

Understanding Training Data Attribution in RLVR: From the Perspective of Causal Inference

Chi Zhang^{1,2*}, Jian Zhao^{1*}, Wei Wang^{2†}, Kai Wang¹, Yao Zhao², Xuelong Li^{1,†}

¹Institute of Artificial Intelligence (TeleAI), China

²Institute of Information Science, Beijing Jiaotong University, Beijing, China

{chi.zhang, wei.wang, yzhao}@bjtu.edu.cn

{zhaoj90, wangk115}@chinatelecom.cn xuelong_li@ieee.org

Abstract

Training data attribution is increasingly used to filter, reweight, and prioritize rollouts in reinforcement learning with verifiable rewards (RLVR). However, these interventions often assume that individual samples, transitions, or trajectories can be attributed independently of the other data used in the same update. We show that this assumption is not merely a property of the attribution estimator, but a structural property of the training objective. We recast training data attribution as individual treatment effect estimation and use the Stable Unit Treatment Value Assumption (SUTVA) to determine when a proposed attribution unit is well-defined. This yields a simple criterion that attribution is only valid at the level where the algorithm assigns a stable contribution to that unit. Applying this criterion to RLVR, we show that the PPO objective admits trajectory-level attribution, whereas naive trajectory-only attribution is not well-defined for GRPO unless peer context is modeled. Experimental results support this distinction and provide preliminary evidence for group-level attribution under GRPO.

1 Introduction

Reinforcement Learning with Verifiable Rewards (RLVR) has emerged as a pivotal paradigm for unlocking the complex reasoning capabilities of large language models (LLMs), especially in domains such as mathematics and code where outputs can be automatically checked (Schulman et al., 2017; Shao et al., 2024). Unlike supervised fine-tuning, RLVR optimizes over sampled reasoning trajectories, allowing the model to explore alternative solution paths and learn from verifier feedback. Among recent RLVR methods, Group Relative Policy Optimization (GRPO) and its variants have become especially prominent because they avoid a learned

value model and estimate advantages from groups of sampled responses (Shao et al., 2024).

Recently, training data attribution is increasingly being used as an optimization tool rather than only a diagnostic tool. Recent methods score rollouts, transitions, or prompts by estimated influence and then use these scores to filter, reweight, replay, or prioritize training data (Hu et al., 2025b; Zhu et al., 2026; Shu et al., 2026). This line of work inherits the intuition from supervised learning attribution: if we can estimate which training units help or hurt a target objective, we can intervene on those units to make training more efficient.

As illustrated in Figure 1, this intuition relies on a structural assumption that is often left implicit: the attributed unit must be a meaningful unit of optimization. Under the trajectory-local PPO objective studied here, this assumption holds because each trajectory receives an advantage estimate computed independently from that trajectory’s own reward and value function. A trajectory-level influence score can therefore be interpreted as approximating the effect of changing the weight of that trajectory. However, this interpretation breaks down under group-relative optimization.

GRPO violates this condition because it does not assign credit to a trajectory in isolation. For each prompt, GRPO samples a group of responses and normalizes every reward against the rewards of the other responses in the same group. The resulting advantage of one trajectory therefore changes when its peers change, even if the trajectory itself is fixed. This group coupling means that the optimization effect of a trajectory is not an intrinsic property of that trajectory. And it is a property of the trajectory together with the sampled group.

This tension leads to our central question: **Can trajectory-level attribution remain well-defined under group-relative optimization without modeling peer context?** At first glance, the answer may appear to be yes. A trajectory-level score can

*Equal contribution.

†Corresponding authors.

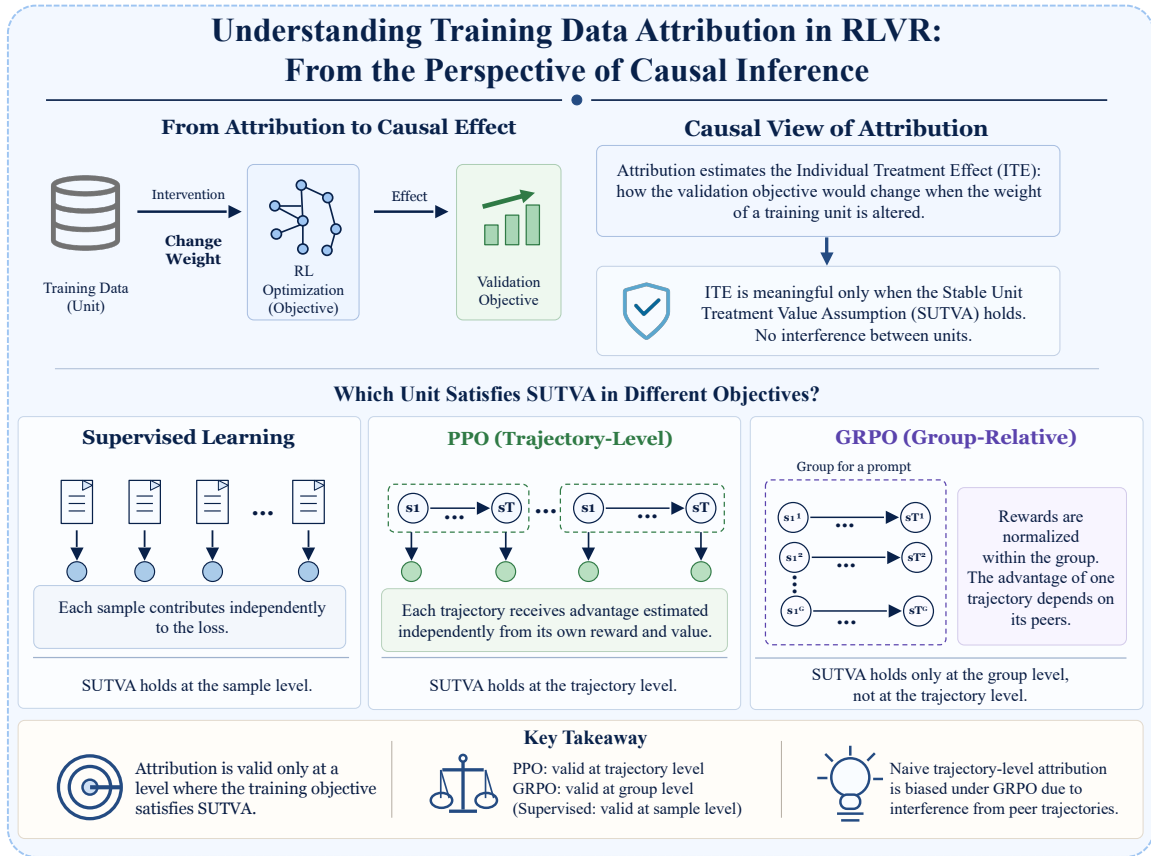


Figure 1: Overview of the attribution unit mismatch studied in this work. Under the trajectory-local PPO objective studied here, each trajectory receives an independently defined advantage, so trajectory-level attribution is well-defined. In GRPO-style optimization, the advantage of each trajectory depends on peer responses sampled for the same prompt, so the trajectory-only TD-ITE varies with peer context. Our framework formalizes this distinction through SUTVA and shows that the group level serves as the natural level of attribution for standard GRPO unless peer interference is explicitly modeled.

still correlate with short term validation loss or reward improvement, and such local property seems to justify filtering or reweighting individual trajectories. We show that this reasoning is misleading because a locally predictive score need not identify an actionable unit of optimization under GRPO.

