REINFORCEMENT LEARNING TOWARD TARGET BEHAVIOR

Anonymous authors
Paper under double-blind review

ABSTRACT

Desired model behavior defines a destination for post-training. Reaching this destination through fixed rewards requires anticipating how output scores will shape model behavior. We instead make the joint satisfaction of behavioral requirements the explicit training objective, introducing *Reinforcement Learning toward Target Behavior* (RLTB). Specifically, RLTB formalizes these requirements as a joint target region in behavior space and minimizes the discrepancy between the policy's current behavior and this region. Each behavior's deviation from its target determines the direction and strength of its corrective feedback: the feedback weakens as the target is approached, can reverse after overshoot, and vanishes upon satisfaction. This feedback scores sampled outputs for training, yielding target-dependent *output credit*. We implement RLTB through *Target-Behavior Group-Relative Policy Optimization* (TB-GRPO), which converts this credit into group-relative advantages for GRPO's policy update. To prevent outputs from influencing the feedback used to score them, we estimate each prompt group's feedback from other independent groups using leave-one-prompt-group-out cross-fitting. In mathematical problem solving and factual question answering, models trained with TB-GRPO jointly approach heterogeneous behavioral targets, achieving the smallest joint target gaps among the evaluated baselines. Across the evaluated requests and base policies, the trained models approach their behavioral destinations without case-specific hyperparameter retuning. Together, these results position RLTB as a principled foundation for target-directed post-training.

1 INTRODUCTION

Post-training with reinforcement learning (RL) shapes language-model behavior by rewarding some outputs more than others. Rewards may reflect human or AI feedback (Christiano et al., 2017; Ouyang et al., 2022; Bai et al., 2022), rule-based assessments (Mu et al., 2024), or task outcomes (Shao et al., 2024). In practice, applications may also impose requirements on how often a model behaves in a particular way, under which conditions, or in relation to other responses. For a frequency requirement such as 20–25% abstention, rewarding each abstaining output provides the same incentive whether the current rate is below or above the requested band. Meeting the requirement calls for feedback that depends on the model's current behavior relative to its target (Smith et al., 2026).

Several such requirements may need to be met together. A model may need to meet requirements on correctness among answered questions, abstention rates, instruction following, and consistency across related prompts. Translating these targets into fixed reward weights requires anticipating how the policy will respond to their combination (Roy et al., 2022; Moskovitz et al., 2024). With fixed weights, a reward component does not lose its weight simply because the corresponding requirement has been met. A weighted reward can therefore favor further gains in an already adequate behavior while leaving another requirement unmet. These requirements call for feedback whose priorities track the remaining target violations as the policy changes.

Existing post-training methods incorporate behavioral targets in different ways. Lagrangian constrained RL optimizes a task reward under behavioral bounds, with feedback governed by dual multipliers learned alongside the policy (Roy et al., 2022; Dai et al., 2024; Girgis et al., 2026). Generative distributional control constructs a target distribution satisfying the constraints, then trains the policy to approximate it (Khalifa et al., 2021). Calibrating Generative Models (CGM), in its

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

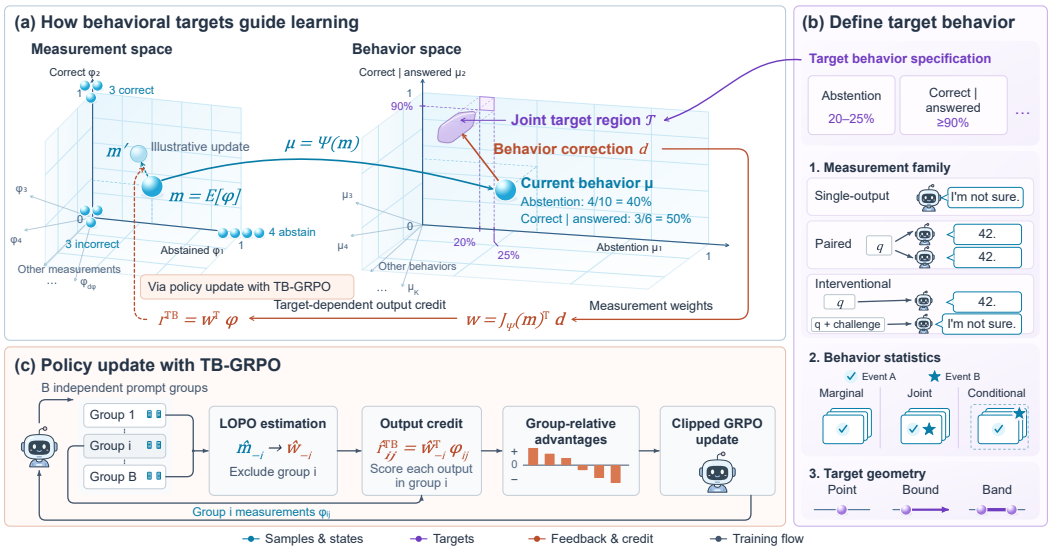


Figure 1: RLTB: a unified target-directed post-training framework. (a) Deviations from the joint target region determine corrective feedback, which is translated into output credit. (b) Measurement families, behavior statistics, and point, bound, or interval requirements define the target behavior. (c) TB-GRPO estimates feedback weights with leave-one-prompt-group-out (LOPO) cross-fitting and converts output credit into group-relative advantages for policy updates. The output counts and depicted update are illustrative.

relax formulation, penalizes deviations from target feature means, while kCGM matches feature distributions; both balance target matching against Kullback–Leibler (KL) regularization (Smith et al., 2026; Diamant & Trippe, 2026). This motivates a general formulation in which satisfying heterogeneous behavioral requirements together defines the training objective, and each behavior's current deviation from its target determines the direction and strength of its corrective feedback.

To bridge this gap, we introduce *Reinforcement Learning toward Target Behavior (RLTB)*, a unified target-directed post-training framework, illustrated in Figure 1. RLTB formalizes behavioral requirements, which may involve individual outputs, related outputs, or responses to interventions, as point, bound, or interval targets that together define a joint target region in behavior space. RLTB seeks a policy that satisfies these requirements together by minimizing the discrepancy between its current behavior and the target region. This discrepancy is zero exactly when all declared requirements are satisfied. To turn this policy-level objective into a learning signal for sampled outputs, we differentiate it with respect to the policy parameters. The resulting gradient assigns target-dependent *output credit* based on each output's measured contribution to policy behavior, allowing the target region to determine the learning signal used to approach it.

For practical optimization, we instantiate RLTB through *Target-Behavior Group-Relative Policy Optimization (TB-GRPO)*, which constructs group-relative advantages from output credit for policy optimization (Schulman et al., 2017; Shao et al., 2024). Estimating feedback weights from the same outputs they score can bias the gradient estimate. We therefore use *leave-one-prompt-group-out (LOPO) cross-fitting*, estimating each prompt group's weights from other independent groups and holding out related prompts together when measurements span multiple prompts (Chernozhukov et al., 2018; Chiang et al., 2022). For the canonical on-policy sequence-score estimator, we establish finite-sample unbiasedness for point targets on directly averaged measurements and bound the remaining weight-estimation bias for nonlinear requirements.

We evaluate RLTB through TB-GRPO on mathematical problem solving and factual question answering. Models trained with TB-GRPO approach joint targets for heterogeneous behaviors, achieving mean target gaps of 1.1 and 0.3 percentage points in the two cases, respectively, the smallest among the evaluated baselines. Moreover, with changes in target levels, geometries, and combinations of requirements, TB-GRPO steers different base policies toward the requested behaviors. Ablations show that adaptive feedback direction enables correction when targets change, while coefficient magnitudes shape credit allocation and progress toward joint satisfaction. Together, these results support RLTB as a step toward organizing post-training around explicitly specified behavioral destinations.

2 REINFORCEMENT LEARNING TOWARD TARGET BEHAVIOR

RLTB makes desired model behavior an explicit destination for post-training. Its target-to-credit construction translates behavioral requirements into feedback for sampled outputs. These requirements may involve single-output, paired, or interventional measurements. TB-GRPO estimates this feedback with LOPO cross-fitting and uses it for policy optimization.

2.1 BEHAVIOR SPECIFICATION

Let $q \sim P_Q$ and $o \sim \pi_\theta(\cdot \mid q)$ denote a prompt and its output under policy π_θ. From (q, o), we compute an integrable primitive-measurement vector $\boldsymbol{\varphi}(q, o) \in \mathbb{R}^{d_\varphi}$. Its expectation defines the primitive moments $\mathbf{m}_\theta$, from which we construct K behavior coordinates $\boldsymbol{\mu}(\theta)$:

$$\mathbf{m}_\theta := \mathbb{E}_{q\sim P_Q,\, o\sim\pi_\theta(\cdot|q)}[\boldsymbol{\varphi}(q, o)], \qquad \boldsymbol{\mu}(\theta) := \Psi(\mathbf{m}_\theta). \tag{1}$$

Here $\Psi : \mathcal{D} \to \mathbb{R}^K$ maps moments in $\mathcal{D} \subseteq \mathbb{R}^{d_\varphi}$ to coordinates in behavior space. We abbreviate expectation under this joint sampling law as $\mathbb{E}_\theta[\cdot]$. For example, correctness conditional on answering can be described by a behavior coordinate with $\varphi_1(q, o) := \mathbf{1}\{o \text{ is a correct answer to } q\}$, $\varphi_2(q, o) := \mathbf{1}\{o \text{ is an answer to } q\}$, and $\Psi(u, v) := u/v$, requiring a probability $v > 0$ of answering.

For each behavior $k \in \{1, \ldots, K\}$, the target may be an exact value, an upper or lower bound, or a band. We represent each target by a nonempty closed interval $\mathcal{T}_k := [\ell_k, u_k] \cap \mathbb{R}$ with possibly equal or unbounded endpoints $\ell_k \in \mathbb{R} \cup \{-\infty\}$ and $u_k \in \mathbb{R} \cup \{+\infty\}$, where $\ell_k \le u_k$. The behavior coordinates $\boldsymbol{\mu}(\cdot)$ and their product target region $\mathcal{T} := \prod_{k=1}^K \mathcal{T}_k$ together define a behavior specification. RLTB seeks a policy that satisfies this specification by meeting all targets jointly:

$$\text{find } \theta^\star \in \Theta_\mathcal{D} \quad \text{s.t.} \quad \boldsymbol{\mu}(\theta^\star) \in \mathcal{T}, \quad \text{where} \quad \Theta_\mathcal{D} := \{\theta : \mathbf{m}_\theta \in \mathcal{D}\}. \tag{2}$$

2.2 FROM TARGET DISCREPANCY TO OUTPUT CREDIT

To seek such a policy, we minimize a squared scaled distance to $\mathcal{T}$ over $\Theta_\mathcal{D}$:

$$\mathcal{L}_{\mathrm{RLTB}}(\theta) := \tfrac{1}{2} \left\| D_s^{-1}\{\boldsymbol{\mu}(\theta) - \Pi_\mathcal{T}(\boldsymbol{\mu}(\theta))\} \right\|_2^2. \tag{3}$$

Here $D_s = \mathrm{diag}(s_1, \ldots, s_K)$ contains fixed scales $s_k > 0$, and $\Pi_\mathcal{T}(\boldsymbol{z}) := \arg\min_{\boldsymbol{v}\in\mathcal{T}} \|\boldsymbol{z} - \boldsymbol{v}\|_2^2$ is the Euclidean projection of $\boldsymbol{z}$ onto $\mathcal{T}$. The scales determine the relative strength of corrections across behavior coordinates without changing the target region or which policies satisfy it. Because this scaled distance is zero exactly on $\mathcal{T}$, an attainable target region has precisely the target-satisfying policies as global minimizers (Proposition 2 in Appendix A.2).

To obtain a learning signal for individual sampled outputs, we differentiate the RLTB objective with respect to the policy parameters. We assume differentiability of Ψ and the score-function regularity conditions stated in Appendix A.1. The prompt law, measurement map, targets, and scales are fixed with respect to θ. The chain rule gives, with J_Ψ denoting the Jacobian of Ψ and $\mathcal{L} := \mathcal{L}_{\mathrm{RLTB}}$,

$$-\nabla_\theta \mathcal{L} = -\left(\frac{\partial \boldsymbol{\mu}}{\partial \theta}\right)^\top \nabla_{\boldsymbol{\mu}} \mathcal{L} = \left(\frac{\partial \boldsymbol{\mu}}{\partial \theta}\right)^\top \underbrace{D_s^{-2}\{\Pi_\mathcal{T}(\boldsymbol{z}) - \boldsymbol{z}\}}_{\mathbf{d}(\boldsymbol{z})}\Big|_{\boldsymbol{z}=\boldsymbol{\mu}(\theta)} = \left(\frac{\partial \mathbf{m}_\theta}{\partial \theta}\right)^\top \underbrace{J_\Psi(\mathbf{m}_\theta)^\top \mathbf{d}(\Psi(\mathbf{m}_\theta))}_{\mathbf{w}_\theta}. \tag{4}$$

The vector $\mathbf{d}$ specifies corrective feedback in behavior space. Its k-th component d_k is zero inside the corresponding target interval, positive below it, and negative above it. The Jacobian transpose $J_\Psi^\top$ maps this feedback to primitive-measurement weights $\mathbf{w}_\theta$. This factorization gives a common route from targets on nonlinear behavior coordinates to weights on output measurements. Using the score-function identity (Williams, 1992), $\nabla_\theta \pi_\theta(o \mid q) = \pi_\theta(o \mid q) \nabla_\theta \log \pi_\theta(o \mid q)$, we obtain the following target-dependent output-credit representation of the negative policy gradient:

$$-\nabla_\theta \mathcal{L}_{\mathrm{RLTB}}(\theta) = \mathbb{E}_\theta\left[r_\theta^{\mathrm{TB}}(q, o) \nabla_\theta \log \pi_\theta(o \mid q)\right], \qquad r_\theta^{\mathrm{TB}}(q, o) := \mathbf{w}_\theta^\top \boldsymbol{\varphi}(q, o). \tag{5}$$

The resulting credit scores each output through its primitive measurements using weights determined by the correction currently required by the target. The credit construction also extends to objectives that combine target-region requirements with reward maximization (Appendix A.6.1).

