite step size.

This distinction motivates an eligibility audit for process rewards and its application to advantage injection in all-wrong groups. The framework first extracts process evidence from task-provided structures, such as supporting facts. A separate audit branch then measures group-relative reference alignment under the current policy. Finally, a rule-based score that does not access the expensive reference determines whether, and how strongly, process advantages are used. The training mechanism retains an abstention option: evidence outside the scope of the checks does not automatically induce a preference. We do not interpret the rule-based score as a reliability probability; whether it predicts and improves actual updates remains an empirical question.

The draft organizes three candidate contributions around this question:

1. **A layered problem definition.** We distinguish advantage recovery, reference-opposed updates, and actual outcome harm, establishing separately auditable objects for evaluating process rewards in all-wrong groups.

2. **A policy-relative audit protocol.** Under matched-static-quality controls, we combine current-policy gradients, independent continuation references, and paired actual updates to test which information incrementally predicts training consequences.

3. **Audit-qualified advantage injection.** We separate process scores from their eligibility for use, define a downweighting and abstention interface without an additional learned evaluator, and test its value against coverage, random-selection, and matched-update-magnitude controls.

The first two form the diagnostic contribution; the third is a method component that requires support from that diagnosis. If the experiments support the measurement protocol but not low-cost gating, the conclusions must remain diagnostic.

2 Problem Setting: Degenerate Outcomes and Partial Process Evidence

2.1 All-wrong groups in GRPO

Let x denote a question and π_θ a language-model policy parameterized by θ. For a fixed x, the sampling policy π_{θ_0} generates G responses, forming $\mathcal{G}_x = \{y_i\}_{i=1}^G$. An executable final-answer check assigns an outcome reward $o_i \in \{0, 1\}$. The outcome advantage is

$$a_i^o = \frac{o_i - \bar{o}}{s_o + \varepsilon}, \qquad \bar{o} = \frac{1}{G}\sum_{i=1}^{G} o_i. \tag{1}$$

Here, s_o is the within-group sample standard deviation and ε is a numerical stabilizer. If all responses are incorrect, every $o_i = 0$ and hence every $a_i^o = 0$. We use $D_x = \mathbf{1}[\sum_i o_i = 0]$ to indicate an all-wrong group. The candidate process term is active only when $D_x = 1$; mixed-outcome and all-correct groups retain the prescribed outcome-reward rule.

This degeneracy concerns the task-reward advantage, not necessarily the complete parameter update. KL regularization, entropy terms, weight decay, and historical momentum can still produce a displacement. Isolating the auxiliary process term therefore requires controlling these sources rather than merely checking whether parameters changed.

2.2 Process evidence is not complete reasoning ground truth

We consider settings with partial structure available from the task or dataset, such as retrievable supporting facts, valid action sets, or explicit relational constraints. A task *contract* specifies checkable rules over this structure: whether a fact identifier exists, whether evidence is duplicated, whether a representational transformation preserves identity, or whether a unit matches an available annotation. The checks distinguish pass, fail, and outside-scope outcomes rather than treating every unknown as an error.

Deterministic process evidence means that identical inputs receive identical check results; it does not establish complete semantic coverage. A valid fact identifier does not imply relevance, relevance does not imply sufficiency for the answer, and matching one annotated explanation does not identify all valid explanations. We therefore call the training-visible structure *process evidence*, not complete process ground truth.

2.3 From process scores to group-relative advantages

A process scorer assigns a scalar r_i to response y_i. When within-group variation is distinguishable, the standardized process advantage is

$$a_i^p = \frac{r_i - \bar{r}}{s_r + \varepsilon}. \tag{2}$$

$\bar{r}$ and s_r are the group mean and sample standard deviation of process scores. When s_r falls below a prespecified numerical tolerance, the process term is explicitly set to zero. We call a nonzero a^p in an all-wrong group *advantage recovery*. It establishes that the scorer distinguishes responses, but guarantees neither a nonzero aggregate gradient nor a task-beneficial preference.

This distinction motivates group-relative fidelity. The absolute score of a response does not determine its training effect: reinforcement depends on the other responses in its group, the normalization rule, and the parameter geometry of the current policy.

3 From Static Quality to Actual Update Consequences

3.1 Static accuracy does not specify an update direction

Static scoring accuracy is usually evaluated on a fixed labeled set. It aggregates over where errors occur, whereas policy updates depend on their locations. Responses have different probability gradients, so incorrectly promoting one response within a group may matter more than another scoring error. Similar static quality should therefore define a controlled comparison, not establish that two process rewards are equivalent for training.