We make this failure precise by recasting training data attribution as an individual treatment effect (ITE) problem from the perspective of causal inference. In this view, attribution asks how the validation objective would change under an intervention on the weight of a training unit. Such a per-unit effect is meaningful only when the unit satisfies the Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1986) that changing one unit’s weight should not change the contribution assigned to another unit. The trajectory-local PPO objective satisfies this condition at the trajectory level, whereas standard GRPO does not satisfy it because the contribution of each response is defined through

group-level reward statistics.

This formulation yields a simple criterion for choosing the right attribution unit: **unit-only attribution is valid only at a level where the training objective satisfies the SUTVA**. We show that supervised learning satisfies this condition at the sample level and the trajectory-local PPO objective satisfies it at the trajectory level, whereas standard GRPO satisfies it at the group level. We then prove that a trajectory-level influence score under GRPO is a biased estimator of the corresponding leave-one-out effect, with a bias term induced exactly by the peer trajectories used in group normalization.

Our experiments confirm this proposition by comparing the same attribution protocol under PPO and GRPO. Under PPO, where the trajectory is a valid attribution unit, trajectory-level reweighting behaves consistently with the attribution signal. Under GRPO, the same protocol becomes unreliable because scores change when only the

peer trajectories are resampled, rankings are unstable under regrouping, and trajectory-level filtering or reweighting does not yield consistent optimization gains. These results support the central claim that the problem is not attribution noise alone, but a mismatch between the attribution unit and the credit-assignment structure of the algorithm.

The key results and main contributions are three-fold:

- We identify attribution-unit mismatch as a structural failure mode in RLVR: trajectory-level attribution is well aligned with PPO-style objectives, but becomes ill-defined under standard GRPO.
- We formulate training data attribution as an ITE estimation problem and use SUTVA to characterize the valid attribution unit of a training objective. This framework classifies supervised learning, PPO, and GRPO-family objectives under a common perspective.
- We provide theoretical and empirical evidence that standard GRPO naturally supports unit-only attribution at the group level, and the trajectory-local PPO objective supports it at the trajectory level.

2 Related Work

2.1 RLVR

With the rapid advancement of large language models, reinforcement learning has become a crucial paradigm in their post-training (Ziegler et al., 2019; Stiennon et al., 2020; An et al., 2025; Bai et al., 2022). Early LLM reinforcement learning largely adopted PPO as the default optimizer for RLHF, following its role as a stable policy gradient method and its use in instruction following systems (Schulman et al., 2017; Ouyang et al., 2022). However, in reasoning oriented RLVR, verifiable outcome rewards and large rollout batches make critic-based PPO increasingly expensive. Thus GRPO becomes the core algorithm, which removes the learned value model by estimating advantages from multiple responses sampled for the same prompt (Shao et al., 2024). And DeepSeek-R1 further demonstrated that verifier driven RL can elicit strong reasoning behavior at scale (DeepSeek-AI, 2025). Recent work has therefore refined this critic-free, group-relative family: REINFORCE++ stabilizes critic-free optimization with global advantage normalization (Hu et al., 2025a), DAPO

improves large scale GRPO training through dynamic sampling and clipping changes (Yu et al., 2025), Dr. GRPO analyzes biases in R1-zero-like training (Liu et al., 2025), and GSPO moves optimization to the sequence level (Zheng et al., 2025). Our work studies this same shift toward group-relative RLVR, but asks a different question: what attribution unit is valid when the optimizer itself assigns credit through a response group?

2.2 Data Attribution for RL

Training data attribution traces model behavior or validation performance back to the data used during training. Classical methods such as influence functions and TracIn estimate the effect of individual training examples through either local sensitivity at an optimum or gradient similarity along the training trajectory (Koh and Liang, 2017; Pruthi et al., 2020). These methods were developed primarily for supervised learning, where examples enter the objective as separable terms. RL complicates this picture because the data distribution changes with the policy, and because a training unit may be a transition, a trajectory, a prompt, or a rollout group rather than a static example. Hu et al. (2025b) adapt attribution to online RL by scoring recent rollout-buffer records and using the scores for influence based filtering. Following this line of research, subsequent works tailored attribution approaches to the distinct phases of RLVR for LLMs. For example, CROPI (Zhu et al., 2026) utilized off-policy influence guidance for data selection to enable more efficient RLVR training. Operating at the trajectory level, IPPO (Shu et al., 2026) leverages influence based attribution to filter or reweight transitions to improve training efficiency. These works show that attribution can be used inside RL training loops, but it does not by itself specify which unit remains valid under a given RL objective. We further study this question in RLVR, where PPO-style and GRPO-style objectives provide a concrete setting for analyzing when trajectory-level attribution is well-defined.

2.3 Causal Inference and Interference

Causal inference provides the tool for asking when a unit-level effect is well-defined. In the potential outcomes framework, causal effects are defined by comparing the outcomes that would be observed under different treatments for the same unit (Rubin, 1974; Holland, 1986). This view underlies a large body of work on treatment effect estima-

tion, including propensity score methods for observational studies (Rosenbaum and Rubin, 1983). A key assumption in this framework is that each unit’s potential outcome is unaffected by the treatment assignment of other units, commonly formalized through SUTVA (Rubin, 1986). When this assumption fails, causal effects are no longer purely unit-wise objects. Prior work on interference studies this setting in randomized and observational designs, including housing-mobility experiments, partial-interference models, and more general exposure mappings (Sobel, 2006; Hudgens and Halloran, 2008; Tchetgen and VanderWeele, 2012; Aronow and Samii, 2017). We borrow this perspective for training data attribution. The treatment is the weight assigned to a training unit, and interference arises when the contribution of one unit depends structurally on other units in the same optimization group.

3 A Causal Inference Perspective on Training Data Attribution

In this section, we first give a causal interpretation of training data attribution. The key idea is to view the weight assigned to a training unit as a treatment and the validation objective as the outcome. Under this view, attribution estimates how the outcome would change if we intervened on the weight of a unit. We then identify the no-interference condition required for such a unit-level effect to be well-defined, and use it to determine the valid attribution unit of different training objectives.

3.1 Training Data Attribution as an Individual Treatment Effect

Training data attribution methods aim to answer a counterfactual question: *how would the validation loss at z_v change if training sample z_i were removed or upweighted?* To formalize this as a causal estimand, we adopt the standard reweighting parameterization (Koh and Liang, 2017). Let $\mathcal{D} = \{z_1, \dots, z_n\}$ denote a training dataset and $L_{\text{val}}(\theta; z_v)$ a validation loss at z_v . We treat each training sample as a unit whose treatment is its inclusion weight $w_i \in \mathbb{R}$, and consider the weighted training problem:

$$\begin{aligned} L_\epsilon(\theta; w) &:= \sum_{i=1}^n w_i \cdot \ell(z_i, \theta), \\ \theta^*(w) &:= \arg \min_{\theta} L_\epsilon(\theta; w), \end{aligned} \quad (1)$$

where $w = \mathbf{1}$ denotes the standard training process $\hat{\theta} := \theta^*(\mathbf{1})$, $w = \mathbf{1} - \mathbf{e}_i$ corresponds to leave-one-out retraining, and $w_i = 1 + \epsilon$ corresponds to infinitesimal upweighting of training sample z_i .