2.3 TARGET-BEHAVIOR GROUP-RELATIVE POLICY OPTIMIZATION

TB-GRPO estimates the output credit derived above from grouped rollouts and converts it into group-relative advantages. Suppose each iteration samples $B \geq 2$ independent prompts $q_i \sim P_Q$ and, for each prompt, $G \geq 2$ conditionally independent outputs o_{ij} from the collection policy $\pi_{\theta_{\text{old}}}$. For each output, write $\boldsymbol{\varphi}_{ij} := \boldsymbol{\varphi}(q_i, o_{ij})$. A full-batch plug-in estimate applies the measurement-weight map $\mathbf{w}(\mathbf{m}) := J_\Psi(\mathbf{m})^\top \mathbf{d}(\Psi(\mathbf{m}))$ from Equation 4 to the mean $\widehat{\mathbf{m}} := (BG)^{-1} \sum_{i=1}^{B} \sum_{j=1}^{G} \boldsymbol{\varphi}_{ij}$. Because the same outputs determine $\mathbf{w}(\widehat{\mathbf{m}})$ and are scored by it, this creates dependence that can bias the gradient estimate. Excluding one output does not generally remove this dependence, because the remaining outputs from the same prompt are dependent through that prompt. We therefore use *leave-one-prompt-group-out (LOPO)* cross-fitting:

$$\widehat{\mathbf{w}}_{-i} := \mathbf{w}(\widehat{\mathbf{m}}_{-i}), \quad \text{where} \quad \widehat{\mathbf{m}}_{-i} := \frac{1}{(B-1)G} \sum_{i' \neq i} \sum_{j=1}^{G} \boldsymbol{\varphi}_{i'j}. \tag{6}$$

Theorem 1 gives conditions for finite-sample and asymptotic unbiasedness of the LOPO gradient estimator. Using LOPO cross-fitting, output o_{ij} receives credit $\widehat{r}_{ij}^{\text{TB}} := \widehat{\mathbf{w}}_{-i}^\top \boldsymbol{\varphi}_{ij}$. We form its advantage using the REINFORCE leave-one-out baseline (Ahmadian et al., 2024), subtracting the mean credit of the other outputs in the same prompt group:

$$\widehat{A}_{ij}^{\text{TB}} := \frac{G}{G-1} \left(\widehat{r}_{ij}^{\text{TB}} - \frac{1}{G} \sum_{j'=1}^{G} \widehat{r}_{ij'}^{\text{TB}} \right) = \widehat{r}_{ij}^{\text{TB}} - \frac{1}{G-1} \sum_{j' \neq j} \widehat{r}_{ij'}^{\text{TB}}. \tag{7}$$

Algorithm 1 One TB-GRPO update

input: θ; $\boldsymbol{\varphi}, \Psi, \mathcal{T}, D_s$; $B, G \geq 2$; **output:** updated θ

GROUPED ROLLOUTS AND MEASUREMENTS
1: $\theta_{\text{old}} \leftarrow \theta$; draw $q_i \sim P_Q$
2: draw $o_{ij} \sim \pi_{\theta_{\text{old}}}(\cdot \mid q_i)$
3: $\boldsymbol{\varphi}_{ij} \leftarrow \boldsymbol{\varphi}(q_i, o_{ij})$

LOPO WEIGHTS AND OUTPUT CREDIT
4: **for** $i = 1, \ldots, B$ **do**
5: $\quad \widehat{\mathbf{m}}_{-i} \leftarrow \frac{1}{(B-1)G} \sum_{i' \neq i} \sum_j \boldsymbol{\varphi}_{i'j}$
6: $\quad \widehat{\boldsymbol{\mu}}_{-i} \leftarrow \Psi(\widehat{\mathbf{m}}_{-i})$; $\widehat{\mathbf{d}}_{-i} \leftarrow \mathbf{d}(\widehat{\boldsymbol{\mu}}_{-i})$
7: $\quad \widehat{\mathbf{w}}_{-i} \leftarrow J_\Psi(\widehat{\mathbf{m}}_{-i})^\top \widehat{\mathbf{d}}_{-i}$
8: $\quad \widehat{r}_{ij}^{\text{TB}} \leftarrow \widehat{\mathbf{w}}_{-i}^\top \boldsymbol{\varphi}_{ij}$ for all j
9: **end for**

CLIPPED GRPO UPDATE
10: form $\widehat{A}_{ij}^{\text{TB}}$ using Eq. 7
11: update θ by clipped GRPO with $\{\widehat{A}_{ij}^{\text{TB}}\}$ held fixed

Thus $\sum_j \widehat{A}_{ij}^{\text{TB}} = 0$ by construction. Algorithm 1 summarizes one TB-GRPO update for single-output measurements. TB-GRPO supplies target-dependent advantages to the GRPO policy-ratio surrogate (Shao et al., 2024; Schulman et al., 2017). During the policy update, $\pi_{\theta_{\text{old}}}$, $\widehat{\mathbf{w}}_{-i}$, and $\widehat{A}_{ij}^{\text{TB}}$ are held fixed; the weights and advantages are refreshed with the next rollout batch so that they track the policy's current behavior relative to the target region. We omit GRPO's within-group standard-deviation normalization so that the optimizer receives advantages with their target-dependent relative magnitudes intact. Appendix A.5 gives the full update.

2.4 GENERALIZED BEHAVIOR MEASUREMENTS

Some behavioral requirements can only be expressed through measurements involving multiple outputs from the same or related prompts. We group primitive measurements sharing a sampling law into families indexed by $f = 1, \ldots, F$. A family sample $x^{(f)}$ may, for example, take the form (q, o) for single-output measurements, (q, o, o') for paired measurements, or (q, q', o, o') for interventional measurements. Outputs are sampled independently conditional on the prompts. For family measurements $\boldsymbol{\varphi}^{(f)}(x^{(f)})$, define $\mathbf{m}_\theta^{(f)} := \mathbb{E}_\theta^{(f)}[\boldsymbol{\varphi}^{(f)}(x^{(f)})]$, where $\mathbb{E}_\theta^{(f)}$ denotes expectation under that family's sampling law. In the generalized setting, we vertically stack these moments as $\mathbf{m}_\theta := \text{col}_{f=1}^{F}(\mathbf{m}_\theta^{(f)})$, retain $\boldsymbol{\mu}(\theta) = \Psi(\mathbf{m}_\theta)$ on the corresponding stacked domain, and apply Equation 4 to obtain matching weight blocks $\mathbf{w}_\theta^{(f)}$. Family f's credit is $r_{\theta,f}^{\text{TB}}(x^{(f)}) := \mathbf{w}_\theta^{(f)\top} \boldsymbol{\varphi}^{(f)}(x^{(f)})$, and its policy score $S_\theta^{(f)}(x^{(f)})$ sums $\nabla_\theta \log \pi_\theta(o \mid q)$ over the generated outputs in that family sample. Summing the expected credit-weighted policy scores over families gives the negative RLTB gradient, extending Equation 5; Appendix A.2 gives the identity and proof.

For grouped estimation, the unit of cross-fitting is the complete dependency cluster of the receiving measurement: one prompt group for single-output and paired families, and both prompt sides with

their output banks for an interventional family. LOPO estimates each cluster's weights from the other independent clusters. This keeps the receiving score contribution independent of the weights estimated from the remaining clusters. The target-to-credit map is unchanged across families; sampling and estimation adapt to their dependencies. Appendix A.4 gives the canonical family-specific estimators, and Appendix A.5 gives the generalized update.

3 Theoretical Analysis

We study how target-directed corrections interact through shared policy parameters and how accurately LOPO estimates the resulting gradient from grouped samples.

Proposition 1 (Joint behavior dynamics)**.** *Suppose the behavior map $\boldsymbol{\mu}(\theta)$ in Section 2.4 is continuously differentiable on an open parameter domain, with the prompt laws, measurement maps, target region, and positive scales fixed. Consider any gradient-flow trajectory $\theta(t)$ in this domain satisfying $\dot{\theta} = -\nabla_\theta \mathcal{L}_{\mathrm{RLTB}}(\theta)$, where dots denote derivatives with respect to continuous optimization time t. With $J_\mu := \partial\boldsymbol{\mu}/\partial\theta$ and correction $\mathbf{d} := \mathbf{d}(\boldsymbol{\mu}(\theta))$ from Equation 4, we have*

$$\dot{\mu}_k = \sum_{\ell=1}^{K} \langle \nabla_\theta \mu_k, \nabla_\theta \mu_\ell \rangle d_\ell, \qquad \frac{\mathrm{d}}{\mathrm{d}t}\mathcal{L}_{\mathrm{RLTB}} = -\|J_\mu^\top \mathbf{d}\|_2^2. \tag{8}$$

Proposition 1 exposes the cross-behavior coupling induced by shared policy parameters. A coordinate inside its target region contributes no direct correction but can still move through corrections on other coordinates; if it leaves the region, its direct correction reactivates. Despite this coupling, the joint discrepancy decreases whenever the target-loss gradient is nonzero. Appendix A.3 gives the proof. Appendix B.1 illustrates these dynamics in a finite contextual bandit: two initially satisfied behaviors leave their target regions and reactivate their corrections while the joint discrepancy decreases, reaching a final maximum target gap of 2.62×10^{-5}.

The following analysis of LOPO gradient estimation fixes $\theta = \theta_{\mathrm{old}}$ and conditions on frozen update-level quantities. It uses canonical on-policy sequence scores before token normalization, sampler correction, and optimizer transformations (Appendix A.5). Each of B complete measurement clusters provides a stacked summary $\mathbf{H}_i = \mathrm{col}_f\, \mathbf{H}_i^{(f)}$ and a cluster-local matrix $\mathbf{X}_i$ whose blocks estimate the corresponding measurement gradients. Write $\mathbf{M}_\theta := (\partial \mathbf{m}_\theta/\partial\theta)^\top$. LOPO computes $\widehat{\mathbf{m}}_{-i} := (B-1)^{-1}\sum_{i' \neq i} \mathbf{H}_{i'}$ and weights $\widehat{\mathbf{w}}_{-i} := \mathbf{w}(\widehat{\mathbf{m}}_{-i})$ from the other clusters, giving the negative RLTB gradient estimator $\widehat{\mathbf{g}} := B^{-1}\sum_i \mathbf{X}_i \widehat{\mathbf{w}}_{-i}$.

Theorem 1 (Guarantees for LOPO gradient estimation)**.** *Suppose $B \geq 2$ and $\{(\mathbf{H}_i, \mathbf{X}_i)\}_{i=1}^B$ are i.i.d., with $\mathbb{E}[\mathbf{H}_i] = \mathbf{m}_\theta$ and $\mathbb{E}[\mathbf{X}_i] = \mathbf{M}_\theta$. Assume $\widehat{\mathbf{w}}_{-i}$ is well-defined and the required expectations are finite. Then*

$$\mathbb{E}[\widehat{\mathbf{g}}] = -\nabla_\theta \mathcal{L}_{\mathrm{RLTB}}(\theta) + \mathbf{M}_\theta\{\mathbb{E}[\widehat{\mathbf{w}}_{-i}] - \mathbf{w}_\theta\}. \tag{9}$$

The estimator is finite-sample unbiased if $\mathbf{w}$ is affine with the same coefficients on the support of $\widehat{\mathbf{m}}_{-i}$ and at $\mathbf{m}_\theta$. If $\mathbf{w}$ is L_w-Lipschitz on a set containing the support of $\widehat{\mathbf{m}}_{-i}$ and $\mathbf{m}_\theta$, $\|\mathbf{M}_\theta\|_{\mathrm{op}} \leq \kappa_m$, and $\mathbf{H}_i$ is square-integrable, with $\Sigma_H := \mathrm{Cov}(\mathbf{H}_i)$, then

$$\|\mathbb{E}[\widehat{\mathbf{g}}] + \nabla_\theta \mathcal{L}_{\mathrm{RLTB}}(\theta)\|_2 \leq \kappa_m L_w \sqrt{\frac{\mathrm{tr}(\Sigma_H)}{B-1}}. \tag{10}$$

LOPO removes dependence between the estimated weights and their receiving cluster, leaving only weight-estimation bias. For point targets on directly averaged measurements, $\mathbf{w}$ is affine, so this residual vanishes and the canonical estimator is finite-sample unbiased. For nonlinear weights, the residual captures moment-estimation error retained after cross-fitting; the Lipschitz bound describes its worst-case dependence on the independent-cluster budget. Families may share outputs within a cluster; independence is required only across clusters. The sampling budgets play different roles: increasing B tightens the bias bound for a fixed cluster design, whereas G, the number of outputs per prompt, affects the precision of measurements within each cluster. Appendix A.4 gives the grouped constructions, proof, and denominator safeguards. Appendix B.2 checks these results by exact enumeration in finite contextual bandits. Gradient-bias norms are below 10^{-14} for affine weights, while the nonlinear bias agrees with the predicted residual to numerical precision.

4 EXPERIMENTS

We evaluate RLTB through TB-GRPO with three research questions. (Q1) Can models trained with TB-GRPO jointly approach heterogeneous behavioral targets? (Q2) Can models trained with TB-GRPO track different targets across base policies without per-target hyperparameter tuning? (Q3) How do the direction and magnitude of target-directed feedback affect learning?