To make this distinction explicit, let z_i be the mean score-function gradient over the evidence tokens of response i, restricted to the parameters that are actually trainable:

$$z_i = \frac{1}{T_i^{\mathrm{ev}}} \sum_{t \in \mathrm{ev}} \nabla_\theta \log \pi_\theta(y_{i,t} \mid x, y_{i,<t})|_{\theta=\theta_0} . \tag{3}$$

T_i^{ev} is the number of evidence tokens. Under this length-weighting convention, the process ascent direction is

$$g_p = \frac{1}{G} \sum_i a_i^p z_i. \tag{4}$$

Intuitively, a_i^p determines which responses are reinforced or suppressed, while z_i determines how reinforcing each response changes the parameters. Examining scores without their gradient directions omits part of the update.

3.2 Group-relative reference alignment under the current policy

Let a_i^{ref} be a group-relative advantage obtained from an independent reference. Using the same token span, parameter subspace, and length weighting, define $g_{\mathrm{ref}} = G^{-1} \sum_i a_i^{\mathrm{ref}} z_i$. When both gradients are nonzero and distinguishable from noise, their reference alignment is

$$F(\mathcal{G}_x; \theta_0) = \frac{g_p^\top g_{\mathrm{ref}}}{\|g_p\|_2 \|g_{\mathrm{ref}}\|_2}. \tag{5}$$

F describes a directional relationship conditional on the group, policy, and reference. It is not an intrinsic property of the scorer independent of the policy. We refer to it as *reference fidelity* for brevity. Undefined zero-gradient cases and reference uncertainty are reported separately rather than encoded as a numerical zero suggesting neutral reliability.

Writing $Z = [z_1, \ldots, z_G]$ gives

$$g_p^\top g_{\mathrm{ref}} = \frac{1}{G^2} (a^p)^\top K a^{\mathrm{ref}}, \qquad K = Z^\top Z. \tag{6}$$

This identity separates within-group advantages from policy geometry. Identical marginal score distributions do not determine which response receives each advantage, and advantage rankings alone do not determine the alignment induced by K. We use the identity to guide the audit, not as a new policy-improvement theorem.

3.3 Directional disagreement is not actual harm

We distinguish two outcomes. A **reference-opposed rescue** occurs when a process term yields a distinguishable nonzero direction with $F < 0$. An **outcome-harmful update** occurs when applying an update produces a distinguishable decrease in held-out task performance relative to fresh samples from the unchanged policy. The first is a directional quantity; the second is an empirical consequence. They should not share an unqualified false-rescue label.

Let $J(\theta)$ be the expected task success rate. Under differentiability and local smoothness, the first-order expansion for the actual parameter displacement $\Delta\theta$ is

$$J(\theta_0 + \Delta\theta) - J(\theta_0) = \nabla J(\theta_0)^\top \Delta\theta + O(\|\Delta\theta\|_2^2). \tag{7}$$

The reference gradient need not equal ∇J, and an Adam displacement need not be parallel to the unpreconditioned process gradient. Equation (5) therefore supplies a testable predictor of consequences, not a substitute for actual updates and reevaluation. The policy dependence of reward errors likewise cautions against interpreting reference disagreement as performance harm. Shang et al. (2026)

4 Method: Auditing Process Evidence and Qualifying Advantage Injection

4.1 Framework overview

The framework separates training from offline auditing. The training branch uses only current responses, task structure, and fixed rules to compute process scores and their usage weights. The offline branch constructs a reference from independent continuations under the same policy, measures the directional quantity in Equation (5), and tests whether the weights predict actual update consequences. This information boundary is part of the method: the deployed gate must not consult the offline reference to decide which groups to update.

Figure 1: Framework schematic—placeholder. Naturally sampled response groups enter from the left. The upper, training branch applies the outcome-degeneracy check, contract checks, process-advantage normalization, and qualification gate before the policy update. The lower, independent audit branch uses bounded continuation sampling, reference-direction measurements, and fresh held-out evaluation to assess the training branch. The reference does not feed back into the deployed gate. This conceptual figure contains no empirical values.