The reweighting parameterization in Eq. 1 recasts removal or upweighting as an intervention on the inclusion weight w_i . In the potential outcomes framework (Rubin, 1974), each training sample is a unit, w_i is a continuous treatment, and $L_{\text{val}}(\theta^*(w); z_v)$ is the outcome under treatment vector w . The infinitesimal unit-level effect at the observed training run is therefore the derivative of the validation loss with respect to w_i . We call this quantity the *training data individual treatment effect* (TD-ITE):

Definition 1 (TD-ITE) For training sample z_i and validation point z_v ,

$$\text{ITE}(z_i; z_v) := \left. \frac{\partial L_{\text{val}}(\theta^*(w); z_v)}{\partial w_i} \right|_{w=\mathbf{1}}. \quad (2)$$

For standard supervised objectives, $\text{ITE}(z_i; z_v)$ is the local first-order version of the discrete leave-one-out retraining effect $L_{\text{val}}(\theta^*(\mathbf{1} - \mathbf{e}_i); z_v) - L_{\text{val}}(\hat{\theta}; z_v)$. This is the causal object that many attribution methods implicitly approximate. Different attribution methods estimate this same underlying quantity with different approximations:

- **Influence functions** (Koh and Liang, 2017) use the implicit function theorem at the trained optimum, yielding a closed-form estimator in terms of the validation gradient and the inverse Hessian of the training loss.
- **TracIn** (Pruthi et al., 2020) accumulates per-step inner products of training and validation gradients along the SGD trajectory.
- **TRAK** (Park et al., 2023) combines random projections of per-sample gradients with a kernel-based scoring rule, scaling influence-style attribution to large models.

Despite these methodological differences, all three can be viewed as estimators of the TD-ITE in Eq. 2. Training data attribution can therefore be formalized as individual treatment effect estimation.

3.2 SUTVA as the Identifiability Prerequisite

The TD-ITE in Eq. 2 is a per-unit effect. Such an effect is meaningful only if the chosen unit is a stable optimization unit. This means that intervening

on one unit’s weight should not change the contribution assigned to another unit. In causal inference, this is the no-interference component of the *Stable Unit Treatment Value Assumption* (SUTVA) (Rubin, 1986). The SUTVA requires that unit i ’s potential outcome depend only on its own treatment T_i , not on the treatment assignments of other units.

In the training data attribution setting, the validation loss is the common outcome shared by all training units. We therefore express SUTVA as a no-interference condition on the unit contribution inside the weighted training objective. After differentiating the objective with respect to a unit’s intervention weight, the resulting contribution must be determined by that unit and the common model state alone.

Definition 2 (SUTVA for Data Attribution) *Let $\mathcal{U} = \{u_1, \dots, u_m\}$ be a candidate attribution unit level, and let w_{u_i} denote the intervention weight assigned to unit u_i . A training regime $(\mathcal{A}, L_\epsilon)$ satisfies SUTVA at level $\mathcal{U}$ if, for every unit $u_i \in \mathcal{U}$, there exists a function g_{u_i} such that*

$$\left. \frac{\partial L_\epsilon(\theta; w)}{\partial w_{u_i}} \right|_{w=1} = g_{u_i}(z_{u_i}, \theta), \quad (3)$$

where g_{u_i} depends only on the data of unit u_i and the model state θ , and not on the data or weights of any other unit u_j with $j \neq i$.

Definition 2 gives a criterion for when an attribution unit is well-defined. We call attribution at level $\mathcal{U}$ *valid* if the TD-ITE of each unit $u_i \in \mathcal{U}$ is determined by z_{u_i} and the common trained state, rather than by the other units that happen to share the same update or group. In this paper, “valid attribution” strictly refers to the unit-only TD-ITE that is invariant to peer context. This definition applies solely to attributions interpretable as individual causal effects. And it does not deny that the utility of cross-unit attribution methods or heuristic scores for ranking may still be useful.

Proposition 1 (Validity of Per-Unit Attribution) *Under the regularity conditions stated in Appendix D, a training regime $(\mathcal{A}, L_\epsilon)$ admits valid unit-only attribution at level $\mathcal{U}$ if and only if it satisfies SUTVA (Definition 2) at that level. When SUTVA holds, the ITE for each unit $u_i \in \mathcal{U}$ is given by*

$$\text{ITE}(z_{u_i}; z_v) = -\nabla_\theta L_{\text{val}}(\theta^*; z_v)^\top H_{\theta^*}^{-1} \nabla_\theta g_{u_i}(z_{u_i}, \theta^*). \quad (4)$$

When SUTVA fails, the TD-ITE at level $\mathcal{U}$ depends on the data of other units in the interfering cluster

containing u_i , so no estimator that only observes z_{u_i} can identify the unit-level effect.

The forward direction follows the standard influence function derivation of (Koh and Liang, 2017). When the unit contribution is separable, differentiating the optimality condition yields Eq. 4. The converse follows the logic of partial interference (Hudgens and Halloran, 2008). If the contribution assigned to u_i changes with peer units, then two contexts with the same z_{u_i} can induce different TD-ITEs. Thus, when SUTVA fails, attribution at that level is not merely noisy, the unit-level estimand is undefined unless the interfering context is included in the unit. (Full proofs in Appendix D.)

3.3 SUTVA of Common Training Regimes

Proposition 1 reduces the question “at what level is attribution valid?” to “at what level does SUTVA hold?” This level is determined by the structure of the training objective, not by the attribution estimator alone. As a result, the same trajectory-level score can estimate a stable unit-only effect under one algorithm but become invalid under another. Table 1 summarizes this distinction.

4 Trajectory-Level SUTVA in RL Training

We now apply the framework to two representative RL training regimes: GRPO (Section 4.1) and PPO with per-trajectory GAE (Section 4.2). They differ in the level at which the advantage is defined. In GRPO, the advantage of a trajectory is normalized against peer responses from the same prompt group. In PPO, the advantage is computed from the reward of the trajectory itself and a value function. This yields the central contrast that **trajectory-level attribution is valid for PPO-style objectives, but not for GRPO-style objectives**.

4.1 GRPO: A SUTVA-Failure Case

We first formalize GRPO in the weighted form of Eq. 1. For a prompt-conditioned group $g = \{\tau_1, \dots, \tau_n\}$ with trajectory rewards $\{r_i\}_{i \in g}$, GRPO computes group statistics and assigns each trajectory a normalized advantage:

$$\mu_g = \frac{1}{n} \sum_{j \in g} r_j, \sigma_g = \sqrt{\frac{1}{n} \sum_{j \in g} (r_j - \mu_g)^2}, \quad (5)$$

$$A_i(g) = \frac{r_i - \mu_g}{\sigma_g + \epsilon}.$$

Table 1: SUTVA structure of common training regimes for training data attribution. The atomic unit at which per-unit attribution is valid coincides with the level at which SUTVA holds. Supervised learning losses are decomposable across samples, whereas PPO’s advantage depends solely on the current trajectory. Therefore, SUTVA holds at the sample level and trajectory level, respectively. In standard GRPO, the advantage depends on peer rewards. Consequently, SUTVA does not hold at the trajectory level, but rather is satisfied exclusively at the group level.