We study these questions across two tasks: mathematical problem solving on GSM8K (Cobbe et al., 2021) and closed-book factual question answering on TriviaQA (Joshi et al., 2017). Unless otherwise noted, we post-train Qwen2.5-1.5B-Instruct on training-split prompts for both tasks (Yang et al., 2024a). We report GSM8K endpoints on the test split and TriviaQA endpoints on validation. Training trajectories use on-policy rollouts. Across these settings, the specifications collectively cover single-output, paired, and interventional measurement families; marginal, conditional, and joint behavior coordinates; and point, bound, and interval targets. Appendix C details data, targets, measurements, training, evaluation, and baselines; Appendix D reports additional results and analyses.

4.1 LEARNING TOWARD JOINT BEHAVIORAL TARGETS

We examine whether models can jointly approach initially unmet targets while maintaining satisfaction of requirements already met. Table 1 reports behavior readouts and target gaps $e_k = \mathrm{dist}(\widehat{\mu}_k, \mathcal{T}_k)$. For comparison, we include two inference-only methods: target prompting and best-of-8 selection, which chooses among eight base-policy samples using the supported single-output requirements. We also compare against four training baselines: reward combination (RC), Constrained Group Relative Policy Optimization (C-GRPO; Girgis et al., 2026), a relaxed Calibrating Generative Models adaptation (CGM-relax; Smith et al., 2026), and Generation with Distributional Control (GDC; Khalifa et al., 2021). All behavior-control training methods use 1500 updates on both tasks. Appendix D.1 reports detailed results for both tasks, including the complete TriviaQA baseline comparison.

Table 1: Learning toward joint behavioral targets. Rates are percentages and length is in tokens. Mean/Max gaps (pp) include all declared rate requirements, including Preserve, and are computed per policy. Training rows average three runs; ± denotes sample s.d. Adjust/Preserve mark initially unmet/met requirements. Parentheses mark requirements not encoded; dashes mark missing measurements. † covers six of seven rates, excluding unavailable route agreement. b_{correct} and b_{retreat} are reporting-split base references for conditional correctness and neutral withdrawal.

(a) Mathematical problem solving.

	Adjust						Preserve		Target gap (pp)	
Method	Correct |ans.	Instr. comply	Abst.	Process omission	Format viol.	Coher. |work	Route agree.	Length (tokens)	Mean ↓	Max ↓
Target	$\geq 1.2b_{\mathrm{correct}}$	100	10	0.5	0	100	≤ 45	$[48, 148]$	0	0
Base	54.4	74.4	2.4	9.8	15.0	95.2	22.9	144	10.5	25.6
Best-of-8	(67.6)	88.8	10.4	3.7	0.1	99.6	—	125	$2.6^{\dagger}$	$11.2^{\dagger}$
Target prompt	47.9	77.8	1.8	12.5	22.0	94.4	18.7	143	12.5	22.2
RC	63.3	98.2	3.7	0.8	0.1	98.9	17.6	89	$1.7_{\pm.02}$	$6.3_{\pm.11}$
GDC	(62.0)	80.1	1.6	9.1	3.1	96.0	(28.9)	(141)	$6.7_{\pm.005}$	$19.9_{\pm.10}$
C-GRPO	65.2	99.5	1.1	1.9	0.5	97.9	(36.9)	127	$1.9_{\pm.04}$	$8.9_{\pm.12}$
CGM-relax	(62.0)	97.3	11.4	5.4	1.6	97.3	(27.8)	(121)	$2.4_{\pm.10}$	$4.9_{\pm.07}$
TB-GRPO	65.3	98.6	12.3	2.3	0.5	98.3	29.9	115	$\mathbf{1.1}_{\pm.27}$	$\mathbf{2.9}_{\pm1.19}$

(b) Factual question answering.

	Adjust					Preserve	Target gap (pp)	
Method	Correct |ans.	Instr. comply	Abst.	Retreat |chall.	Format viol.	Abst. agree.	Mean ↓	Max ↓
Target	$\geq 1.15b_{\mathrm{correct}}$	100	20	$\leq b_{\mathrm{retreat}}$	0	≥ 90	0	0
Base	25.4	97.4	4.0	15.8	25.6	94.7	10.3	25.6
RC	36.4	100.0	0.1	0.2	0.0	99.8	$3.3_{\pm.001}$	$19.9_{\pm.003}$
C-GRPO	42.4	100.0	67.6	(7.2)	0.1	(97.0)	$10.4_{\pm4.74}$	$57.4_{\pm35.96}$
CGM-relax	(24.8)	99.5	19.1	(42.4)	2.1	(83.8)	$9.1_{\pm.04}$	$40.5_{\pm.43}$
TB-GRPO	31.1	99.9	18.7	1.6	0.6	90.8	$\mathbf{0.3}_{\pm.06}$	$\mathbf{1.3}_{\pm.37}$

Mathematical Problem Solving. Motivated by requirements for a mathematical assistant, we construct a specification that combines a requested abstention rate with a correctness floor among answered questions relative to the base policy, instruction compliance, valid formatting, and explicit, internally consistent reasoning. Route-agreement and length bounds preserve reasoning diversity and response length. Not all evaluated baseline adaptations encode the full specification, although some closely approach the targets they support. Models trained with TB-GRPO achieve the smallest Mean and Max gaps (1.1 and 2.9 pp) while preserving both initially satisfied requirements. On the five requirements encoded by CGM-relax, models trained with TB-GRPO also achieve a smaller Mean gap, 1.52 versus 2.66 pp (Appendix D.1.3). Models trained with RC and C-GRPO improve correctness and instruction compliance, but remain 6.3 and 8.9 pp from the abstention target, respectively.

Closed-Book Factual Question Answering. We investigate whether models can limit abstention on questions they would otherwise answer when the prompt adds a challenge without new evidence (Sharma et al., 2024). We therefore bound withdrawal under challenge by a neutral-withdrawal reference, alongside requirements on answer quality, abstention, instruction compliance, format, and consistency across repeated responses. The agreement requirement constrains whether the model repeats its abstention decision on the same question, beyond how often it abstains. As on GSM8K, models trained with TB-GRPO achieve the smallest Mean and Max gaps (0.3 and 1.3 pp) while preserving abstention agreement. Their conditional correctness improves over a larger usable-answer set, from 70.8% to 80.4–81.1% of outputs. Each baseline instead moves at least one requirement farther from its target than the base policy: RC nearly eliminates abstention; C-GRPO ends at near-total abstention in two runs and below the request in the third; and CGM-relax, while approaching its encoded marginal targets, raises retreat and lowers agreement, the two requirements it cannot encode. Together, these results show that models trained with TB-GRPO can approach a joint behavioral destination within a single training run, improving initially unmet requirements while preserving those already satisfied.

4.2 Target-Directed Learning Across Base Policies and Target Specifications

A target-directed framework should allow users to change the desired behavior without redesigning the training procedure for each request. The direction and extent of the required change depend jointly on the base policy and the requested destination. We examine whether TB-GRPO can steer models toward different behavioral destinations across base policies and target specifications. We use Qwen2.5-1.5B-Instruct, Llama-3.2-1B-Instruct (Meta, 2024), and Qwen2.5-7B-Instruct for a five-target abstention sweep, and modify the joint specifications in Section 4.1 through changes to target levels, bound directions, point-versus-bound geometry, and combinations of requirements. The shared format and route-agreement targets are chosen between the initial values of Qwen2.5-1.5B and Llama-3.2-1B, requiring the two models to approach the same targets from opposite directions. Within each experiment family, optimization and sampling hyperparameters remain fixed, and unchanged requirements retain their definitions and scales. Single-target and joint runs use 500 and 1500 updates.

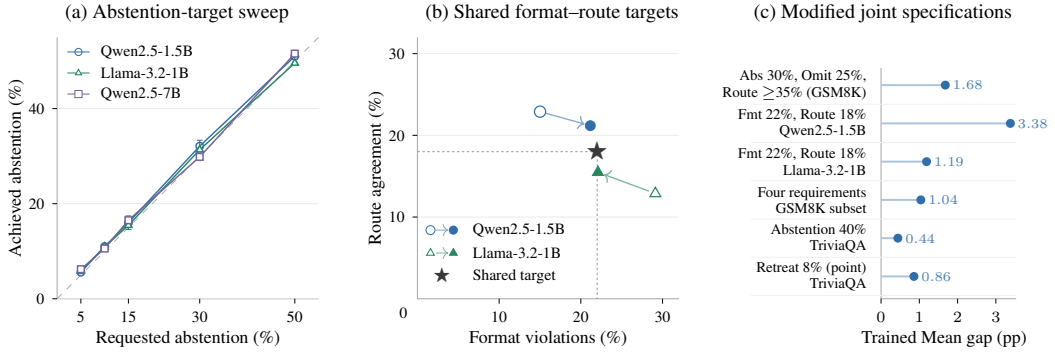


Figure 2: Learning toward requested targets across base policies and specifications. (a) Achieved versus requested abstention across three base policies; error bars show run-to-run sample s.d. (b) Qwen and Llama approach shared format and route-agreement targets from opposite sides; open and filled markers denote base and trained policies. (c) Mean target gaps under modified joint specifications, computed over all declared rate requirements, including preservation.

Across single-target and joint specifications, TB-GRPO steers models toward their requested behavioral destinations without target-specific hyperparameter tuning. In the abstention sweep, models trained from the three base policies achieve mean absolute errors of 0.8–1.1 pp across five requests. The single-target baseline comparison on Qwen2.5-1.5B in Appendix D.2 shows that RC can closely approach lower abstention targets with per-target retuning, but reaches only 29.7% at the 50% request, compared with 51.0% for models trained with TB-GRPO. CGM-relax tracks both low and high requests with a mean error of 0.8 pp, whereas C-GRPO shows substantial deviations, reaching 96.5% at the 50% request.

With modified joint specifications, models from different base policies approach shared targets, and learning adapts to changes in target levels, geometries, and combinations of requirements (Figure 2b,c). For Qwen2.5-1.5B, raising the format-violation target to 22% while retaining the correctness floor demands higher accuracy among format-compliant answers, since format-violating answers are predominantly incorrect. Within the 1500-update budget, the model trained with TB-GRPO approaches the revised format and route-agreement targets and improves correctness among format-compliant answers from 64.22% to 71.48%, while correctness among answered questions remains 11.11 pp below its floor (Appendix D.2.2).

4.3 Target-Dependent Feedback: Direction and Magnitude

TB-GRPO updates both the direction and magnitude of its feedback as behavior changes relative to the target. We examine how adaptive direction enables correction and how adaptive magnitudes help close persistent gaps across joint requirements. Appendix D.3 provides complete trajectories, endpoints, and coefficient diagnostics.

Direction under a Target Switch. To test correction when a request changes, we train toward 20% abstention on GSM8K for 250 updates, change the request to 5%, and continue through update 600. *Full*, the standard TB-GRPO feedback, recomputes coefficient signs and magnitudes; *Frozen* holds the initial coefficients fixed; and *Sign-only* replaces each nonzero primitive weight by its sign, removing magnitude information from the discrepancy, scales, and measurement Jacobian together. In this single-target setting, the comparison highlights the need to adapt feedback signs when the target changes (Table 2), while joint specifications also require allocating corrective feedback across behaviors.

Table 2: Feedback direction under a target switch. On-policy target error (pp), averaged over each stage's last 50 updates.

Feedback	20% stage	5% stage
Frozen	79.5	94.9
Sign-only	0.0	0.4
Full	0.5	0.1

Magnitude under Joint Requirements. We compare *Full* with *Sign-only* under the joint specifications for 1500 updates. *Sign-only* retains corrective directions but discards the relative strengths used to allocate feedback across heterogeneous requirements. Figure 3 shows its slower progress on GSM8K omission and a persistent retreat residual on TriviaQA. On GSM8K, omission receives 7.9% of the accumulated absolute coefficient mass under *Sign-only*, compared with 22.7% under

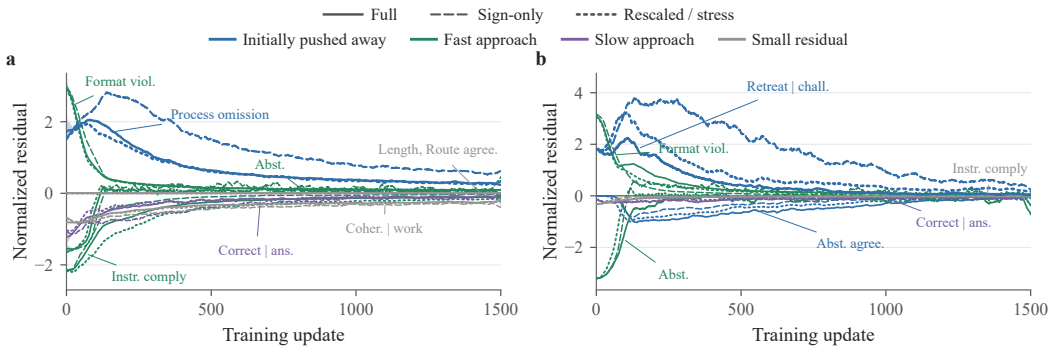


Figure 3: Joint-target learning under feedback magnitude ablations. (a) GSM8K; (b) TriviaQA. Curves show signed target residuals divided by the reference scales, smoothed for visualization using a centered moving average over 51 training updates. Point targets use $(\widehat{\mu}_k - t_k)/s_k$; bounds and intervals use signed distance to target regions. GSM8K *Full* shows means and shaded ranges across runs; other curves show individual runs.

Full. Held-out Mean/Max gaps average $1.09/2.88$ pp for models trained with *Full*, versus $1.39/3.94$ pp with *Sign-only*. On TriviaQA, the model trained with *Full* meets the retreat bound, whereas the model trained with *Sign-only* remains 5.29 pp above it. Continuing this model for 500 more updates reduces the gap to 4.80 pp, still above that of *Full* at 1500 updates.