4.2 From task structure to unit-level process evidence

We instantiate the framework through supporting-fact selection in WorldTree. The actor receives a question, all answer options, and candidate facts retrieved without using the correct answer. It first selects fact IDs and then produces an answer. This interface separates checkable evidence selections from the final outcome without requiring an additional language model to translate free-form reasoning into a verifiable structure. Xie et al. (2020)

Let S_x be the annotated support set available during training, and let $U_{i,t}$ be the deduplicated set formed by the first t evidence selections in response i. The basic proxy is set F1, whose changes define unit-level increments:

$$r_{i,t} = \frac{2|U_{i,t} \cap S_x|}{|U_{i,t}| + |S_x|}, \qquad e_{i,t} = r_{i,t} - r_{i,t-1}, \qquad r_{i,0} = 0. \tag{8}$$

The final score is $r_i = \sum_t e_{i,t}$. Repeated evidence earns no additional reward, and unknown IDs are recorded as parsing failures. This proxy exposes the preferences induced by a deterministic scorer for inspection; it does not make the gold support set a complete semantic judge. Valid facts absent from the annotations can still reduce F1. Independent reference measurements and failure-case analysis must address this limitation.

In the primary instantiation, unit-level increments are aggregated into a sequence score, which yields the relative advantage in Equation (2). The process term applies only to evidence tokens, excluding final-answer tokens. The method therefore uses unit-level evidence, sequence-level advantages, and evidence-token injection. It does not claim to solve stepwise causal credit assignment.

4.3 From rule checks to usage qualification

We distinguish the process score r_i from the qualification of that signal for use in training. For each current group, the rule audit computes three quantities: annotated-evidence coverage c_x, stability of within-group preferences under meaning-preserving transformations s_x, and the pass rate on explicit structural negative controls n_x. Coverage is defined as

$$c_x = \frac{1}{G} \sum_{i=1}^{G} \frac{|U_i \cap S_x|}{\max(1, |U_i|)}. \tag{9}$$

Meaning-preserving transformations include whitespace normalization, bijective fact-ID renaming, and changes to candidate presentation order, with existing outputs mapped accordingly. The statistic s_x is the proportion of pairwise preference signs preserved across transformations. These checks rescore the same semantic objects; they do not regenerate responses or establish invariance of the model itself. The negative controls contributing to n_x require that adding duplicates not increase the set reward, unknown IDs not receive positive reward, and inconsistent mappings be rejected.

Let $b_x \in \{0, 1\}$ indicate whether the entire group is parseable, untruncated, and covered by all required checks. The initial audit weight is

$$q_x = b_x \, c_x \, s_x \, n_x. \tag{10}$$

The weight q_x is a computable rule-audit score, not an unbiased estimator of Equation (5) or a calibrated probability of beneficial updating. Whether it provides information beyond coverage is an empirical question. For a correctly canonicalized set-F1 scorer, s_x and n_x may both be identically one, reducing q_x to c_x on auditable groups. Such a result would mean that this instantiation has no independent mechanism for identifying reliability; the coverage-only control is designed to expose that possibility.

4.4 Audit-qualified process advantages

We call the gated process term a fidelity-qualified advantage (FQA). The name describes a mechanism intended to qualify signals through reliability auditing, not an established reliability guarantee. For token t in response i,

$$A^p_{i,t} = \lambda D_x q_x a^p_i M^{\text{ev}}_{i,t}, \tag{11}$$

where M^{ev} is the evidence-token mask and $\lambda \geq 0$ controls process strength. Process scores, advantages, and q_x are all treated as stop-gradient quantities. Groups lacking checkable evidence

receive $q_x = 0$, while groups that are not all-wrong receive $D_x = 0$. Abstention does not imply that these groups can never be useful; it means that the current checks do not justify using their process signal.

To make the normalization explicit, the process loss is averaged separately over evidence tokens. Let $\rho_{i,t} = \pi_\theta(y_{i,t} \mid x, y_{i,<t})/\pi_{\theta_0}(y_{i,t} \mid x, y_{i,<t})$ be the token probability ratio and ϵ_c the PPO clipping width. Then

$$\ell_x^p = -\frac{1}{G}\sum_i \frac{1}{T_i^{\text{ev}}}\sum_{t\in\text{ev}} \min(\rho_{i,t}a_i^p, \text{clip}(\rho_{i,t}, 1-\epsilon_c, 1+\epsilon_c)a_i^p)\,. \tag{12}$$

For a batch of B question groups, the full objective is

$$\mathcal{L} = \mathcal{L}_{\text{out}} + \frac{\lambda}{B}\sum_x D_x q_x \ell_x^p + \beta\mathcal{L}_{\text{KL}}. \tag{13}$$

Here, $\mathcal{L}_{\text{out}}$ is the outcome-reward policy loss over complete responses and excludes the separately listed KL term. The KL coefficient β is shared across comparison arms. The proposed change concerns eligibility for the auxiliary process term; it does not redefine all-wrong groups by adding a format reward.