Regime	L_ϵ structure	SUTVA holds at	Attribution valid at	Work family
Supervised learning	$\sum_i w_i \ell(z_i, \theta)$	per-sample	per-sample	(Koh and Liang, 2017; Ilyas et al., 2022)
PPO family	$\sum_\tau w_\tau A_\tau(\theta) \log \pi(\tau)$	per-trajectory	per-trajectory	(Hu et al., 2025b; Shu et al., 2026)
GRPO family	$\sum_g L_g(\{w_i, r_i\}_{i \in g}, \theta)$	per-group only	per-group only	(Zhu et al., 2026; Yang et al., 2026)

The corresponding weighted GRPO loss is

$$L_\epsilon^{\text{GRPO}}(\theta; w) = - \sum_g \sum_{i \in g} w_i A_i(g) \log \pi_\theta(\tau_i). \quad (6)$$

which recovers the standard GRPO objective at $w = \mathbf{1}$.

This objective does not satisfy SUTVA at the trajectory level.

Proposition 2 (GRPO violates SUTVA) *For the weighted GRPO loss in Eq. 6, for any trajectory τ_i in group g ,*

$$\left. \frac{\partial L_\epsilon^{\text{GRPO}}(\theta; w)}{\partial w_i} \right|_{w=\mathbf{1}} = -A_i(g) \log \pi_\theta(\tau_i), \quad (7)$$

where $A_i(g)$ depends on the rewards $\{r_j\}_{j \in g}$ through μ_g and σ_g . The contribution assigned to τ_i therefore changes when its peer trajectories change. Consequently, GRPO does not satisfy SUTVA (Definition 2) at the trajectory level.

Proposition 2 and Proposition 1 together imply that naive per-trajectory attribution in GRPO is not well-defined as an individual causal effect independent of the peer group. The TD-ITE of τ_i is not a function of τ_i alone, but of the entire prompt conditioned group containing it. By contrast, if the whole group is treated as the unit, the contribution is determined by that group and SUTVA is restored.

4.2 PPO: A SUTVA-Holding Case

We next consider PPO with per-trajectory GAE and no cross-trajectory normalization as the contrasting case. For a batch of trajectories $\{\tau_1, \dots, \tau_N\}$, the weighted PPO loss is

$$L_\epsilon^{\text{PPO}}(\theta; w) = - \sum_{i=1}^N w_i A_{\tau_i}(\theta) \log \pi_\theta(\tau_i), \quad (8)$$

where $A_{\tau_i}(\theta)$ is the generalized advantage estimate (Schulman et al., 2015) computed from trajectory τ_i :

$$A_{\tau_i}(\theta) = \sum_{t=0}^{T_i-1} (\gamma \lambda)^t \delta_t^{\tau_i}, \quad (9)$$

$$\delta_t^{\tau_i} = r_t^{\tau_i} + \gamma V_\theta(s_{t+1}^{\tau_i}) - V_\theta(s_t^{\tau_i}).$$

The key contrast with GRPO is that A_{τ_i} is computed from τ_i and the common model state θ only. It does not depend on the rewards, states, or weights of any other trajectory τ_j with $j \neq i$.

Therefore, under this decomposed objective, PPO satisfies SUTVA at the trajectory level.

Proposition 3 (PPO satisfies SUTVA) *For the weighted PPO loss in Eq. 8, for any trajectory τ_i ,*

$$\left. \frac{\partial L_\epsilon^{\text{PPO}}(\theta; w)}{\partial w_i} \right|_{w=\mathbf{1}} = -A_{\tau_i}(\theta) \log \pi_\theta(\tau_i), \quad (10)$$

where $A_{\tau_i}(\theta)$ depends only on τ_i and the common model state θ , the contribution assigned to τ_i is invariant to changes in other trajectories. Consequently, PPO satisfies SUTVA (Definition 2) at the trajectory level.

In contrast to GRPO (Proposition 2), PPO assigns each trajectory a stable contribution. Therefore, PPO is suited for trajectory-level attribution.

5 Empirical Experiments

5.1 Experimental Setup

Model and Dataset. We fine-tune Qwen2.5-1.5B-Instruct (Yang et al., 2024) and Gemma 3-1B (Team, 2025) on the GSM8K math reasoning benchmark (Cobbe et al., 2021) using the ver1 framework. Training uses all 7,473 GSM8K training problems, and evaluation is performed on the 1,319 test problems. We evaluate model performance using GSM8K Acc@1 at each training

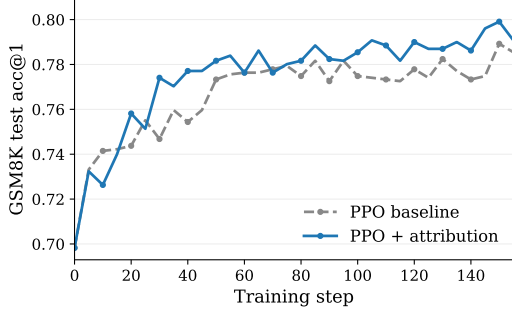


Figure 2: Trajectory-level attribution is effective under PPO. PPO with attribution guided reweighting improves sample efficiency over the PPO baseline, which is consistent with our framework’s prediction that the studied PPO objective supports trajectory-level attribution.

checkpoint using sampling temperature 1.0. To verify the consistency across different test datasets, we further evaluate the trained checkpoints on MATH500 (Hendrycks et al., 2021).

Training Regimes. We compare a matched set of four training configurations: PPO, PPO with attribution guided reweighting, GRPO, and GRPO with attribution guided reweighting. The PPO runs use per-trajectory GAE, while the GRPO runs sample $n = 5$ responses per prompt and compute group relative advantages. All runs follow the standard ver1 implementation, using KL regularization and a rule based verifier based on boxed answer matching with `math verify`. For more details, please refer to the Appendix.

Attribution score. For every rollout trajectory τ_i , we compute a TracIn-style (Pruthi et al., 2020) influence score by measuring the alignment between its training loss gradient and a held-out validation gradient:

$$\text{Score}(\tau_i) = \langle \nabla_{\theta} \mathcal{L}_{\text{train}}^{\text{alg}}(\tau_i; \theta), g_V \rangle, \quad (11)$$

where $\mathcal{L}_{\text{train}}^{\text{alg}}(\tau_i; \theta)$ is the per-trajectory loss term of Eq. 8 (PPO) or Eq. 6 (GRPO), and the validation gradient $g_V = \nabla_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}_{\text{val}}} [-\log \pi_{\theta}(y | x)]$ is the supervised gradient on a held out reasoning set $\mathcal{D}_{\text{val}}$ of N_{val} GSM8K prompt response pairs disjoint from training data. Scores are recomputed every training steps.