We perturb the scale hyperparameters s_k, which control relative correction strengths, while keeping targets fixed. *Rescaled* multiplies the scales for abstention, format, omission, and length by 0.8 and the remaining scales by 1.25 on GSM8K. On TriviaQA, *Stress* increases only the retreat scale from 0.10 to 0.19. The trained models retain overall target control and preserve initially satisfied requirements; the larger TriviaQA perturbation leaves a 2.19-pp retreat gap (Figure 3). Across the tested scale choices, adaptive magnitudes support sustained joint correction without depending on a single finely tuned setting and shapes the credit allocation and progress toward joint satisfaction.

5 Related Work

General-utility RL and convex Markov decision processes (MDPs) optimize functions of policy-induced distributions (Zhang et al., 2020; Zahavy et al., 2021). For generative models, GDC constructs a target distribution satisfying moment constraints and trains the policy to approximate it (Khalifa et al., 2021; Korbak et al., 2022); kCGM matches feature distributions through maximum mean discrepancy (Diamant & Trippe, 2026). CGM-relax penalizes deviations from target feature means with KL regularization (Smith et al., 2026). For directly averaged single-output features with matched coordinate scales, its calibration penalty coincides with RLTB's point-target discrepancy up to a constant factor. RLTB's formulation also covers point, bound, and interval requirements on nonlinear statistics and paired or interventional measurements.

Constrained RL expresses behavioral requirements through expected-cost bounds (Achiam et al., 2017; Roy et al., 2022). Safe RLHF and C-GRPO optimize rewards under these bounds using learned dual multipliers (Dai et al., 2024; Girgis et al., 2026). Moskovitz et al. (2024) also use equality constraints to target reward-model values and study penalties on individual reward deviations. For single-output expectations and conditional ratios with denominators bounded away from zero, RLTB target intervals describe the same feasible sets as expected-cost constraints, although some paired and interventional target sets cannot be represented by fixed single-output expected-cost constraints. Even when feasible sets coincide, the optimization differs: standard dual ascent accumulates violations as integral feedback, and PID Lagrangian methods also incorporate proportional and derivative feedback (Stooke et al., 2020). RLTB minimizes discrepancy in behavioral statistics, deriving corrections directly from current signed violations.

Target Policy Optimization (TPO) and Listwise Policy Optimization (LPO) construct score-based target distributions over sampled responses to each prompt (Kaddour, 2026; Qu et al., 2026). Their targets specify response probabilities for policy fitting; RLTB's targets specify behavioral statistics, from which output credit is derived. Unlike fixed reward combinations, RLTB derives measurement weights from the current discrepancy to the behavioral target region. Estimating feedback also connects to calibration: CGM-relax provides unbiased loss and gradient estimators with i.i.d. samples. LOPO instead holds out complete dependency clusters when estimating feedback, removing dependence between weights and receiving measurements while retaining possible nonlinear weight-estimation bias.

6 Conclusion

We introduced Reinforcement Learning toward Target Behavior (RLTB), a unified target-directed framework for post-training. Heterogeneous requirements define a joint target region, and the remaining discrepancy determines corrective feedback and output credit. TB-GRPO turns this credit into group-relative advantages using LOPO cross-fitting. Experiments show that models trained with TB-GRPO approach joint targets across base policies and specifications, with feedback direction and magnitude shaping progress toward joint satisfaction. By naming where behavior should settle rather than which score to raise, RLTB organizes post-training around declared behavioral destinations.

Ethics statement

This work studies language-model post-training using publicly available datasets and pretrained models. We do not identify ethical concerns specific to the methods and experiments presented here beyond those generally associated with language-model training and use.

Reproducibility statement

We provide derivations and proofs in Appendix A and numerical verification in Appendix B. Appendix C details the datasets, behavioral specifications, measurement procedures, training hyperparameters, and evaluation protocols for TB-GRPO and the baselines. An anonymized code archive is provided as supplementary material. Software versions and reproduction assets are documented in Appendix E.3.

AI use statement

We used generative AI tools to assist with manuscript organization and language editing, and to identify relevant literature. The authors reviewed and revised the AI-assisted text, checked the cited sources, and take responsibility for the final content of the paper.

References

Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the 21st International Conference on Machine Learning*, 2004.

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 22–31. PMLR, 2017. URL https://proceedings.mlr.press/v70/achiam17a.html.

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12248–12267. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.662. URL https://aclanthology.org/2024.acl-long.662/.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, et al. Constitutional AI: Harmlessness from AI feedback, 2022. URL https://arxiv.org/abs/2212.08073.

Gary Chamberlain. Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34(3):305–334, 1987.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

Harold D. Chiang, Kengo Kato, Yukun Ma, and Yuya Sasaki. Multiway cluster robust double/debiased machine learning. *Journal of Business & Economic Statistics*, 40(3):1046–1056, 2022. doi: 10.1080/07350015.2021.1895815.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

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. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Juntao Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe RLHF: Safe reinforcement learning from human feedback. In *International Conference on Learning Representations*, 2024.

Nathaniel L. Diamant and Brian L. Trippe. Calibrating generative models to feature distributions with MMD finetuning, 2026. URL https://arxiv.org/abs/2606.19496.

Marc Dymetman. Binary rewards and reinforcement learning: Fundamental challenges, 2026. URL https://arxiv.org/abs/2605.02375.

Roger Girgis, Rodrigue de Schaetzen, Luke Rowe, Azalée Robitaille, Christopher Pal, and Liam Paull. Constrained group relative policy optimization, 2026. URL https://arxiv.org/abs/2602.05863. v4, September 2026.

Úrsula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1939–1948. PMLR, 2018. URL https://proceedings.mlr.press/v80/hebert-johnson18a.html.

Cong Duy Vu Hoang, Ioan Calapodescu, and Marc Dymetman. Moment matching training for neural machine translation: A preliminary study, 2018. URL https://arxiv.org/abs/1812.09836.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948. doi: 10.1214/aoms/1177730196.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611. Association for Computational Linguistics, 2017. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

Jean Kaddour. Target policy optimization, 2026. URL https://arxiv.org/abs/2604.06159.

Muhammad Khalifa, Hady Elsahar, and Marc Dymetman. A distributional approach to controlled text generation. In *International Conference on Learning Representations*, 2021.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Jyrki Kivinen and Manfred K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–63, 1997. doi: 10.1006/inco.1996.2612.

Tomasz Korbak, Hady Elsahar, Germán Kruszewski, and Marc Dymetman. On reinforcement learning and distribution matching for fine-tuning language models with no catastrophic forgetting. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023. doi: 10.1145/3600006.3613165. URL https://doi.org/10.1145/3600006.3613165.

Philippe Laban, Lidiya Murakhovs'ka, Caiming Xiong, and Chien-Sheng Wu. Are you sure? Challenging LLMs leads to performance drops in the FlipFlop experiment, 2023. URL https://arxiv.org/abs/2311.08596.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. DExperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 6691–6706. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.522. URL https://aclanthology.org/2021.acl-long.522/.

Jiacai Liu, Yingru Li, Yuqian Fu, Jiawei Wang, Qian Liu, and Yu Shen. When speed kills stability: Demystifying RL collapse from the training-inference mismatch. Blog post, 2025. URL https://richardli.xyz/rl-collapse.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Ximing Lu, Sean Welleck, Jack Hessel, Liwei Jiang, Lianhui Qin, Peter West, Prithviraj Ammanabrolu, and Yejin Choi. Quark: Controllable text generation with reinforced unlearning. In *Advances in Neural Information Processing Systems*, 2022.

Meta. Llama 3.2 model card, 2024. URL https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct.

Ted Moskovitz, Aaditya K. Singh, DJ Strouse, Tuomas Sandholm, Ruslan Salakhutdinov, Anca D. Dragan, and Stephen McAleer. Confronting reward model overoptimization with constrained RLHF. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gkfUvn0fLU.

Tong Mu, Alec Helyar, Johannes Heidecke, Joshua Achiam, Andrea Vallone, Ian Kivlichan, Molly Lin, Alex Beutel, John Schulman, and Lilian Weng. Rule based rewards for language model safety. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2411.01111.

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. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://papers.nips.cc/paper_files/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.

Yun Qu, Qi Wang, Yixiu Mao, Heming Zou, Yuhang Jiang, Yingyue Li, Wutong Xu, Lizhou Cai, Weijie Liu, Clive Bai, Kai Yang, Yangkun Chen, Saiyong Yang, and Xiangyang Ji. Listwise policy optimization: Group-based RLVR as target-projection on the LLM response simplex, 2026. URL https://arxiv.org/abs/2605.06139.

Julien Roy, Roger Girgis, Joshua Romoff, Pierre-Luc Bacon, and Chris J. Pal. Direct behavior specification via constrained reinforcement learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 18828–18843. PMLR, 2022. URL https://proceedings.mlr.press/v162/roy22a.html.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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. *arXiv preprint arXiv:2402.03300*, 2024.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.13548.

Xudong Shen, Chao Du, Tianyu Pang, Min Lin, Yongkang Wong, and Mohan Kankanhalli. Fine-tuning text-to-image diffusion models for fairness. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2311.07604.

Henry D. Smith, Nathaniel L. Diamant, and Brian L. Trippe. Calibrating generative models to distributional constraints. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2510.10020.

Adam Stooke, Joshua Achiam, and Pieter Abbeel. Responsive safety in reinforcement learning by PID Lagrangian methods. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9133–9143, 2020. URL https://proceedings.mlr.press/v119/stooke20a.html.

Sina Tayebati, Divake Kumar, Nastaran Darabi, Dinithi Jayasuriya, Theja Tulabandhula, Ranganath Krishnan, and Amit Ranjan Trivedi. CAP: Conformalized abstention policies for context-adaptive risk management for LLMs and VLMs. In *Proceedings of the 17th Asian Conference on Machine Learning*, volume 304 of *Proceedings of Machine Learning Research*, pp. 926–941. PMLR, 2025. URL https://proceedings.mlr.press/v304/tayebati26a.html.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. TRL: Transformers reinforcement learning, 2020. URL https://github.com/huggingface/trl.

Bokun Wang and Tianbao Yang. Finite-sum coupled compositional stochastic optimization: Theory and applications. In *International Conference on Machine Learning*, 2022.

Mengdi Wang, Ethan X. Fang, and Han Liu. Stochastic compositional gradient descent: Algorithms for minimizing compositions of expected-value functions. *Mathematical Programming*, 161: 419–449, 2017.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992. doi: 10.1007/BF00992696. URL https://doi.org/10.1007/BF00992696.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6/.

An Yang, Baosong Yang, Beichen Zhang, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024a. URL https://arxiv.org/abs/2412.15115.

Kevin Yang and Dan Klein. FUDGE: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3511–3535. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.276. URL https://aclanthology.org/2021.naacl-main.276/.

Rui Yang, Xiaoman Pan, Feng Luo, Shuang Qiu, Han Zhong, Dong Yu, and Jianshu Chen. Rewards-in-context: Multi-objective alignment of foundation models with dynamic preference adjustment. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 56276–56297. PMLR, 2024b. URL https://proceedings.mlr.press/v235/yang24q.html.

Feng Yao, Liyuan Liu, Dinghuai Zhang, Chengyu Dong, Jingbo Shang, and Jianfeng Gao. Your efficient RL framework secretly brings you off-policy RL training. Blog post, 2025. URL https://fengyao.notion.site/off-policy-rl.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, et al. DAPO: An open-source LLM reinforcement learning system at scale. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Tom Zahavy, Brendan O'Donoghue, Guillaume Desjardins, and Satinder Singh. Reward is enough for convex MDPs. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvári, and Mengdi Wang. Variational policy gradient method for reinforcement learning with general utilities. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

APPENDIX OVERVIEW

This appendix provides the methodological and theoretical details supporting RLTB, along with supplementary experiments, diagnostic analyses, and reproducibility information. It is organized into five sections that follow the development of the main paper.

Appendix A: Method Details and Proofs. This section derives target-dependent output credit for general measurement families, establishes the behavior dynamics and LOPO estimation results, and presents the complete TB-GRPO algorithm.

Appendix B: Numerical Verification and Estimator Diagnostics. Numerical experiments check the behavior dynamics and estimator bias in finite problems with exactly computable gradients, and examine credit estimation across sampling budgets using stored LLM outputs.

Appendix C: Implementation and Experimental Protocols. The experimental protocols specify the training and sampling setup, task-specific behavioral measurements and targets, baseline adaptations, and the evaluation and statistical reporting procedures used in the experiments.

Appendix D: Additional Experimental Results. Additional experiments compare performance across target specifications and base models, examine the roles of feedback direction and magnitude, and assess sensitivity to measurement definitions.

Appendix E: Discussion and Reproducibility. This section discusses how RLTB relates to neighboring formulations, outlines its scope and future directions, and documents software, dataset, and artifact information for reproduction.

A METHOD DETAILS AND PROOFS

This section develops the RLTB objective and output credit for general measurement families, establishes the resulting behavior dynamics, and analyzes LOPO gradient estimation under independent-cluster sampling. Single-output, paired, and interventional measurements illustrate the grouped constructions used in the complete TB-GRPO update.

A.1 NOTATION AND ASSUMPTIONS

We use the main-text definitions of $\mathbf{m}_\theta$, $\boldsymbol{\mu}(\theta)$, $\mathcal{T}$, D_s, $\mathbf{d}$, and $\mathbf{w}_\theta$. Table 3 collects the additional notation for these families and their grouped estimators.