The gate must act after the normalization in Equation (2). If all scores in a group are first multiplied by the same positive q_x, then, ignoring the numerical stabilizer, standardize$(q_x r)$ = standardize(r): normalization cancels the gate. Conversely, scaling a loss gradient does not guarantee proportional scaling of the Adam displacement. Actual update strength must therefore be measured separately.

4.5 Independent offline references and actual consequences

The offline audit truncates each response after its complete evidence prefix and before answer generation. It draws K independent continuations from the same initial policy and defines

$$\widehat{Q}_i = \frac{1}{K}\sum_{k=1}^{K} o_{i,k}^{\text{cont}}, \qquad a_i^{\text{ref}} = \frac{\widehat{Q}_i - \overline{\widehat{Q}}}{s_{\widehat{Q}} + \varepsilon}. \tag{14}$$

The variable $o_{i,k}^{\text{cont}}$ denotes the final outcome of continuation k. Neither the original incorrect answer nor the correct answer may be included in the continuation prompt. This reference estimates how likely the current model is to answer correctly after the given evidence; it does not determine whether that evidence is sufficient and correct for a human reasoner. We report finite-K binomial uncertainty, near-zero variation, and directional stability across two independent halves of the continuation samples.

Actual consequences are measured using cloned training states. Each arm starts from identical model parameters, adapters, and optimizer state, performs one update on the same initial batch, and answers held-out questions afresh. The primary endpoint is

$$\Delta J_{\text{heldout}} = \widehat{J}_{\text{heldout}}(\theta') - \widehat{J}_{\text{heldout}}(\theta_0). \tag{15}$$

The unchanged policy also receives fresh evaluation samples. The old outputs used to select all-wrong groups must not serve as the evaluation baseline: even without an update, resampling can appear to improve performance through regression to the mean. Actual displacement norm, KL divergence, and alignment with the reference are auxiliary endpoints, not substitutes for Equation (15).

4.6 Computational and information boundaries

The training branch deterministically parses and rescores existing responses a bounded number of times. It requires no additional learned evaluator, LLM judge, or teacher model. Existing annotations nevertheless constitute human supervision supplied by the task and must be disclosed as such. Offline continuations and gradient audits add research costs but are not per-iteration operations of the candidate gate. These costs are reported separately. Offline auditing is not free, and a gate that consults the offline reference for online selection cannot be described as reference-free at deployment.

5 Experiments: Three Questions for the Evidence

This section presents proposed experiments and reserves space for their results. Every [TBD] in a table denotes an unavailable measurement. The settings below are initial designs, not configurations that have already been executed.

5.1 Tasks, models, and evaluation principles

The primary task combines structured evidence selection with multiple-choice question answering in WorldTree. We use Qwen2.5-1.5B-Instruct as the proposed diagnostic model, with training and comparisons restricted to the same LoRA subspace. Candidate facts are retrieved symmetrically from the question and all options; neither gold answers nor gold explanations enter the actor's input. Data are partitioned by question and near-duplicate cluster, with separate training, calibration, and final-evaluation roles. WorldTree V2, official Qwen model card

The experiments begin by measuring the phenomenon in raw, unfiltered groups, proceed to reference audits and paired updates, and then extend to short training runs with independent seeds. During short-run training, each policy samples its own on-policy responses; only the one-step diagnostic arms with identical starting states share the initial responses. The primary comparison matches the total generated-token budget, including groups rejected by the gate. Update counts and GPU hours are reported separately.

The statistical unit is the independent question for prevalence estimates, the independent update trial for one-step consequences, and the training run or seed for algorithmic stability. Multiple responses within a group and multiple checkpoints within a run are not independent replications. We report effect sizes and paired hierarchical confidence intervals, counting undefined references and technical failures separately.