5.2 Trajectory-Level Attribution Improves PPO

As established in Sec 4.2, our framework predicts that PPO is compatible with trajectory-level attribution. Figure 2 shows the GSM8K Acc@1 over

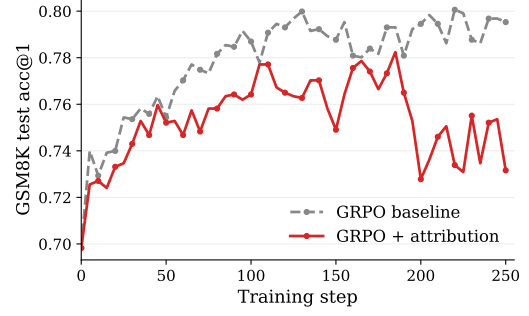


Figure 3: Trajectory-level attribution fails under GRPO. In contrast to PPO, applying the same trajectory-level attribution protocol to GRPO does not yield reliable optimization gains, which is consistent with the prediction that standard GRPO does not admit trajectory-level attribution unless peer context is modeled.

training checkpoints. As can be seen, attribution guided reweighting improves Acc@1 relative to the PPO baseline, which is consistent with trajectory-level attribution being well-defined under the studied PPO objective. This behavior is consistent with Proposition 3 and helps explain the empirical success of trajectory-level attribution in PPO-style training (Hu et al., 2025b; Shu et al., 2026).

5.3 Trajectory-Level Attribution Fails for GRPO

In contrast to PPO, our framework predicts that standard GRPO does not admit trajectory-level attribution unless peer context is modeled (Section 4.1). Figure 3 reports GSM8K Acc@1 for the corresponding GRPO runs. As can be seen, applying the same trajectory-level attribution and reweighting protocol to GRPO fails to improve over the GRPO baseline and instead degrades Acc@1. This contrast with PPO supports Proposition 2. And it shows that when a trajectory’s contribution is defined relative to its peer responses, a trajectory-level attribution score is not a stable intervention signal for GRPO optimization.

5.4 Consistency across Test Datasets

To verify whether our findings are consistent across test datasets, we evaluate the Qwen2.5-1.5B-Instruct checkpoints trained on GSM8K using the MATH500 test set. As shown in Table 2, trajectory-level attribution fails to improve Acc@1 over the GRPO baseline, whereas group-level attribution improves Acc@1 by 5.90 percentage points to 54.80%. These results indicate that our main finding holds across the various test datasets.

Table 2: Cross-dataset evaluation of the Qwen2.5-1.5B-Instruct checkpoints trained on GSM8K. TA and GA denote trajectory-level and group-level attribution, respectively. Δ are computed relative to GRPO.

Method	Unit	Acc@1	Δ
GRPO	–	48.90%	–
GRPO + TA	trajectory	48.80%	−0.10%
GRPO + GA	group	54.80%	+5.90%

5.5 Consistency across Model Backbones

To verify the consistency of our findings across various model backbones, we repeat the GRPO comparison using Gemma 3-1B on GSM8K. As can be seen in Table 3, trajectory-level attribution reduces the best Acc@1 from 64.90% to 61.71%, whereas group-level attribution improves it to 66.26%. These results demonstrate that the observed attribution-unit mismatch is not specific to a specific model backbone.

Table 3: GRPO attribution comparison using Gemma 3-1B on GSM8K. TA and GA denote trajectory-level and group-level attribution, respectively. Δ are computed relative to GRPO.

Method	Final Acc@1	Best Acc@1	Δ
GRPO	64.90%	64.90%	–
GRPO + TA	60.12%	61.71%	−3.19%
GRPO + GA	65.96%	66.26%	+1.36%

6 Discussion

Mechanism behind GRPO degradation. The structural origin of the performance degradation caused by trajectory-level attribution in GRPO is biased intervention estimation. In standard GRPO, reweighting or removing a trajectory changes not only its own loss contribution but also the advantages assigned to its peers through the group statistics. Trajectory-level scores ignore these peer-induced changes, thereby incorrectly estimating the magnitude or direction of the actual intervention effect. This bias does not necessarily mean that trajectory ranking is entirely invalidated. While attribution scores might still retain partial ranking information, they still yield flawed filtering or reweighting decisions. As these flawed decisions are repeatedly applied throughout training, the resulting errors can compound across policy updates and ultimately degrade the overall training

performance.

Extension to Other RLVR Algorithms. Our analysis is not specific to the exact GRPO objective in Section 4.1. Applying SUTVA-based criteria, Table 4 compares representative RLVR algorithms in terms of their credit-assignment structures and attribution units. Methods such as DAPO (Yu et al., 2025), Dr. GRPO (Liu et al., 2025), and GSPO (Zheng et al., 2025) modify the details of sampling, clipping, or normalization, but they still define the model parameter update of one response through statistics computed from other responses in the same prompt conditioned group. As long as this group-relative dependence remains, the contribution assigned to a single trajectory is not determined by that trajectory alone. The SUTVA diagnosis therefore extends beyond standard GRPO and suggests that unit-only attribution for this family should operate at the group level unless peer context is explicitly modeled. Beyond the GRPO family, other RLVR algorithms exhibit distinct dependency structures. For example, REINFORCE++ (Hu et al., 2025a) normalizes advantages over the global optimization batch, causing the contribution assigned to each trajectory to depend on batch-level statistics. Therefore, it does not satisfy the condition required for trajectory-level attribution. In contrast, PPO is compatible with trajectory-level attribution under the trajectory-local assumptions. These analyses demonstrate that our theoretical framework is applicable to RLVR algorithms with varying credit-assignment structures, rather than being confined to standard GRPO.

Reconciling existing attribution for RL successes. This distinction also helps reconcile our negative GRPO result with prior attribution guided RL successes. When the optimizer treats the intervened records as separable update units, trajectory- or transition-level scores can be meaningful, which is the regime targeted by PPO based attribution methods such as Shu et al. (2026) and local online RL attribution methods such as Hu et al. (2025b). In contrast, attribution guided data selection under GRPO family training is better understood as acting on coarser units, such as prompts or prompt conditioned rollout groups, rather than as an intervention on a single trajectory in isolation (Zhu et al., 2026; Yang et al., 2026). These successes therefore do not contradict our SUTVA analysis. They instead suggest that attribution is most reli-

Table 4: SUTVA-based attribution-unit analysis of representative RLVR algorithms. “Conditional” assumes a fixed critic, normalization statistics, and loss-reduction convention, without recomputing cross-sample statistics.

Algorithm	Credit-assignment structure	Trajectory-level SUTVA	Suggested attribution unit
PPO (Schulman et al., 2017)	Trajectory-local rewards and value predictions through GAE	Conditional	Trajectory
GRPO (Shao et al., 2024)	Advantage normalized by the mean and standard deviation of a prompt-conditioned response group	No	Prompt-conditioned rollout group
Dr. GRPO (Liu et al., 2025)	Group-mean reward baseline without standard-deviation normalization	No	Prompt-conditioned rollout group
DAPO (Yu et al., 2025)	GRPO-style group-relative advantage with modified sampling, clipping, and loss aggregation	No	Prompt-conditioned rollout group
GSPO (Zheng et al., 2025)	Group-relative advantage with sequence-level importance ratios and clipping	No	Prompt-conditioned rollout group
REINFORCE++ (Hu et al., 2025a)	Global advantage normalization over the optimization batch	No under global normalization	Normalization batch

Table 5: Multi-seed preliminary comparison of attribution units under GRPO. TA denotes trajectory attribution, and GA denotes group attribution.

Method	Attribution unit	Final Acc@1 (mean $\pm$ std)	Best Acc@1 (mean $\pm$ std)	Δ (Best vs. Base)
GRPO	–	79.76% $\pm$ 0.23%	79.80% $\pm$ 0.19%	–
GRPO + TA	trajectory	76.55% $\pm$ 3.05%	78.77% $\pm$ 0.50%	–1.03%
GRPO + GA	group	80.40% $\pm$ 0.19%	80.86% $\pm$ 0.05%	+1.06%

able when the intervention unit is aligned with the optimizer’s credit assignment unit.