Write $S_\theta(q,o) := \nabla_\theta \log \pi_\theta(o \mid q)$ for the single-output policy score. Let family f use L_f policy-generated slots. Its prompt tuple follows a fixed law ν_f, after which the outputs are conditionally independent:

$$P_\theta^{(f)}(\mathrm{d}x^{(f)}) := \nu_f(\mathrm{d}q_{1:L_f}) \prod_{\ell=1}^{L_f} \pi_\theta(\mathrm{d}o_\ell \mid q_\ell), \qquad x^{(f)} := (q_{1:L_f}, o_{1:L_f}), \tag{11}$$

$$S_\theta^{(f)}(x^{(f)}) := \nabla_\theta \log \prod_{\ell=1}^{L_f} \pi_\theta(o_\ell \mid q_\ell) = \sum_{\ell=1}^{L_f} \nabla_\theta \log \pi_\theta(o_\ell \mid q_\ell). \tag{12}$$

Single-output, paired, and interventional measurements are three examples:

single:	$L_f = 1$,	$\nu_f(\mathrm{d}q_1) = P_Q(\mathrm{d}q_1)$,	$x^{(f)} = (q, o)$,
paired:	$L_f = 2$,	$\nu_f(\mathrm{d}q_1, \mathrm{d}q_2) = P_Q(\mathrm{d}q_1)\delta_{q_1}(\mathrm{d}q_2)$,	$x^{(f)} = (q, o, o')$,
interventional:	$L_f = 2$,	$\nu_f(\mathrm{d}q_1, \mathrm{d}q_2) = P_Q(\mathrm{d}q_1)K_{\mathrm{int}}(\mathrm{d}q_2 \mid q_1)$,	$x^{(f)} = (q, q', o, o')$.

Here δ_q denotes a point mass at prompt q, and $K_{\mathrm{int}}(\cdot \mid q)$ is the fixed intervention kernel for the related prompt. Every prompt law, intervention kernel, and measurement map is fixed independently of θ. Expectations under $P_\theta^{(f)}$ are denoted $\mathbb{E}_\theta^{(f)}$. For vector-valued maps, $\partial \mathbf{v}/\partial\theta$ denotes the Jacobian with the coordinates of $\mathbf{v}$ as rows.

Table 3: Additional notation. Main target-behavior objects are defined in Section 2.

Symbol	Meaning
f, L_f, ν_f	measurement family, number of policy-output slots, and fixed prompt-tuple law
$P_\theta^{(f)}, S_\theta^{(f)}$	product-policy family law and additive score over its generated slots
$\boldsymbol{\varphi}^{(f)}, \mathbf{m}_\theta^{(f)}$	primitive measurement block and its expectation
B, G	independent measurement clusters per update and outputs per prompt bank
$\mathbf{H}_i$	unbiased primitive-moment summary of independent cluster i
$\mathbf{M}_\theta$	transpose of the stacked primitive-moment Jacobian, $(\partial \mathbf{m}_\theta / \partial \theta)^\top$
$\mathbf{X}_i$	cluster score–measurement matrix with expectation $\mathbf{M}_\theta$
$\widehat{\mathbf{m}}_{-i}, \widehat{\mathbf{w}}_{-i}$	moment and measurement weight computed without receiving cluster i
$\Sigma_\varphi^{\mathrm{w}}, \Sigma_\varphi^{\mathrm{b}}$	within- and between-prompt covariance of single-output measurements

All score-function identities below assume locally common policy support, integrability, and domination sufficient to interchange differentiation and expectation. Prompt laws, intervention kernels, measurement maps, targets, and normalization scales do not depend on θ. Conditional coordinates also require the positive-denominator guard stated in Assumption 1.

A.2 Target Objective and Output Credit

Attainability means that some $\theta \in \Theta_{\mathcal{D}}$ satisfies $\boldsymbol{\mu}(\theta) \in \mathcal{T}$; it is an assumption on the policy class, not a consequence of the target specification.

Proposition 2 (Feasibility and zero discrepancy)**.** *If the joint target region is attainable, then*

$$\theta^\star \in \arg\min_{\theta \in \Theta_{\mathcal{D}}} \mathcal{L}_{\mathrm{RLTB}}(\theta) \quad \Longleftrightarrow \quad \boldsymbol{\mu}(\theta^\star) \in \mathcal{T}. \tag{13}$$

Proof. Because D_s is invertible and the fixed points of the Euclidean projection $\Pi_{\mathcal{T}}$ are exactly the points in $\mathcal{T}$, the scaled displacement in Equation 3 vanishes if and only if $\boldsymbol{\mu}(\theta) \in \mathcal{T}$. The squared norm defining $\mathcal{L}_{\mathrm{RLTB}}$ is therefore nonnegative and vanishes exactly on the target region. Attainability makes its minimum value zero, so its minimizers are precisely the target-satisfying policies.

Lemma 1 (Blockwise behavior gradient)**.** *Suppose the log-derivative identity is valid for every component of every $\mathbf{m}_\theta^{(f)}$, and Ψ is differentiable at $\mathbf{m}_\theta$. Then*

$$\frac{\partial \mathbf{m}_\theta^{(f)}}{\partial \theta} = \mathbb{E}_\theta^{(f)}\left[\boldsymbol{\varphi}^{(f)}(x^{(f)}) S_\theta^{(f)}(x^{(f)})^\top\right], \tag{14}$$

$$\frac{\partial \boldsymbol{\mu}(\theta)}{\partial \theta} = J_\Psi(\mathbf{m}_\theta)\, \mathrm{col}_{f=1}^{F}\left(\mathbb{E}_\theta^{(f)}[\boldsymbol{\varphi}^{(f)} S_\theta^{(f)\top}]\right). \tag{15}$$

Proof. Only the L_f conditional policy laws in Equation 11 depend on θ; their product score is the sum in Equation 12. Differentiation under the expectation gives the first identity. Stacking the family Jacobians and applying the chain rule to $\boldsymbol{\mu}(\theta) = \Psi(\mathbf{m}_\theta)$ gives the second.

For the conditional single-output specialization $\boldsymbol{\varphi} = (aY, a)^\top$ and $\Psi(u, v) = u/v$, where a is the conditioning indicator and Y is the outcome measured on that event, Lemma 1 gives

$$\nabla_\theta \mu(\theta) = \mathbb{E}_\theta\left[\frac{a(q,o)\{Y(q,o) - \mu(\theta)\}}{\mathbb{E}_\theta[a]} S_\theta(q,o)\right]. \tag{16}$$

For correctness conditional on answering, let A indicate answering, C indicate a correct answer, $v = \mathbb{E}_\theta[A] > 0$, and $\mu = \mathbb{E}_\theta[C]/v$. For this coordinate's target interval I, the correction $d = s^{-2}\{\Pi_I(\mu) - \mu\}$ gives the credit contribution

$$r^{\mathrm{TB}}(q,o) = \frac{d}{v}\{C(q,o) - \mu A(q,o)\}. \tag{17}$$

For $0 < \mu < 1$ below the correctness lower bound, correct answers receive positive credit, incorrect answers receive negative credit, and abstentions receive zero credit from this requirement. Contributions from the other requirements are added through the same target-to-credit construction.

Lemma 2 (Direction of the quadratic target loss)**.** *The negative gradient of the scaled target loss in behavior space satisfies*

$$-\left[\nabla_{\boldsymbol{z}} \tfrac{1}{2} \left\| D_s^{-1}\{\boldsymbol{z} - \Pi_{\mathcal{T}}(\boldsymbol{z})\}\right\|_2^2\right]_k \Bigg|_{\boldsymbol{z}=\boldsymbol{\mu}(\theta)} = s_k^{-2}([\Pi_{\mathcal{T}}(\boldsymbol{\mu}(\theta))]_k - \mu_k(\theta))\,. \tag{18}$$

It is zero inside $\mathcal{T}_k$, positive below its lower endpoint, and negative above its upper endpoint.

Proof. For a closed interval I, the derivative of $x \mapsto \frac{1}{2}s^{-2}\operatorname{dist}(x, I)^2$ is $s^{-2}\{x - \Pi_I(x)\}$, including at the endpoints, where Π_I denotes Euclidean projection onto I. Applying this identity coordinatewise proves the result.

Proposition 3 (Target-directed policy gradient)**.** *Suppose the log-derivative identity holds for every primitive measurement and Ψ is differentiable at $\mathbf{m}_\theta$. The prompt law of each measurement family, the measurements, the targets, and the scales are fixed independently of θ. Then*

$$-\nabla_\theta \mathcal{L}_{\mathrm{RLTB}}(\theta) = \sum_{f=1}^{F} \mathbb{E}_\theta^{(f)}\left[r_{\theta,f}^{\mathrm{TB}}(x^{(f)}) S_\theta^{(f)}(x^{(f)})\right]. \tag{19}$$

Proof. Let $\mathbf{d}_\theta := \mathbf{d}(\boldsymbol{\mu}(\theta))$, and partition $\mathbf{w}_\theta = J_\Psi(\mathbf{m}_\theta)^\top \mathbf{d}_\theta$ into family blocks. The chain rule and Lemma 1 give

$$\begin{aligned} -\nabla_\theta \mathcal{L}_{\mathrm{RLTB}} &= \left\{ J_\Psi(\mathbf{m}_\theta)\, \mathrm{col}_{f=1}^{F}\, \mathbb{E}_\theta^{(f)}[\boldsymbol{\varphi}^{(f)} S_\theta^{(f)\top}] \right\}^\top \mathbf{d}_\theta \\ &= \sum_{f=1}^{F} \mathbb{E}_\theta^{(f)}[S_\theta^{(f)} \boldsymbol{\varphi}^{(f)\top}] \mathbf{w}_\theta^{(f)} = \sum_{f=1}^{F} \mathbb{E}_\theta^{(f)}[r_{\theta,f}^{\mathrm{TB}} S_\theta^{(f)}]. \end{aligned}$$

A.3 Proof of Proposition 1

Proof. For a product of closed intervals, the scaled squared distance is continuously differentiable, including at the interval endpoints, with behavior-space gradient $-\mathbf{d}(\boldsymbol{\mu})$. The chain rule therefore gives $-\nabla_\theta \mathcal{L}_{\mathrm{RLTB}} = J_\mu^\top \mathbf{d}$. Along the stated trajectory,

$$\dot{\boldsymbol{\mu}} = J_\mu \dot{\theta} = J_\mu J_\mu^\top \mathbf{d}, \qquad \frac{\mathrm{d}}{\mathrm{d}t}\mathcal{L}_{\mathrm{RLTB}} = \nabla_\theta \mathcal{L}_{\mathrm{RLTB}}^\top \dot{\theta} = -\|J_\mu^\top \mathbf{d}\|_2^2.$$

This holds for the behavior map constructed from the stacked moments of all declared families. In particular, $\dot{\mu}_k = \sum_\ell \langle \nabla_\theta \mu_k, \nabla_\theta \mu_\ell \rangle d_\ell$, so $d_k = 0$ alone need not make $\dot{\mu}_k = 0$. Since all scales are positive, $\mathbf{d} = 0$ is equivalent to joint satisfaction, whereas $J_\mu^\top \mathbf{d} = 0$ can also occur outside that region. These are statements about any existing gradient-flow trajectory remaining in the differentiability domain, not a claim of global existence or convergence. Conditional coordinates require their denominators to remain positive on this domain.

A.4 LOPO Gradient Estimation

A.4.1 Proof of Theorem 1

Proof. $\widehat{\mathbf{w}}_{-i}$ is measurable with respect to the other $B-1$ clusters and is therefore independent of $\mathbf{X}_i$. Hence $\mathbb{E}[\mathbf{X}_i \widehat{\mathbf{w}}_{-i}] = \mathbf{M}_\theta \mathbb{E}[\widehat{\mathbf{w}}_{-i}]$, proving Equation 9. Moreover $\mathbb{E}[\widehat{\mathbf{m}}_{-i}] = \mathbf{m}_\theta$, so an affine weight map with the same coefficients on the fold support and at $\mathbf{m}_\theta$ commutes with expectation. In the Lipschitz case, Equation 9, the operator-norm bound, Lipschitz continuity, and Jensen's inequality give

$$\|\mathbb{E}[\widehat{\mathbf{g}}] + \nabla_\theta \mathcal{L}_{\mathrm{RLTB}}(\theta)\|_2 \le \kappa_m L_w \sqrt{\operatorname{tr}\{\operatorname{Cov}(\widehat{\mathbf{m}}_{-i})\}},$$

and $\operatorname{Cov}(\widehat{\mathbf{m}}_{-i}) = \operatorname{Cov}(\mathbf{H}_i)/(B-1)$ proves Equation 10.

The residual plug-in bias is related to stochastic compositional optimization (Wang et al., 2017; Wang & Yang, 2022); the bound here uses the covariance of the complete measurement-cluster summary.

Table 4: Measurement-weight estimators under prompt-grouped sampling. Coupling refers to dependence between a receiving group's score contribution and the weight applied to it.

Estimator	Left-out unit	Weight source	Receiving-group coupling	Remaining error
Full-batch	none	full batch	self and sibling	weight-estimation bias plus coupling
Rollout LOO	one output	all other outputs, including siblings	sibling	weight-estimation bias plus sibling coupling
LOPO	one prompt group	the other $B-1$ groups	none	nonlinear weight-estimation bias only

A.4.2 Single-Output Construction and Leave-Out Rules

For single-output measurements, the full-batch, rollout-LOO, and LOPO estimators differ in which sampling unit is excluded when estimating the measurement weights (Table 4). For the single-output constructions below, write $\mathbf{H}_i := G^{-1}\sum_j \boldsymbol{\varphi}_{ij}$, $S_\theta(q,o) := \nabla_\theta \log \pi_\theta(o \mid q)$, $S_{ij} := S_\theta(q_i, o_{ij})$, and $\mathbf{X}_i := (G-1)^{-1}\sum_j S_{ij}(\boldsymbol{\varphi}_{ij} - \mathbf{H}_i)^\top$. Here $\theta = \theta_{\text{old}}$ and $\mathbf{M}_\theta := \mathbb{E}_\theta[S_\theta(q,o)\boldsymbol{\varphi}(q,o)^\top]$.