5.2 Question 1: How much analyzable process signal occurs in natural all-wrong groups?

This experiment asks whether enough relevant groups occur in the target setting to support the study; it does not assume that every model wastes large numbers of all-wrong groups. Under independent sampling for a given question, the all-wrong probability is $(1 - p_x)^G$, where p_x is that question's single-sample success probability. Population prevalence depends on the distribution of question difficulty and cannot be obtained by substituting the overall mean accuracy into this expression. We therefore report pre-filter denominators and successively count all-wrong, parseable, nonzero-process, and reference-distinguishable groups.

Table 1: Observability in naturally sampled groups. Raw groups provide the common starting population, and each column states its conditional denominator. Reference distinguishability is measured only in the sampled audit subset, not presented as an exact population prevalence.

Model / setting	Raw independent groups	All-wrong / raw	Parseable / all-wrong	Nonzero process signal / parseable all-wrong	Distinguishable reference / audited subset
Primary setting	[TBD]	[TBD]	[TBD]	[TBD]	[TBD]
Prespecified robustness setting	[TBD]	[TBD]	[TBD]	[TBD]	[TBD]

Results paragraph placeholder: [TBD: Report natural prevalence and intervals, the contribution of formatting and truncation errors, reference coverage, and whether the available sample supports subsequent analyses.]

We distinguish groups with no process-score variation from groups whose process scores vary but whose reference remains indistinguishable. The former limits the supply of training signals; the latter limits what can be established about those signals. Collapsing both into a single failure-to-rescue category would obscure the source of the limitation.

5.3 Question 2: At matched static quality, what predicts actual updates?

The natural comparison first matches scoring quality at the unit level. Several process proxies share the same fact-level judgments but aggregate them using precision, recall, or F1. Because their unit outputs are identical, their unit accuracy is identical on the same labeled set, while their sequence-level preferences may differ. This comparison isolates how the same unit judgments enter an update; it does not establish equal semantic quality for the complete sequence scorers.

Independent semantic labels are used only within their coverage. Unrated facts are not treated as negatives. If only structural-property labels are available, claims are restricted to matched structural scoring quality. Synthetic proxies are reported separately as additional controlled counterexamples. Their perturbations are designed on the calibration partition, without tailoring them against unseen questions or checkpoints.

Predictive comparisons start from static quality, annotation coverage, group variance, length, confidence, and update norm. We then separately add advantage-level reference agreement, reference-gradient fidelity, or the inexpensive q_x. The target is the independently evaluated consequence in Equation (15), with prediction-training and held-out sets separated by question cluster and update trial. Any low-dimensional regressor trained for this analysis remains outside the deployed gate and does not become an implicit learned evaluator.

Table 2: Incremental prediction of actual update consequences. All rows use the same held-out trials. Prediction error is primary; rank correlation is auxiliary. The table presupposes no winning predictor.

Predictive information	Static-matching level / gap	Held-out prediction error	Error change relative to static baseline, with CI	Rank correlation with actual consequences
Static quality and confounders	[TBD]	[TBD]	Reference	[TBD]
Add advantage-level reference agreement	Same setting	[TBD]	[TBD]	[TBD]
Add policy-gradient reference fidelity	Same setting	[TBD]	[TBD]	[TBD]
Add reference-free audit score q	Same setting	[TBD]	[TBD]	[TBD]

Results paragraph placeholder: [TBD: Establish whether incremental predictive value is observed, whether it comes from natural proxies or synthetic counterexamples, whether it transfers to held-out questions or policies, and whether reference error and actual effects are distinguishable.]

5.4 Question 3: Can simple explanations account for gating gains?

This experiment tests whether qualification auditing adds independent value. We compare outcome-only GRPO, uniform process rescue for all-wrong groups, coverage-only weighting, the full q gate, and randomly matched weights. The random arm permutes the gate's weight multiset across groups, separating useful group selection from a mere reduction in updating. An additional one-step, displacement-norm-matched control scales the uniform-rescue direction to the gated arm's parameter-displacement norm. We also report actual KL divergence: equal Euclidean displacement does not imply identical policy change.

Table 3: Task performance and cost under matched budgets. One-step geometric controls are kept separate from long-run optimizer comparisons. All short-run arms share a prespecified seed set, raw generated-token budget, and endpoint evaluation.