What group-level attribution implies. Moving from trajectory-level to group-level attribution does not simply mean averaging trajectory scores within a group. Rather, it changes the intervention being estimated. Under the GRPO advantage estimator, the relevant counterfactual is how the validation loss would change if a prompt-conditioned rollout group were reweighted, removed, or resampled as a whole. This perspective preserves the group-relative normalization structure used by the optimizer and avoids assigning an intrinsic effect to a trajectory whose contribution is defined only relative to its peers.

As a preliminary validation, we additionally evaluate group-level attribution for GRPO by reweighting prompt-conditioned rollout groups as whole units. Table 5 reports the best GSM8K Acc@1 under the fixed-seed protocol. Compared with trajectory-level attribution, group-level attribution improves the best Acc@1 of GRPO, providing initial empirical support for the unit-alignment principle suggested by our SUTVA analysis. The additional evaluations with MATH500 and Gemma

3-1B (Tables 2 and 3) provide further evidence for this unit-alignment principle across various model backbones and test datasets. We provide the corresponding extension of influence functions to group-level interventions in Appendix E. These preliminary results suggest that future data selection methods for GRPO should prioritize prompts, rollout groups, or group construction policies rather than isolated responses.

7 Conclusion

In this paper, we recast training data attribution as individual treatment effect estimation and identified SUTVA as the structural condition that determines whether attribution at the unit is valid. Our proposed framework illustrates that under the objectives studied here, trajectory-level attribution is well-defined for trajectory-local PPO, whereas trajectory-level attribution is invalid under standard GRPO. And our experiments provide empirical support for both predictions. The framework further suggests group-level attribution as a constructive direction for the GRPO family, and offers a default diagnostic for choosing the right attribution unit in any new training regime.

Limitations

In this section, we honestly discuss several limitations of our work, including empirical scope, applicability to large-scale RL training, scope of trajectory-only attribution under GRPO, and theoretical assumptions.

Empirical scope. Our experiments focus on math oriented RLVR with a fixed model scale, verifier reward, and attribution protocol. Although this setting directly reflects the GRPO-style training regime that motivates our analysis, it does not cover the full range of RLVR tasks, model sizes, reward functions, or attribution estimators. Therefore, our empirical results should be interpreted as evidence for the predicted unit mismatch in the studied setting, rather than as a universal claim about the quantitative behavior of all attribution methods under all group-relative algorithms.

Applicability to large-scale RL training. Our SUTVA diagnostic focuses on the structure of the optimization objective. As long as the algorithm defines the credit assigned to a trajectory through its peer responses, the trajectory-level contribution inherently depends on its peer responses regardless of model scale. However, this structural conclusion does not guarantee that group-level attribution will always yield practical benefits across all large-scale training regimes. In large-scale RL training, distributed optimization, larger and more heterogeneous rollout batches, estimation variance, non-stationary sampling, and the computational cost of group-level estimation can all impact its practical utility. As our current experiments have not yet covered larger-scale models or extended training schedules, evaluating the effectiveness and scalability of group-level attribution in these settings remains an important direction for future research.

Scope of trajectory-only attribution under GRPO. Our framework shows that trajectory-level attribution is not well-defined for GRPO-style objectives unless peer context is modeled, because the contribution of a trajectory depends on the peer responses in its sampled group. Therefore, this paper does not propose a trajectory-level attribution method for GRPO. Instead, it suggests that future attribution methods for GRPO should operate at the group level or explicitly model within group interference. And our framework does not deny that the utility of cross-unit attribution methods or heuristic scores for ranking may still be useful.

Theoretical scope and idealized assumptions.

Our theoretical analysis has two main limitations. First, the attribution quantities studied in this paper are local. The TD-ITE formulation follows the standard infinitesimal reweighting view used in influence based attribution. Therefore, it analyzes first-order sensitivity around the local training buffer rather than the effect of large data interventions or full retraining. Second, the derivations rely on idealized regularity assumptions, including differentiability of the weighted objective, the existence of a well-defined local optimum, and the nonsingular Hessian matrix. These assumptions are standard in influence-style analyses and help isolate the structural role of the attribution unit, but they ignore the finite-step optimizer dynamics, rollout sampling noise, and other non-asymptotic effects in practical RLVR training.

Ethical Considerations

No ethical concerns have been identified in connection with this study.

Acknowledgments

We sincerely thank Mingzhi Wang at the University of Copenhagen for numerous constructive suggestions and insightful discussions. We also thank the anonymous reviewers and AC for their constructive comments and suggestions. We used GPT-5.5 as a writing assistance tool solely to improve the grammar, wording, and readability of our paper. This work was supported in part by the National Natural Science Foundation of China under Grant 62372033, and in part by the Natural Science Foundation of Beijing Municipality under Grant L252025.

References

- Hongjun An, Sida Huang, Siqu Huang, Ruanjun Li, Yuanzhi Liang, Jiawei Shao, Zihan Wang, Cheng Yuan, Chi Zhang, Hongyuan Zhang, Wenhao Zhuang, and Xuelong Li. 2025. [AI flow: Perspectives, scenarios, and approaches](#). *CoRR*, abs/2506.12479.
- Peter M. Aronow and Cyrus Samii. 2017. [Estimating average causal effects under general interference, with application to a social network experiment](#). *The Annals of Applied Statistics*, 11(4):1912 – 1947.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion,

- Tom Conerly, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022. [Training a helpful and harmless assistant with reinforcement learning from human feedback](#). *CoRR*, abs/2204.05862.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#). *CoRR*, abs/2110.14168.
- DeepSeek-AI. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *CoRR*, abs/2501.12948.
- Amirata Ghorbani and James Zou. 2019. Data shapley: Equitable valuation of data for machine learning. In *International conference on machine learning*, pages 2242–2251. PMLR.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.
- Paul W Holland. 1986. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960.
- Jian Hu, Jason Klein Liu, Haotian Xu, and Wei Shen. 2025a. Reinforce++: Stabilizing critic-free policy optimization with global advantage normalization. *arXiv preprint arXiv:2501.03262*.
- Yuzheng Hu, Fan Wu, Haotian Ye, David Forsyth, James Zou, Nan Jiang, Jiaqi W. Ma, and Han Zhao. 2025b. [A snapshot of influence: A local data attribution framework for online reinforcement learning](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Michael G Hudgens and M Elizabeth Halloran. 2008. Toward causal inference with interference. *Journal of the american statistical association*, 103(482):832–842.
- Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. 2022. Data-models: Understanding predictions with data and data with predictions. In *International Conference on Machine Learning*, pages 9525–9587. PMLR.
- Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR.
- Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. 2023. [TRAK: Attributing model behavior at scale](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 27074–27113. PMLR.
- Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. 2020. Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems*, 33:19920–19930.
- Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.
- Donald B Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688–701.
- Donald B Rubin. 1986. Comment: Which ifs have causal answers. *Journal of the American statistical association*, 81(396):961–962.
- John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. 2015. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. [Deepseekmath: Pushing the limits of mathematical reasoning in open language models](#). *CoRR*, abs/2402.03300.
- Dong Shu, Denghui Zhang, and Jessica Hullman. 2026. Learning from the right rollouts: Data attribution for ppo-based llm post-training. *arXiv preprint arXiv:2604.01597*.
- Michael E Sobel. 2006. What do randomized studies of housing mobility demonstrate? causal inference in the face of interference. *Journal of the American Statistical Association*, 101(476):1398–1407.
- Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford,

Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021.