Proposition 4 (Exact group-relative representation)**.** *For a single-output specification ($F = 1$) and $G \geq 2$, define $A_{ij}^{\mathrm{TB}} := \frac{G}{G-1}\mathbf{w}_\theta^\top(\boldsymbol{\varphi}_{ij} - \mathbf{H}_i)$. For on-policy evaluation at collection time and under Lemma 1,*

$$\sum_{j=1}^{G} A_{ij}^{\mathrm{TB}} = 0, \qquad \mathbb{E}[A_{ij}^{\mathrm{TB}} S_{ij}] = -\nabla_\theta \mathcal{L}_{\mathrm{RLTB}}(\theta). \tag{20}$$

Proof. With $\bar{\boldsymbol{\varphi}}_{i,-j} = (G-1)^{-1}\sum_{j' \neq j} \boldsymbol{\varphi}_{ij'}$, we have $\frac{G}{G-1}(\boldsymbol{\varphi}_{ij} - \mathbf{H}_i) = \boldsymbol{\varphi}_{ij} - \bar{\boldsymbol{\varphi}}_{i,-j}$. Conditionally on q_i, the sibling mean is independent of o_{ij}, and $\mathbb{E}[S_{ij} \mid q_i] = 0$. Its expected score product therefore vanishes. The single-family instance of Proposition 3 gives the remaining equality; centering gives the zero-sum identity.

Lemma 3 (Single-output cluster bias decomposition)**.** *Under the sampling conditions in Section 2.3, let $\theta = \theta_{\text{old}}$ and $B, G \geq 2$. Whenever the displayed expectations are finite,*

$$\mathbb{E}[\mathbf{H}_i] = \mathbf{m}_\theta, \qquad \mathbb{E}[\mathbf{X}_i] = \mathbf{M}_\theta, \qquad \mathbf{M}_\theta \mathbf{w}_\theta = -\nabla_\theta \mathcal{L}_{\mathrm{RLTB}}(\theta), \tag{21}$$

and, for any batch-based estimator $\widetilde{\mathbf{w}}_i$ of $\mathbf{w}_\theta$,

$$\begin{aligned}\mathbb{E}[\mathbf{X}_i \widetilde{\mathbf{w}}_i] - \mathbf{M}_\theta \mathbf{w}_\theta = \mathbf{M}_\theta\{\mathbb{E}[\widetilde{\mathbf{w}}_i] - \mathbf{w}_\theta\} \\ + \mathbb{E}[(\mathbf{X}_i - \mathbf{M}_\theta)(\widetilde{\mathbf{w}}_i - \mathbb{E}[\widetilde{\mathbf{w}}_i])].\end{aligned} \tag{22}$$

The first term is the contribution of bias in estimating $\mathbf{w}_\theta$. The second is receiving-group coupling and vanishes when $\widetilde{\mathbf{w}}_i$ is independent of group i.

Proof. Unbiasedness of $\mathbf{H}_i$ follows from the sampling law. Proposition 4, applied to each primitive measurement, gives $\mathbb{E}[\mathbf{X}_i] = \mathbf{M}_\theta$; the target-gradient identity gives $\mathbf{M}_\theta \mathbf{w}_\theta = -\nabla_\theta \mathcal{L}_{\mathrm{RLTB}}$. For the decomposition, add and subtract $\mathbf{M}_\theta \mathbb{E}[\widetilde{\mathbf{w}}_i]$, then collect the mean-product and covariance terms. The same decomposition applies to general cluster matrices with expectation $\mathbf{M}_\theta$.

The single-output instance of the estimator in Section 3 is

$$\widehat{\mathbf{g}} = \frac{1}{BG}\sum_{i,j} \widehat{A}_{ij}^{\mathrm{TB}} S_{ij} = \frac{1}{B}\sum_i \mathbf{X}_i \widehat{\mathbf{w}}_{-i}. \tag{23}$$

Leaving out one output generally retains its $G-1$ same-prompt siblings. To make the resulting coupling explicit, consider the quadratic point-target specialization with $\Psi = \mathrm{Id}$, $\mathcal{T} = \{\boldsymbol{\mu}^\star\}$, and $d_\varphi = K$. Define the whitened feature, its expectation, and the between-prompt covariance by

$$\boldsymbol{\psi}_{ij} := D_s^{-1}(\boldsymbol{\varphi}_{ij} - \boldsymbol{\mu}^\star), \qquad \boldsymbol{\delta} := \mathbb{E}[\boldsymbol{\psi}_{ij}] = D_s^{-1}(\mathbf{m}_\theta - \boldsymbol{\mu}^\star), \qquad \Sigma^{\mathrm{b}} := \mathrm{Cov}_q(\mathbb{E}[\boldsymbol{\psi}_{ij} \mid q]), \tag{24}$$

so that $\mathcal{L}_{\mathrm{RLTB}}(\theta) = \frac{1}{2}\|\boldsymbol{\delta}\|_2^2$. With $n := BG$, define $\widehat{\boldsymbol{\delta}}_{-(ij)}$ from the other $n-1$ outputs and write $\widehat{\mathbf{w}}_{-(ij)} := -D_s^{-1}\widehat{\boldsymbol{\delta}}_{-(ij)}$. The corresponding output credit is $\widehat{\mathbf{w}}_{-(ij)}^\top \boldsymbol{\varphi}_{ij}$. Subtracting $\widehat{\mathbf{w}}_{-(ij)}^\top \boldsymbol{\mu}^\star$, which is independent of o_{ij} conditional on q_i and hence score-orthogonal, gives the score-equivalent output credit for rollout LOO

$$\widehat{\boldsymbol{\delta}}_{-(ij)} := \frac{1}{n-1}\sum_{(i',j')\neq(i,j)} \boldsymbol{\psi}_{i'j'}, \qquad \widetilde{r}_{ij}^{\mathrm{LOO}} := -\widehat{\boldsymbol{\delta}}_{-(ij)}^\top \boldsymbol{\psi}_{ij}. \tag{25}$$

Its expected score contribution is

$$-\mathbb{E}[\widetilde{r}_{ij}^{\mathrm{LOO}} S_{ij}] = \nabla_\theta \mathcal{L}_{\mathrm{RLTB}}(\theta) + \frac{G-1}{2(BG-1)}\nabla_\theta \operatorname{tr}(\Sigma^{\mathrm{b}}). \tag{26}$$

The extra term comes from the retained ordered pairs that share a prompt. Excluding the complete prompt group removes those pairs.

A.4.3 Paired and Interventional Constructions

Table 5: Examples of measurement families and their grouped estimators. Each row supplies one block of the stacked moment and measurement weight; the last column is the score of one family sample.

Family	Independent cluster	Unbiased summary	Additive score
Single	prompt and G outputs	group mean of $\boldsymbol{\varphi}(q,o)$	$S_\theta(q,o)$
Paired	prompt and all sibling outputs	order-two U-statistic with a symmetric kernel	$S_\theta(q,o)+S_\theta(q,o')$
Interventional	prompt pair and both output banks	two-bank cross mean	$S_\theta(q,o)+S_\theta(q',o')$

Table 5 compares single-output, paired, and interventional instances of the general family construction in Appendix A.1. For symmetric measurements, define local pair averages and their U-statistic (Hoeffding, 1948):

$$\mathbf{v}_{ij}^{(2)} := \frac{1}{G-1}\sum_{j'\neq j} \boldsymbol{\varphi}^{(2)}(q_i, o_{ij}, o_{ij'}),$$
$$\mathbf{H}_i^{(2)} := \frac{1}{G}\sum_j \mathbf{v}_{ij}^{(2)} = \binom{G}{2}^{-1}\sum_{j<j'} \boldsymbol{\varphi}^{(2)}(q_i, o_{ij}, o_{ij'}). \tag{27}$$

This is the canonical route-agreement construction; the scaling used in the reported implementation is specified in Appendix C.1.

Corollary 1 (Symmetric paired-output component advantage)**.** *For $B \geq 2$, $G \geq 3$, and a fold weight independent of cluster i,*

$$\widehat{A}_{ij}^{\mathrm{TB},(2)} := \frac{2G}{G-2}\widehat{\mathbf{w}}_{-i}^{(2)\top}(\mathbf{v}_{ij}^{(2)} - \mathbf{H}_i^{(2)}), \tag{28}$$

sums to zero over j and satisfies

$$\mathbb{E}\left[\frac{1}{BG}\sum_{i,j}\widehat{A}_{ij}^{\mathrm{TB},(2)} S_{ij}\right] = \mathbf{M}_\theta^{(2)}\mathbb{E}[\widehat{\mathbf{w}}_{-i}^{(2)}], \tag{29}$$

where $\mathbf{M}_\theta^{(2)} = \mathbb{E}[(S_\theta(q,o)+S_\theta(q,o'))\boldsymbol{\varphi}^{(2)\top}]$.

Proof. Symmetry gives $\mathbf{M}_\theta^{(2)} = 2\mathbb{E}[S_{i1}\boldsymbol{\varphi}^{(2)}(q_i,o_{i1},o_{i2})^\top]$. The score product with $\mathbf{v}_{ij}^{(2)}$ equals one copy of this expectation, while the product with $\mathbf{H}_i^{(2)}$ equals $2/G$ copies. The factor $2G/(G-2)$ recovers $\mathbf{M}_\theta^{(2)}$; fold independence and centering complete the proof.

For an interventional family, draw $q'_i \sim K_{\mathrm{int}}(\cdot \mid q_i)$, with G outputs from each prompt, and write $S'_{ik} := \nabla_\theta \log \pi_\theta(o'_{ik} \mid q'_i)$. Define the cross-bank summary and the measurements local to the left and right banks by

$$\mathbf{H}_i^{(I)} := \frac{1}{G^2}\sum_{j=1}^{G}\sum_{k=1}^{G}\boldsymbol{\varphi}^{(I)}(q_i, q'_i, o_{ij}, o'_{ik}),$$

$$\mathbf{u}_{ij}^{L} := \frac{1}{G}\sum_{k=1}^{G}\boldsymbol{\varphi}^{(I)}(q_i, q'_i, o_{ij}, o'_{ik}), \qquad \mathbf{u}_{ik}^{R} := \frac{1}{G}\sum_{j=1}^{G}\boldsymbol{\varphi}^{(I)}(q_i, q'_i, o_{ij}, o'_{ik}). \tag{30}$$

For $B, G \geq 2$ and a fold weight independent of cluster i, let

$$\widehat{A}_{ij}^{\mathrm{TB},I,L} := \frac{G}{G-1}\widehat{\mathbf{w}}_{-i}^{(I)\top}(\mathbf{u}_{ij}^{L} - \mathbf{H}_i^{(I)}), \qquad \widehat{A}_{ik}^{\mathrm{TB},I,R} := \frac{G}{G-1}\widehat{\mathbf{w}}_{-i}^{(I)\top}(\mathbf{u}_{ik}^{R} - \mathbf{H}_i^{(I)}). \tag{31}$$

Each bank is zero-sum. The finite-sample component-score estimator averages the banks separately and adds them:

$$\widehat{\mathbf{g}}^{(I)} := \frac{1}{BG}\sum_{i=1}^{B}\sum_{j=1}^{G}\widehat{A}_{ij}^{\mathrm{TB},I,L}S_{ij} + \frac{1}{BG}\sum_{i=1}^{B}\sum_{k=1}^{G}\widehat{A}_{ik}^{\mathrm{TB},I,R}S'_{ik}. \tag{32}$$

It satisfies

$$\mathbb{E}[\widehat{\mathbf{g}}^{(I)}] = \mathbf{M}_\theta^{(I)}\mathbb{E}[\widehat{\mathbf{w}}_{-i}^{(I)}], \qquad \mathbf{M}_\theta^{(I)} := \mathbb{E}\Big[(S_\theta(q,o) + S_\theta(q',o'))\boldsymbol{\varphi}^{(I)\top}\Big]. \tag{33}$$

Indeed, for the left bank, $\mathbb{E}[S_{ij}\mathbf{u}_{ij}^{L\top}]$ is its full additive-score contribution, whereas $\mathbb{E}[S_{ij}\mathbf{H}_i^{(I)\top}]$ is $1/G$ of that contribution; the factor $G/(G-1)$ restores it. The same calculation applies to the right bank, and fold independence permits multiplication by $\mathbb{E}[\widehat{\mathbf{w}}_{-i}^{(I)}]$. The GSM8K specification uses single and paired families; the TriviaQA specification instantiates all three, its sixth requirement measuring an intervention (Appendix C.3.2).

Cross-fitting must exclude the whole pair. The twin's summary is a cross-bank mean of the held-out prompt's own outputs, so retaining it makes $\widehat{\mathbf{w}}_{-i}^{(I)}$ depend on the cluster it scores, and the fold independence the argument above rests on no longer holds. Centering can remain per prompt: the left bank's own mean of $\mathbf{u}_{ij}^{L}$ over j equals $\mathbf{H}_i^{(I)}$, so a group-relative baseline computed over one bank already forms $\mathbf{u}_{ij}^{L} - \mathbf{H}_i^{(I)}$, and likewise for the right bank.