Training rule	Independent seeds	Held-out accuracy, with CI	Difference from uniform rescue	Raw generated tokens	GPU hours
Outcome-only	[TBD]	[TBD]	[TBD]	[TBD]	[TBD]
Uniform process rescue	[TBD]	[TBD]	Reference	[TBD]	[TBD]
Coverage-only	[TBD]	[TBD]	[TBD]	[TBD]	[TBD]
q gate / FQA prototype	[TBD]	[TBD]	[TBD]	[TBD]	[TBD]
Randomly matched weights	[TBD]	[TBD]	[TBD]	[TBD]	[TBD]

Results paragraph placeholder: [TBD: Report the main effect and interval, whether q meaningfully differs from coverage, whether random selection or matched update strength explains the result, and consistency across seeds.]

If q is exactly equivalent to coverage-only weighting, the resulting null ablation is not something to conceal. It is an informative outcome for the method hypothesis: the current contract adds no information for identifying update risk. Likewise, if reference fidelity predicts consequences but

inexpensive q does not, the diagnostic finding should be retained without attributing the expensive reference's value to the deployed mechanism.

5.5 Mechanism analysis and failure cases

The mechanism analysis examines four boundaries. First, how do group size and decoding temperature affect natural all-wrong prevalence and available process variation? Second, does removing or shuffling evidence change the model's answer distribution, indicating whether predictions depend on the scored content? Third, how are structurally valid but irrelevant facts and valid alternative explanations scored? Fourth, do abstention and high q correspond to distinguishable update consequences, rather than merely response length or formatting?

Figure 2: Training curves—placeholder. The horizontal axis is cumulative generated tokens, including rejected groups. The vertical axis is accuracy on a fixed held-out set. The figure will summarize every prespecified seed with uncertainty intervals. No simulated curves are included in this draft.

Figure 3: Audit measurements versus consequences—placeholder. Using trials excluded from tuning, plot reference fidelity or q against actual ΔJ. Mark reference-indistinguishable records, proxy identities, and actual update strength. A same-batch algebraic inner product must not replace independent measurements on the two axes.

Case-analysis paragraph placeholder: [TBD: Select real questions and responses from prespecified failure categories, with original IDs, scores, reference intervals, and update consequences. Do not fabricate illustrative empirical cases.]

6 Related Work

Degenerate groups and training-signal recovery. GRPO constructs a relative baseline from rewards for responses to the same question, and DeepSeekMath discusses both outcome and process supervision. LENS uses confidence-related rewards for negative samples to exploit all-wrong groups. VeriGate introduces a learned PRM when outcomes degenerate and uses cumulative subsequent process credit. Our distinction is not being the first to address all-wrong groups; it is treating evidence qualification and actual consequences after signal recovery as a separate object of study. Shao et al. (2024), Feng et al. (2025), Agrawal et al. (2026)

Same-outcome group splitting and update alignment. In the VLA setting, Prism-GRPO splits same-outcome groups using execution quality and provides alignment conditions under which an additional update preserves local ascent on the task objective. Directional alignment is therefore not a theoretical concept exclusive to this work. We focus on partial process evidence for LLMs: without access to a complete semantic oracle, how can a reference direction be measured, distinguished from actual consequences, and potentially replaced by an inexpensive audit? Deng et al. (2026)

Structured verification and process supervision. WorldTree supplies supporting facts and explanation structures for science questions, while subsequent expert-rating work documents the non-exhaustiveness of gold explanations. NormWorlds-CF uses solver evidence and metamorphic-relation rewards in executable rule worlds, providing another example of structural supervision without an LLM judge. We distinguish properties that these tasks make deterministically checkable from semantic properties that remain unidentified. A deterministic backend should not automatically be interpreted as complete reliability. Xie et al. (2020), Jansen et al. (2021), Zhang (2026)

Reward errors and actual utility. The consequences of imperfect rewards depend on the policy, motivating evaluation beyond static error counts. We adopt this perspective but do not declare an

update harmful merely because its direction disagrees with a reference. Fresh sampling, unchanged policy clones, and actual parameter-displacement controls distinguish proxy disagreement from deterioration in the final task. Shang et al. (2026)

7 Limitations and Discussion

First, what an audit score can identify depends on the information supplied by the task contract. A scorer that systematically favors irrelevant facts may still be invariant to whitespace, renaming, and permutation. Invariance alone establishes neither semantic correctness nor policy improvement. We therefore provide no guarantee that passing the checks makes an update safe. The interpretation of q must remain bounded by its coverage and empirical evidence.