Eric J Tchetgen Tchetgen and Tyler J VanderWeele. 2012. [On causal inference in the presence of interference](#). *Statistical Methods in Medical Research*, 21(1):55–75. PMID: 21068053.

Gemma Team. 2025. [Gemma 3 technical report](#). *CoRR*, abs/2503.19786.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. [Qwen2.5 technical report](#). *CoRR*, abs/2412.15115.

Ningyuan Yang, Weihua Du, Weiwei Sun, Sean Welleck, and Yiming Yang. 2026. Gradalign: Gradient-aligned data selection for llm reinforcement learning. *arXiv preprint arXiv:2602.21492*.

Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gao-hong Liu, Juncui Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, and 17 others. 2025. [DAPO: an open-source LLM reinforcement learning system at scale](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. 2025. [Group sequence policy optimization](#). *CoRR*, abs/2507.18071.

Erle Zhu, Dazhi Jiang, Yuan Wang, Xujun Li, Jiale Cheng, Yuxian Gu, Yilin Niu, Aohan Zeng, Jie Tang, Minlie Huang, and Hongning Wang. 2026. [Data efficient RLVR via off-policy influence guidance](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 46167–46192, San Diego, California, United States. Association for Computational Linguistics.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A Additional Results with TRAK

To verify the consistency of our findings across different attribution estimators, we repeat the GRPO

comparison using TRAK-style scores. As can be seen in Table 6, trajectory-level attribution reduces the best Acc@1 from 77.25% to 76.65%, whereas group-level attribution improves it to 78.17%. These results indicate that the observed attribution-unit mismatch is not unique to any specific attribution estimator.

B Experimental Setup Details

Tables 7 and 8 present the key hyperparameters used for PPO and GRPO training.

C Random seed and reporting protocol

To reduce confounding effects from stochastic training dynamics, all reported training runs use a fixed random seed of 1234 under each experimental configuration. The curves in the main text therefore report single-run Acc@1 values under this fixed-seed protocol. We use the same seed across compared methods to make the PPO/GRPO and attribution/no-attribution comparisons as controlled and reproducible as possible.

D Proofs for the Causal Attribution Framework

Regularity assumptions. All statements in this appendix are local around the unweighted training run $w = 1$ and $\theta^* = \theta^*(1)$. We assume that $L_\epsilon(\theta; w)$ is twice continuously differentiable in θ and continuously differentiable in w in a neighborhood of $(\theta^*, 1)$, that $\theta^*(w)$ is an isolated local minimizer defined by the first-order condition $\nabla_\theta L_\epsilon(\theta^*(w); w) = 0$, and that the Hessian $H_{\theta^*} := \nabla_\theta^2 L_\epsilon(\theta^*; 1)$ is nonsingular. We also assume that the validation loss $L_{\text{val}}(\theta; z_v)$ is differentiable at θ^* .

D.1 Proof of Proposition 1

Let u_i be a candidate attribution unit and define

$$\psi_i(\theta; w) := \frac{\partial L_\epsilon(\theta; w)}{\partial w_{u_i}}. \quad (12)$$

The optimality condition for the weighted training problem is

$$\nabla_\theta L_\epsilon(\theta^*(w); w) = 0. \quad (13)$$

Differentiating this identity with respect to w_{u_i} at $w = 1$ gives

$$H_{\theta^*} \frac{\partial \theta^*(w)}{\partial w_{u_i}} \Big|_{w=1} + \nabla_\theta \psi_i(\theta^*; 1) = 0. \quad (14)$$

Table 6: GRPO attribution results using TRAK-style scores over 100 training steps.. TA denotes trajectory attribution, and GA denotes group attribution.

Method	Attribution Estimator	Best Acc@1	Δ
GRPO	—	77.25%	—
GRPO + TA	TRAK	76.65%	−0.60%
GRPO + GA	TRAK	78.17%	+0.92%

Table 7: Key Hyperparameters for PPO Training.

Hyperparameter	Baseline	+ attribution
backbone	Qwen2.5-1.5B-Instruct	
train / mini batch	256 / 128 prompts	
rollouts per prompt	5	
actor / critic lr	1e−6 / 1e−5	
advantage estimator	GAE ($\gamma=\lambda=1.0$)	
value clip / grad clip	0.5 / 1.0	
KL regularization	in reward, $\beta_{\text{KL}} = 0.01$	
temperature (train / eval)	1.0 / 0 (greedy)	
total epochs / steps	6 / 150	
eval cadence	every 5 steps on full test set	
attribution mode		
score type	—	inner product (TracIn-style)
validation gradient samples	—	$N_{\text{val}} = 200$
shaping β / weight clip	—	0.1 / [0.5, 2.0]

Table 8: Key Hyperparameters for GRPO Training.

Hyperparameter	Baseline	+ attribution
backbone	Qwen2.5-1.5B-Instruct	
train / mini batch	256 / 128 prompts	
rollouts per prompt	5	
actor lr	1e−6 (critic-free)	
advantage estimator	group-relative GRPO ($\gamma=\lambda=1.0$)	
grad clip	1.0	
KL regularization	in loss, $\beta_{\text{KL}} = 0.001$	
temperature (train / eval)	1.0 / 0 (greedy)	
total epochs / steps	10 / 250	
eval cadence	every 5 steps on full test set	
attribution mode		
score type	—	inner product (TracIn-style)
validation gradient samples	—	$N_{\text{val}} = 200$
shaping β / weight clip	—	0.3 / [0.5, 2.0]

Since H_{θ^*} is nonsingular,

$$\left. \frac{\partial \theta^*(w)}{\partial w_{u_i}} \right|_{w=\mathbf{1}} = -H_{\theta^*}^{-1} \nabla_{\theta} \psi_i(\theta^*; \mathbf{1}). \quad (15)$$

Applying the chain rule to the validation loss yields

$$\begin{aligned} & \left. \frac{\partial L_{\text{val}}(\theta^*(w); z_v)}{\partial w_{u_i}} \right|_{w=\mathbf{1}} \\ &= -\nabla_{\theta} L_{\text{val}}(\theta^*; z_v)^{\top} H_{\theta^*}^{-1} \nabla_{\theta} \psi_i(\theta^*; \mathbf{1}). \end{aligned} \quad (16)$$

If SUTVA holds at the unit level, then by Definition 2,

$$\psi_i(\theta; \mathbf{1}) = g_{u_i}(z_{u_i}, \theta), \quad (17)$$

where g_{u_i} depends on the data of unit u_i and on θ , but not on the data or weights of other units. Substituting this expression into Eq. 16 gives

$$\text{ITE}(z_{u_i}; z_v) = -\nabla_{\theta} L_{\text{val}}(\theta^*; z_v)^{\top} H_{\theta^*}^{-1} \nabla_{\theta} g_{u_i}(z_{u_i}, \theta^*). \quad (18)$$

which is the influence-form expression in Proposition 1. The only unit-specific term in this expression is $g_{u_i}(z_{u_i}, \theta^*)$; the Hessian and validation gradient are common to all attribution units for the fixed trained model and validation point. Thus, the per-unit effect is well-defined at the chosen unit level.