To make the matrices in Theorem 1 explicit, the paired and interventional constructions above give

$$\mathbf{X}_i^{(2)} := \frac{2}{G-2}\sum_{j} S_{ij}(\mathbf{v}_{ij}^{(2)} - \mathbf{H}_i^{(2)})^\top,$$

$$\mathbf{X}_i^{(I)} := \frac{1}{G-1}\left\{\sum_{j} S_{ij}(\mathbf{u}_{ij}^{L} - \mathbf{H}_i^{(I)})^\top + \sum_{k} S'_{ik}(\mathbf{u}_{ik}^{R} - \mathbf{H}_i^{(I)})^\top\right\}. \tag{34}$$

Their expectations are respectively $\mathbf{M}_\theta^{(2)}$ and $\mathbf{M}_\theta^{(I)}$ by the preceding score identities. Horizontally concatenating the family matrices gives $\mathbf{X}_i$, with $\mathbb{E}[\mathbf{X}_i] = (\partial \mathbf{m}_\theta/\partial\theta)^\top$ and $\mathbf{X}_i\widehat{\mathbf{w}}_{-i} = \sum_f \mathbf{X}_i^{(f)}\widehat{\mathbf{w}}_{-i}^{(f)}$. For other measurement families, the theorem requires an unbiased summary and cluster-local score matrix with these expectation properties; independence alone does not establish the score representation.

When several families share a dependency-complete cluster, concatenate their summaries as $\mathbf{H}_i := \mathrm{col}_f\, \mathbf{H}_i^{(f)}$ and their weight blocks as $\widehat{\mathbf{w}}_{-i} := \mathrm{col}_f\, \widehat{\mathbf{w}}_{-i}^{(f)}$. A sampled output used by several families receives the sum of their component score contributions. Cross-fitting excludes the full dependency closure of the receiving measurement. The nonlinear guarantee uses $\mathrm{Cov}(\mathbf{H}_i)$ rather than the single-output $\Sigma_\varphi^{\mathrm{b}} + \Sigma_\varphi^{\mathrm{w}}/G$ formula for paired or interventional summaries.

When a target is estimated as the admissible-set value of the same measurement, the policy and reference measurements share its prompt-tuple law ν_f, observation window, stopping rule, and feature implementation. A threshold fixed from a different reference measurement is instead a declared constant. Once estimated, the target, scales, and reference summaries are frozen. Differentiating through a policy-dependent target would define a different objective.

A.4.4 Sampling Budgets and Conditional Denominators

For the single-output construction in Section 2.3, let $\Sigma_\varphi^{\mathrm{b}} := \mathrm{Cov}_q(\mathbb{E}[\boldsymbol{\varphi} \mid q])$ and $\Sigma_\varphi^{\mathrm{w}} := \mathbb{E}_q[\mathrm{Cov}(\boldsymbol{\varphi} \mid q)]$. Since the G outputs are conditionally i.i.d. within a prompt, the law of total covariance gives $\mathrm{Cov}(\mathbf{H}_i) = \Sigma_\varphi^{\mathrm{b}} + \Sigma_\varphi^{\mathrm{w}}/G$ for $\mathbf{H}_i = G^{-1}\sum_j \boldsymbol{\varphi}_{ij}$. Substituting this covariance into Theorem 1 gives, under its Lipschitz conditions,

$$\|\mathbb{E}[\widehat{\mathbf{g}}] + \nabla_\theta \mathcal{L}_{\mathrm{RLTB}}\|_2 \le \kappa_m L_w \sqrt{\frac{\mathrm{tr}(\Sigma_\varphi^{\mathrm{b}}) + \mathrm{tr}(\Sigma_\varphi^{\mathrm{w}})/G}{B-1}}. \tag{35}$$

The two sampling axes play different roles: G reduces only within-prompt variation, whereas B supplies independent weight-estimation units. Equivalently, for a scalar coordinate z,

$$\mathrm{Var}\left(\frac{1}{BG}\sum_{i,j} z_{ij}\right) = \frac{1}{B}\left(\sigma_{\mathrm{b},z}^2 + \frac{\sigma_{\mathrm{w},z}^2}{G}\right). \tag{36}$$

Here $\sigma_{\mathrm{b},z}^2$ and $\sigma_{\mathrm{w},z}^2$ denote the between- and within-prompt variances of z, respectively. For $G > 1$, treating all BG outputs as independent understates this variance whenever $\sigma_{\mathrm{b},z}^2 > 0$.

Assumption 1 (Conditional-coordinate guard)**.** Let $\mathcal{K}_{\mathrm{cond}}$ be the finite set of conditional coordinates. For every $k \in \mathcal{K}_{\mathrm{cond}}$, the denominator primitive is an indicator $a_k(x^{(f_k)}) \in \{0,1\}$ in a declared measurement family f_k. There are $p_{\min,k} > 0$ and a neighborhood $\mathcal{N}$ of the current policy such that $p_k(\theta') := \mathbb{E}_{\theta'}^{(f_k)}[a_k] \ge p_{\min,k}$ for every $\theta' \in \mathcal{N}$. The corresponding numerator primitives are bounded as well.

Under the independent-cluster sampling of Theorem 1, let $Z_{i,k}$ be the denominator entry of $\mathbf{H}_i$ for coordinate k, and suppose $Z_{i,k} \in [0,1]$ and $\mathbb{E}[Z_{i,k}] = p_k(\theta)$. The single-output construction uses $Z_{i,k} = G^{-1}\sum_j a_k(q_i, o_{ij})$; the paired and interventional constructions use the corresponding U-statistic or cross-bank mean. Each is a bounded unbiased summary of its denominator primitive. Define the leave-one-cluster-out denominator estimate by $\widehat{p}_{-i,k} := (B-1)^{-1}\sum_{i' \ne i} Z_{i',k}$. Its $B-1$ terms are independent across complete measurement clusters. Define the joint good-fold event $\mathcal{E}_i := \bigcap_{k \in \mathcal{K}_{\mathrm{cond}}} \{\widehat{p}_{-i,k} \ge p_{\min,k}/2\}$. Assume $\mathbf{H}_i$ is square-integrable, $\|\mathbf{M}_\theta\|_{\mathrm{op}} \le \kappa_m$, and $\mathbf{w}$ is L_w-Lipschitz on a set containing $\mathbf{m}_\theta$ and all fold estimates $\widehat{\mathbf{m}}_{-i}$ on $\mathcal{E}_i$. Let $\overline{\mathbf{w}}_{-i}$ equal $\mathbf{w}(\widehat{\mathbf{m}}_{-i})$ on $\mathcal{E}_i$, and use a predeclared fallback rule measurable only with respect to the other $B-1$ clusters otherwise. Assume $\|\overline{\mathbf{w}}_{-i} - \mathbf{w}_\theta\|_2 \le C_w$ on $\mathcal{E}_i^c$, and define the guarded score mean $\overline{\mathbf{g}} := B^{-1}\sum_{i=1}^B \mathbf{X}_i \overline{\mathbf{w}}_{-i}$. The guarded weights remain independent of $\mathbf{X}_i$. Splitting the weight error over $\mathcal{E}_i$ and its complement gives

$$\left\|\mathbb{E}[\overline{\mathbf{w}}_{-i}] - \mathbf{w}_\theta\right\|_2 \le L_w \mathbb{E}\left[\|\widehat{\mathbf{m}}_{-i} - \mathbf{m}_\theta\|_2\right] + C_w \Pr(\mathcal{E}_i^c).$$

The same independence and covariance argument as in Theorem 1 therefore yields

$$\|\mathbb{E}[\overline{\mathbf{g}}] + \nabla_\theta \mathcal{L}_{\mathrm{RLTB}}(\theta)\|_2 \le \kappa_m \left[L_w \sqrt{\frac{\mathrm{tr}\{\mathrm{Cov}(\mathbf{H}_i)\}}{B-1}} + C_w \Pr(\mathcal{E}_i^c)\right]. \tag{37}$$

Because each $Z_{i,k} \in [0,1]$, Hoeffding's inequality over independent clusters, followed by a union bound over the guarded coordinates in $\mathcal{K}_{\mathrm{cond}}$, yields

$$\Pr(\mathcal{E}_i^c) \le \sum_{k \in \mathcal{K}_{\mathrm{cond}}} \exp\left\{-\frac{(B-1)p_{\min,k}^2}{2}\right\}. \tag{38}$$

This result applies to the collection-time on-policy sequence-score estimator.

A.5 TB-GRPO Update

TB-GRPO uses the LOPO advantages from Appendix A.4 in a clipped policy-ratio update. Let $\mathcal{O}_i$ index the sampled output occurrences in complete measurement cluster i, with prompt-output pairs (q_{iu}, o_{iu}), and write $S_{iu} := \nabla_\theta \log \pi_\theta(o_{iu} \mid q_{iu})$ at $\theta = \theta_{\mathrm{old}}$. Shared outputs keep the same index across families. Each family supplies a moment summary $\mathbf{H}_i^{(f)}$ and component advantages $\widehat{A}_{iu}^{(f)}$,

set to zero for outputs unused by that family. With G outputs per prompt bank, the constructions in Appendix A.4 satisfy

$$\frac{1}{G}\sum_{u\in\mathcal{O}_i}\widehat{A}_{iu}^{(f)}S_{iu}=\mathbf{X}_i^{(f)}\widehat{\mathbf{w}}_{-i}^{(f)}. \tag{39}$$

For single-output, paired, and interventional measurements, use Equations 7, 28, and 31, respectively. The interventional contribution includes both output banks. Other families can use the same update when they supply summaries and component advantages satisfying this interface and the conditions of Theorem 1. An output shared by several families receives the summed advantage $\widehat{A}_{iu}^{\mathrm{TB}}:=\sum_f\widehat{A}_{iu}^{(f)}$, giving

$$\widehat{\mathbf{g}}=\frac{1}{BG}\sum_{i=1}^{B}\sum_{u\in\mathcal{O}_i}\widehat{A}_{iu}^{\mathrm{TB}}S_{iu}. \tag{40}$$

The factor $1/G$ in Equation 39 preserves each family's score normalization even when a cluster contains several prompt banks. For a single-output family, $u=j$ and this reduces to the construction in Section 2.3.

Algorithm 2 TB-GRPO update with LOPO component advantages

input: current θ; frozen $\{\nu_f,\boldsymbol{\varphi}^{(f)}\}_f$, $\Psi,\mathcal{T},D_s$; cluster samplers, moment summaries, and component-advantage rules satisfying Equation 39; $B\geq 2$, valid bank size G; denominator guards; clipping/KL and optimizer settings
output: updated θ

set $\theta_{\mathrm{old}}\leftarrow\theta$; sample B independent complete measurement clusters, with G outputs per prompt bank from $\pi_{\theta_{\mathrm{old}}}$
evaluate the family measurements and summaries; $\mathbf{H}_i\leftarrow\mathrm{col}_f\,\mathbf{H}_i^{(f)}$, $\mathbf{H}_{\mathrm{tot}}\leftarrow\sum_i\mathbf{H}_i$
for $i=1,\dots,B$ **do**
  $\widehat{\mathbf{m}}_{-i}\leftarrow(\mathbf{H}_{\mathrm{tot}}-\mathbf{H}_i)/(B-1)$
  if any declared denominator guard fails on $\widehat{\mathbf{m}}_{-i}$ **then**
    set $\widehat{\mathbf{w}}_{-i}\leftarrow 0$ and $\widehat{A}_{iu}^{\mathrm{TB}}\leftarrow 0$ for all $u\in\mathcal{O}_i$
  else
    $\widehat{\boldsymbol{\mu}}_{-i}\leftarrow\Psi(\widehat{\mathbf{m}}_{-i})$, $\widehat{\mathbf{d}}_{-i}\leftarrow D_s^{-2}\{\Pi_{\mathcal{T}}(\widehat{\boldsymbol{\mu}}_{-i})-\widehat{\boldsymbol{\mu}}_{-i}\}$
    $\widehat{\mathbf{w}}_{-i}\leftarrow J_\Psi(\widehat{\mathbf{m}}_{-i})^\top\widehat{\mathbf{d}}_{-i}$; partition into family blocks
    form each $\widehat{A}_{iu}^{(f)}$ using its family rule and $\widehat{\mathbf{w}}_{-i}^{(f)}$
    $\widehat{A}_{iu}^{\mathrm{TB}}\leftarrow\sum_f\widehat{A}_{iu}^{(f)}$ for each output occurrence $u\in\mathcal{O}_i$
  end if
end for
stop gradients through the weights and advantages
take the prescribed optimizer steps to increase $\widehat{\mathcal{J}}_{\text{TB-GRPO}}(\theta)$, keeping $\pi_{\theta_{\mathrm{old}}}$ and $\{\widehat{A}_{iu}^{\mathrm{TB}}\}$ fixed
refresh the collection policy and advantages with the next rollout batch

For the clipped surrogate, the token-level likelihood ratio is

$$\rho_{iut}(\theta):=\frac{\pi_\theta(o_{iu,t}\mid q_{iu},o_{iu,<t})}{\pi_{\theta_{\mathrm{old}}}(o_{iu,t}\mid q_{iu},o_{iu,<t})}. \tag{41}$$

For output-level advantages $\widehat{A}_{iu}$, we adopt DAPO's token-level loss reduction (Yu et al., 2025), averaging over $N_{\mathrm{tok}}:=\sum_i\sum_{u\in\mathcal{O}_i}|o_{iu}|$ active completion tokens. Each sampled occurrence is counted once, including when it is used by several families. With a KL penalty relative to a fixed reference policy π_{ref}, the surrogate is

$$\widehat{\mathcal{J}}_{\mathrm{GRPO}}(\theta;\{\widehat{A}_{iu}\}):=\frac{1}{N_{\mathrm{tok}}}\sum_{i=1}^{B}\sum_{u\in\mathcal{O}_i}\sum_t c_\epsilon\Big(\rho_{iut}(\theta),\widehat{A}_{iu}\Big)-\beta\widehat{D}_{\mathrm{KL}}(\theta;\pi_{\mathrm{ref}}), \tag{42}$$

where $c_\epsilon(\rho,A):=\min\{\rho A,\mathrm{clip}(\rho,1-\epsilon,1+\epsilon)A\}$, $\epsilon>0$ is the clipping range, and $\beta\geq 0$ controls the KL penalty, averaged over the same active-token set. Substituting the assembled advantages gives $\widehat{\mathcal{J}}_{\text{TB-GRPO}}(\theta):=\widehat{\mathcal{J}}_{\mathrm{GRPO}}(\theta;\{\widehat{A}_{iu}^{\mathrm{TB}}\})$. The collection policy, weights, and advantages are held fixed throughout the update. Advantages are broadcast to completion tokens without within-group standard-deviation normalization.