Second, the WorldTree instantiation restricts the scope of the conclusions. Fact selection is not open-ended natural-language reasoning. Retrieved candidates constrain the evidence the model can express, and existing annotations may omit valid alternative paths. The current model may bypass the selected evidence when answering, while same-policy continuations may inherit the model's own knowledge deficits. These issues require case analysis and evidence interventions; a high reference-alignment score does not resolve them.

Third, directional quantities are defined only for the measured parameter subspace, length weighting, reference precision, and checkpoint. Findings in a LoRA subspace do not automatically extend to full-parameter training, and one-step effects do not guarantee long-term benefits. Limited questions, few seeds, or noisy references may support only an inconclusive finding—not evidence that the method succeeds or that the central hypothesis has been refuted.

Finally, requiring no additional learned evaluator does not mean requiring no supervision or computation. The task's supporting facts come from existing annotations, and offline continuations and gradient audits consume model computation. Both training costs and the diagnostic costs of developing the qualification audit must be disclosed, alongside compliance with the terms of the actual data package. The limits of rule checks also preclude treating this method as a general certification tool for trustworthy reasoning.

8 Conclusion

We reformulate the role of process rewards in all-wrong GRPO groups: beyond asking whether a nonzero advantage can be recovered, we ask which evidence deserves to influence the current policy update. By separating structural evidence, group-relative reference directions, and actual training consequences, we propose a falsifiable audit framework and a process-advantage injection prototype that permits downweighting and abstention. The central distinction is between scoring a response and qualifying its score for use, with both subjected to independent consequence evaluation. [**TBD: Summarize diagnostic value, method value, or scope limitations from actual results. Do not presuppose that gating outperforms the controls.**]

References

Aakriti Agrawal, Minghui Liu, and Furong Huang. VeriGate: Verifier-Gated Step-Level Supervision for GRPO, 2026. URL https://arxiv.org/abs/2605.30451.

Zeyun Deng, Yuzhe Lu, Yawei Wang, Linbo Liu, Qing Ping, Han Ding, Guande Wu, Panpan Xu, and Jun Huan. Prism-GRPO: Faster VLA Policy Optimization via Splitting Same-outcome Groups, 2026. URL https://arxiv.org/abs/2608.17423.

Yunzhen Feng, Parag Jain, Anthony Hartshorn, Yaqi Duan, and Julia Kempe. Don't Waste Mistakes: Leveraging Negative RL-Groups via Confidence Reweighting, 2025. URL https://arxiv.org/abs/2510.08696.

Peter Jansen, Kelly J. Smith, Dan Moreno, and Huitzilin Ortiz. On the Challenges of Evaluating Compositional Explanations in Multi-Hop Inference: Relevance, Completeness, and Expert Ratings. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7529–7542, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.596. URL https://aclanthology.org/2021.emnlp-main.596/.

Shuning Shang, Hubert Strauss, Stanley Wei, Sanjeev Arora, and Noam Razin. When Errors Can Be Beneficial: A Categorization of Imperfect Rewards for Policy Gradient, 2026. URL https://arxiv.org/abs/2604.25872.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models, 2024. URL https://arxiv.org/abs/2402.03300.

Zhengnan Xie, Sebastian Thiem, Jaycie Martin, Elizabeth Wainwright, Steven Marmorstein, and Peter Jansen. WorldTree V2: A Corpus of Science-Domain Structured Explanations and Inference Patterns supporting Multi-Hop Inference. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5456–5473, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.671/.

Xinqi Zhang. NormWorlds-CF: Solver-Verified Counterfactual Normative Reasoning with Metamorphic-Relation GRPO, 2026. URL https://arxiv.org/abs/2607.03957.

A Notation

Symbol	Meaning
$x, \mathcal{G}_x, G$	Question, response group for that question, and group size
$\theta_0, \theta, \theta'$	Sampling / initial parameters, optimization variable, and updated parameters
o_i, a_i^o, D_x	Binary outcome reward, outcome advantage, and all-wrong indicator
$S_x, U_{i,t}, e_{i,t}$	Annotated support set, selected evidence set, and unit-level increment
r_i, a_i^p	Process score and group-normalized process advantage
z_i, g_p, g_{ref}	Response score gradient, process direction, and reference direction
F, K	Reference fidelity and current-policy gradient Gram matrix
b_x, c_x, s_x, n_x, q_x	Auditability indicator, coverage, stability, negative-control pass rate, and audit weight
$\lambda, \beta, \epsilon_c$	Process strength, KL coefficient, and PPO clipping width
$M^{\text{ev}}, T_i^{\text{ev}}$	Evidence-token mask and evidence-token count
$\widehat{Q}_i, a_i^{\text{ref}}$	Independent continuation success rate and reference advantage
$J, \Delta J_{\text{heldout}}$	Expected task success rate and held-out performance difference under fresh sampling

B Algorithm

Inputs: Sampling policy, question batch, training-only support annotations, fixed contracts, and process weight. **Outputs:** Updated policy and complete audit records.