Conversely, suppose SUTVA fails at the same unit level. Then there exist two admissible contexts, with the same focal unit z_{u_i} but different peer units or peer weights in the interfering cluster, such that

$$\psi_i(\theta; \mathbf{1}) = \left. \frac{\partial L_{\epsilon}(\theta; w)}{\partial w_{u_i}} \right|_{w=\mathbf{1}} \quad (19)$$

is not determined by z_{u_i} and θ alone. In the nondegenerate case where this difference is not annihilated by the common linear functional $\nabla_{\theta} L_{\text{val}}(\theta^*; z_v)^{\top} H_{\theta^*}^{-1}$, Eq. 16 assigns different treatment effects to the same focal unit under different peer contexts. Therefore no estimator that is only a function of z_{u_i} can identify the unit-level effect across both contexts. The attribution target at that level is not a unit-only estimand; it must either include the interfering context or be lifted to a coarser unit for which the no-interference condition holds.

A degenerate cancellation can make the numerical derivative coincide for a particular validation point even when SUTVA fails. Such cancellation does not restore a well-defined unit-level estimand, because the contribution assigned to the unit by the training objective remains structurally dependent on peer units. This proves the proposition under the stated local regularity and nondegeneracy conditions.

D.2 Proof of Proposition 2

For a GRPO group g , the group-relative advantage is

$$\begin{aligned} A_i(g) &= \frac{r_i - \mu_g}{\sigma_g + \varepsilon}, \\ \mu_g &= \frac{1}{|g|} \sum_{j \in g} r_j, \\ \sigma_g &= \sqrt{\frac{1}{|g|} \sum_{j \in g} (r_j - \mu_g)^2}. \end{aligned} \quad (20)$$

The weighted GRPO objective in Eq. 6 is

$$L_{\epsilon}^{\text{GRPO}}(\theta; w) = - \sum_g \sum_{i \in g} w_i A_i(g) \log \pi_{\theta}(\tau_i). \quad (21)$$

Taking the partial derivative with respect to the weight of trajectory τ_i gives

$$\left. \frac{\partial L_{\epsilon}^{\text{GRPO}}(\theta; w)}{\partial w_i} \right|_{w=\mathbf{1}} = -A_i(g) \log \pi_{\theta}(\tau_i). \quad (22)$$

Although this expression is linear in w_i , its coefficient is not a function of τ_i alone. The term $A_i(g)$ depends on the peer rewards $\{r_j : j \in g, j \neq i\}$ through both μ_g and σ_g . Hence the contribution assigned to τ_i changes when the peer trajectories in the same group change, even if τ_i itself is fixed. Therefore GRPO violates trajectory-level SUTVA. If the entire response group g is treated as the unit, however, the group contribution is determined by the contents of that group, so the obstruction disappears at the group level.

D.3 Proof of Proposition 3

For PPO with per-trajectory advantages, the weighted objective in Eq. 8 is

$$L_{\epsilon}^{\text{PPO}}(\theta; w) = - \sum_i w_i A_{\tau_i}(\theta) \log \pi_{\theta}(\tau_i). \quad (23)$$

Taking the partial derivative with respect to w_i gives

$$\left. \frac{\partial L_{\epsilon}^{\text{PPO}}(\theta; w)}{\partial w_i} \right|_{w=\mathbf{1}} = -A_{\tau_i}(\theta) \log \pi_{\theta}(\tau_i). \quad (24)$$

With per-trajectory GAE, $A_{\tau_i}(\theta)$ is computed from the rewards, states, and value estimates along τ_i . It does not use rewards, states, or weights from another trajectory τ_j with $j \neq i$. Therefore the partial derivative of the weighted objective with respect to w_i is a function of τ_i and θ only. This is exactly the trajectory-level SUTVA condition in Definition 2, so PPO admits trajectory-level attribution under this objective.

E Extension of Influence Functions to Group-Level Interventions

Existing influence-function frameworks (Koh and Liang, 2017) can be extended to group-level interventions by treating each prompt-conditioned rollout group as a single weighted unit. Let g denote a rollout group and let its contribution to the training objective be:

$$\mathcal{L}_g(\theta) = \sum_{i \in g} \mathcal{L}_i(\theta; g), \quad (25)$$

where each trajectory-level term may depend on the group context through its group-relative advantage. We then assign a shared weight w_g to $\mathcal{L}_g$, rather than assigning independent weights to individual trajectories:

$$L_\epsilon^{\text{group}}(\theta; w) = \sum_{g \in \mathcal{G}} w_g \mathcal{L}_g(\theta), \quad (26)$$

where $w_g = 0$ denotes removing the entire group, while setting $w_g = 1 + \epsilon$ infinitesimally upweights it. Unlike independent trajectory weights, the shared weight w_g preserves the peer context within the group.

Under the regularity assumptions, the influence of upweighting group g on the validation loss is:

$$\begin{aligned} \mathcal{I}_g(z_v) &= \left. \frac{\partial L_{\text{val}}(\theta^*(w); z_v)}{\partial w_g} \right|_{w=1} \\ &= -\nabla_{\theta} L_{\text{val}}(\theta^*; z_v)^\top H_{\theta^*}^{-1} \nabla_{\theta} \mathcal{L}_g(\theta^*). \end{aligned} \quad (27)$$

This expression follows directly from Eq. 16 by taking the rollout group as the intervention unit.

This extension is conceptually direct, but efficient implementation at large scale may still involve nontrivial memory, variance, and Hessian-vector-product costs. Developing an efficient group-level influence estimator is beyond the scope of this paper.

F Applicability Beyond Infinitesimal Reweighting

Our formal derivation of TD-ITE is based on infinitesimal reweighting, thus characterizing the local first-order sensitivity around the observed training process. The underlying attribution-unit diagnostic concerns the more general question of whether interventions on a selected unit can be interpreted without omitting the peer context. This question also arises in non-local attribution methods. We discuss the specific properties of several representative paradigms in the following part.

Leave-group-out retraining. Leave-group-out retraining removes an entire prompt-conditioned rollout group and retrains the model without that group. Because the intervention unit contains all peer responses used to define the group-relative advantages, this procedure is aligned with the group-level formulation.

Leave-one-trajectory-out retraining. Removing a single trajectory also defines a valid procedure when the treatment of the remaining group is specified. For example, the procedure must determine whether the group statistics are recomputed and whether the removed trajectory is replaced by a newly sampled response. The resulting effect depends on the remaining peer trajectories and the specified removal rule. It therefore represents a peer-conditioned effect rather than a context-free effect determined by the removed trajectory alone.

Shapley-value methods. Shapley-value methods assign credit by averaging the marginal contribution of a unit over different coalitions (Ghorbani and Zou, 2019). Under the standard GRPO, these coalitions determine which peer responses accompany the focal trajectory and how the corresponding group statistics are constructed. The resulting allocation therefore depends on the coalition construction, the group-completion rule, and the distribution of peer exposure. Once these choices are specified, a Shapley-value method can define a well-posed allocation score, but that score does not automatically recover the unit-only TD-ITE considered in our framework.

Therefore, non-local attribution methods do not eliminate the need to specify the intervention unit and its peer context. They instead define different causal quantities through distinct intervention and aggregation rules.