1. Sample G complete responses for each question. Save the raw outputs before computing final-outcome rewards.
2. Compute outcome advantages using Equation (1). Inspect process-use eligibility only for all-wrong groups.
3. Parse evidence units and compute the increments in Equation (8), process advantages in Equation (2), and q in Equation (10).
4. Set the process term to zero for groups with parsing failures, truncation, missing required checks, or no process-score variation, and record the reason.
5. Apply q after normalization. Compute Equation (12) with the evidence mask, combine it with the outcome term, and update the policy.
6. Log raw group denominators, process-signal recovery, weights, actual parameter displacement, and costs. Do not hide abstention costs through uncounted replacement sampling.
7. Separately construct the reference in Equation (14) on a presampled independent diagnostic subset, compute fidelity, and test actual consequences using cloned updates and fresh held-out sampling. These references are not available to Step 3.

C Initial Configuration and Reproducibility Placeholders

The following settings are design starting points. A completed manuscript must replace them with frozen configurations that were actually executed.

Item	Initial design
Model	Qwen2.5-1.5B-Instruct; revision [TBD]
Adaptation	BF16 + LoRA; rank 16, alpha 32, dropout 0
LoRA modules	Attention q/k/v/o projections; exact module names [TO VERIFY]
Prompt / output	Maximum total length 2048; maximum new tokens 256
Sampling	G8; temperature 0.8; top_p 0.95
Candidate facts	k24; select at most 8 IDs; no gold-based retrieval
Optimizer	AdamW; lr 1e-5; betas=(0.9,0.999); eps=1e-8
One-step diagnostic	Weight decay 0; KL coefficient 0; no entropy bonus; identical cloned initial states
Process / clipping	λ=0.1; PPO clip=0.2; gradient clip=1.0
Initial diagnostic partition	256 independent questions; A/B/C=128/64/64; updates use A only
Reference	K32 answer continuations; at most 64 new tokens per continuation; at most 32 audited groups
Seeds	E0 sampling: 17; candidate set for formal training: 17/23/41/59/83
Hardware	One RTX 5090 planned initially; at most two for independent parallel runs later; actual specifications and runtime [TBD]
Reproducibility assets	Code version, data hashes, complete commands, raw logs, and checkpoints [TBD]

All thresholds, comparison counts, and stopping rules must be frozen before the confirmatory experiments. The final test set must not inform gate, learning-rate, or prompt selection. Raw materials containing repeated questions must be deduplicated and partitioned by question cluster. Data permissions must be checked against the actual package; a code license is not a substitute for a data license.

D Mathematical Boundaries and Basic Checks

Equation (6) follows directly from $g_p = Za^p/G$ and $g_{\text{ref}} = Za^{\text{ref}}/G$. It describes the inner product between the defined directions; it does not establish that the reference equals the task gradient. Equation (7) requires suitable local smoothness around the current parameters. It does not justify declaring an update beneficial in the presence of large steps, optimizer preconditioning, or reference bias.

Equation (8) satisfies $\sum_t e_{i,t} = r_{i,T} - r_{i,0}$, so the unit-level increments in this prototype still aggregate to a sequence-level set score. This identity limits the granularity of any credit-assignment claim. Equation (10) also has a directly checkable degeneracy: when $s_x = n_x = 1$, q reduces to coverage on auditable groups. Any claim beyond coverage-only weighting requires additional, genuine information.

Implementation checks must cover zero outcome advantages for all-wrong groups, zero process advantages for constant scores, post-normalization q scaling that is not canceled by standardization, exclusion of answer and prompt tokens from the process mask, identical initial cloned parameters and optimizer states, and consistency between gradient direction and the loss sign. These checks validate implementation behavior; they are not empirical performance experiments.