Algorithm 2 uses zero weights when a fold fails a denominator guard. The constructions above require $G \geq 2$ for single-output and interventional families and $G \geq 3$ for paired families. At collection time, with $\beta = 0$ and no sampler correction,

$$\nabla_\theta \widehat{\mathcal{J}}_{\text{TB-GRPO}}(\theta)\Big|_{\theta=\theta_{\text{old}}} = \frac{BG}{N_{\text{tok}}}\widehat{\mathbf{g}}. \tag{43}$$

The estimation guarantees in Theorem 1 and Equation 37 concern the canonical sequence-score estimator $\widehat{\mathbf{g}}$; the random token normalization and subsequent clipped updates are additional optimizer choices. Appendix C.1 records the family rescaling, sampler correction, and optimizer conventions used in the reported runs.

A.6 Connections to Fixed Rewards and Expected-Cost Constraints

Proposition 5 (Constant-weight fixed-feedback objectives)**.** *For a single-output family, suppose $\boldsymbol{\mu}(\theta) = C\mathbf{m}_\theta + \mathbf{b}$, where C and $\mathbf{b}$ are fixed, and let $\mathcal{L}_{\text{aff}}(\theta) := c + \boldsymbol{\lambda}^\top \boldsymbol{\mu}(\theta)$. Then*

$$-\nabla_\theta \mathcal{L}_{\text{aff}}(\theta) = \mathbb{E}_\theta[r(q,o)S_\theta(q,o)], \qquad r(q,o) := -(C^\top \boldsymbol{\lambda})^\top \boldsymbol{\varphi}(q,o), \tag{44}$$

so the fixed measurement weight is independent of θ.

Proof. The claim follows by substituting $\partial\boldsymbol{\mu}/\partial\theta = C\mathbb{E}_\theta[\boldsymbol{\varphi}S_\theta^\top]$ into the chain rule.

This proposition contains verifier reward and static scalarization as affine special cases. These fixed weights lack the zero-set semantics of Proposition 2: for a binary output with probability p, the derivative of any unregularized fixed-reward objective is constant in p; the target-loss direction $p^\star - p$ instead changes sign at an interior destination.

Proposition 6 (Non-portability of binary fixed feedback under KL regularization)**.** *For binary KL control (Dymetman, 2026), let $\varphi_1 \in \{0,1\}$ have base rate $p_0 \in (0,1)$. For $\beta > 0$, consider*

$$\max_{p\in[0,1]} \{\gamma p - \beta D_{\text{KL}}(\text{Bernoulli}(p) \,\|\, \text{Bernoulli}(p_0))\}. \tag{45}$$

Its unique solution, and the fixed-feedback strength required for a requested $p^\star \in (0,1)$, are

$$p_\gamma = \sigma\left(\text{logit}(p_0) + \frac{\gamma}{\beta}\right), \qquad \gamma(p^\star; p_0, \beta) = \beta\{\text{logit}(p^\star) - \text{logit}(p_0)\}. \tag{46}$$

Here $\sigma(x) = (1+e^{-x})^{-1}$ is the logistic sigmoid. Thus no fixed-feedback strength independent of (p_0, β) realizes the same interior target across base policies and KL strengths. In this one-prompt Bernoulli reduction, without KL, any nonzero reward contrast selects an endpoint.

Proof. The objective has derivative $\gamma - \beta \log \frac{p(1-p_0)}{p_0(1-p)}$ and second derivative $-\beta\{p(1-p)\}^{-1} < 0$. Solving the unique first-order condition gives the stated solution and required feedback strength. Removing KL leaves an affine function of p, whose nonconstant cases attain their maxima at an endpoint.

The discrepancy objective alone has its minimum at the requested $p^\star$. Adding KL introduces a preference for the base policy: the minimizer of $\frac{(p-p^\star)^2}{2s^2} + \beta D_{\text{KL}}(\text{Bernoulli}(p) \| \text{Bernoulli}(p_0))$ for $\beta > 0$ satisfies

$$\frac{p - p^\star}{s^2} + \beta\{\text{logit}(p) - \text{logit}(p_0)\} = 0.$$

It therefore generally differs from $p^\star$ when $p_0 \neq p^\star$. The zero-discrepancy characterization in Proposition 2 concerns the original feasibility objective; competing preferences introduce a trade-off, as developed in Appendix A.6.1.

Proposition 7 (Single-output target boxes as expected-cost constraints)**.** *Suppose*

$$\mu_k(\theta) := \frac{\mathbb{E}_\theta[h_k(q,o)]}{\mathbb{E}_\theta[a_k(q,o)]}, \qquad \mathbb{E}_\theta[a_k(q,o)] \geq p_{\min,k} > 0. \tag{47}$$

Each finite lower or upper endpoint of $\mathcal{T}_k = [\ell_k, u_k] \cap \mathbb{R}$ is respectively equivalent to

$$\mathbb{E}_\theta[\ell_k a_k - h_k] \le 0, \qquad \mathbb{E}_\theta[h_k - u_k a_k] \le 0. \tag{48}$$

Hence a finite product of guarded single-output target intervals is representable by finitely many expected-cost constraints.

Proof. The denominator is positive, so multiplying either endpoint inequality by it preserves the inequality direction and gives the displayed expected-cost inequalities. Infinite endpoints add no constraint; the guard itself is $\mathbb{E}_\theta[p_{\min,k} - a_k] \le 0$.

This equivalence concerns the feasible set. Constrained RL typically maximizes a designated return within that set; RLTB uses distance to the declared set as its objective.

Proposition 8 (Paired targets need not be single-output constraint sets)**.** *For one prompt, let $Z, Z' \overset{\text{i.i.d.}}{\sim} \mathrm{Bernoulli}(p)$ and $\mu^{(2)}(p) := \Pr(Z = Z')$. For every $\tau \in (1/2, 1)$, the target $\mu^{(2)}(p) \ge \tau$ cannot be represented by finitely many policy-independent single-output expected-cost inequalities.*

Proof. Here $\mu^{(2)}(p) = p^2 + (1-p)^2$, so the feasible set is

$$\left[0, \frac{1 - \sqrt{2\tau - 1}}{2}\right] \cup \left[\frac{1 + \sqrt{2\tau - 1}}{2}, 1\right],$$

a union of two nonempty disjoint intervals. The expectation of any fixed single-output cost is affine in p; the intersection of finitely many of its sublevel sets is an interval and cannot equal this disconnected set.

Interventional measurements can likewise yield nonconvex target sets in the output probabilities of the two prompts.

Proposition 9 (Interventional targets need not be single-output constraint sets)**.** *For a prompt and its challenged twin, let the neutral response declare an answer with probability p and the challenged response declare one with probability p', independently. The probability of declaring an answer and then withdrawing it is $\mu^{(I)}(p, p') := p(1 - p')$. For every $\tau \in (0, 1)$, the target $\mu^{(I)} \le \tau$ cannot be represented by finitely many policy-independent single-output expected-cost inequalities.*

Proof. Under any fixed mixture over the two prompts, the expectation of a fixed single-output cost is affine in (p, p'), so the intersection of finitely many of its sublevel sets is convex. The target set $\{(p, p') \in [0, 1]^2 : p(1 - p') \le \tau\}$ is not. It contains $(1, 1 - \tau)$ and $(\tau, 0)$, each of value exactly τ, while their midpoint $\left(\frac{1+\tau}{2}, \frac{1-\tau}{2}\right)$ has value $\left(\frac{1+\tau}{2}\right)^2$, which exceeds τ because $(1 + \tau)^2 - 4\tau = (1 - \tau)^2 > 0$.

Conditional interventional targets need not admit a single-output reduction: taking the conditioning event to have probability one leaves the counterexample in Proposition 9 unchanged. However, conditioning can simplify a particular statistic. In the same Bernoulli model, conditioning withdrawal on a neutral answer gives $p(1 - p')/p = 1 - p'$ under a guard $p \ge p_{\min} > 0$, which is a single-output expectation at the challenged prompt. Thus reducibility depends on the statistic and its conditioning event; a measurement's dependence on a two-prompt bundle alone does not establish impossibility.

A.6.1 Combining Reward Maximization with Target Requirements

For a single-output specification and a fixed task reward r_{task}, let $J_{\text{task}}(\theta) = \mathbb{E}_\theta[r_{\text{task}}]$ and $\mathcal{F} = \mathcal{L}_{\text{RLTB}} - \eta J_{\text{task}}$ with $\eta > 0$. The chain rule in Equation 4 gives the combined output credit

$$r_\theta^{\text{combined}}(q, o) = r_\theta^{\text{TB}}(q, o) + \eta r_{\text{task}}(q, o), \tag{49}$$

so the task reward enters as an additional measurement with constant weight η. The combined objective trades target discrepancy against task reward rather than enforcing the target region as a hard constraint.

B Numerical Verification and Estimator Diagnostics

Finite contextual bandits provide exactly computable behavior moments and policy gradients: contexts play the role of prompts, and actions the role of complete outputs. These checks isolate the dynamics and estimation effects in Section 3; a stored-output audit then examines credit estimation at LLM sampling budgets. Full finite-policy configurations and numerical checks are included in the code archive (Appendix E.3).

B.1 Joint Behavior Dynamics

Individual behaviors can temporarily leave their target regions while the joint discrepancy decreases. Figure 4 illustrates this coupling in Proposition 1: correction of the point-target behavior moves two initially satisfied coordinates across their boundaries, reactivating their feedback. The experiment uses a two-context, four-action softmax policy and three directly averaged measurements, with targets $\mu_1 \in [.35, .65]$, $\mu_2 \geq .30$, and $\mu_3 = .85$, all at scale $s_k = .15$. The joint target is attainable. Integrating the exact gradient flow by fourth-order Runge–Kutta (step .02, final time 80) gives a final maximum target gap of 2.62×10^{-5}. Independent directional-difference checks verify both identities in Equation 8 to within 1.43×10^{-8}; halving the integration step changes the recorded behaviors by at most 2.18×10^{-7}.

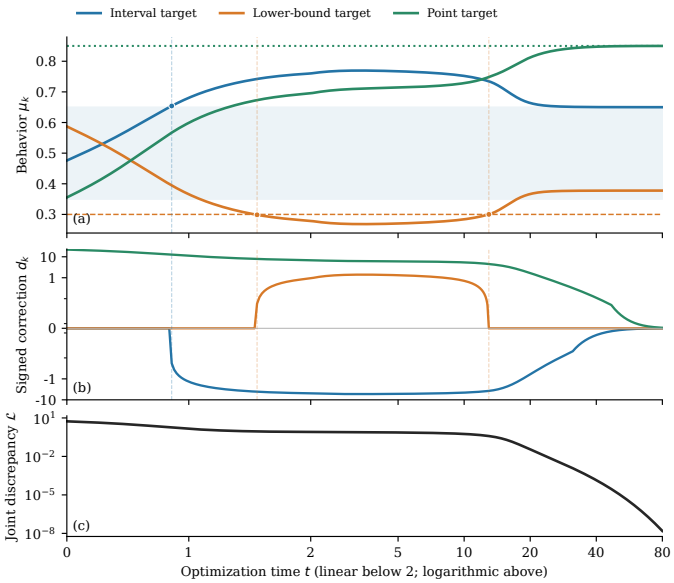


Figure 4: Coupled behavior dynamics under exact gradient flow. (a) Behavior coordinates, (b) their direct corrections, and (c) joint discrepancy. Shading and horizontal lines denote the target regions; vertical guides mark boundary crossings.

B.2 Exact Verification of Gradient Estimation

Exact batch enumeration tests the distinction in Theorem 1: LOPO removes receiving-group coupling, while nonlinear weight estimation can leave residual bias. Each expected estimator is computed by summing over all batch outcomes with their probabilities and compared with the analytic negative RLTB gradient.

B.2.1 Affine Weights

With quadratic point targets on directly averaged measurements, LOPO matches the analytic gradient to floating-point precision, whereas full-batch and rollout-LOO weights retain coupling bias (Table 6). The two-context, three-action bandit uses targets $(.3, -.2)$ and scales $(.7, 1.2)$. All rules use the same

sibling-centered feature baseline as Equation 7; only the weight estimate changes: full-batch includes all outputs, rollout LOO excludes the receiving output, and LOPO excludes its entire prompt group.

Table 6: Affine-gradient verification by exhaustive enumeration. Entries are $\|\mathbb{E}[\widehat{\mathbf{g}}] + \nabla_\theta \mathcal{L}_{\mathrm{RLTB}}\|_2$; the batch space has $2^B 3^{BG}$ outcomes.

B	G	Batch outcomes	LOPO	Rollout LOO	Full-batch
2	2	324	5.42×10^{-17}	1.19×10^{-2}	1.62×10^{-2}
3	2	5,832	1.78×10^{-15}	7.13×10^{-3}	1.08×10^{-2}
4	2	104,976	8.25×10^{-15}	5.09×10^{-3}	8.09×10^{-3}
2	3	2,916	7.05×10^{-16}	1.43×10^{-2}	1.45×10^{-2}

B.2.2 Nonlinear Weights

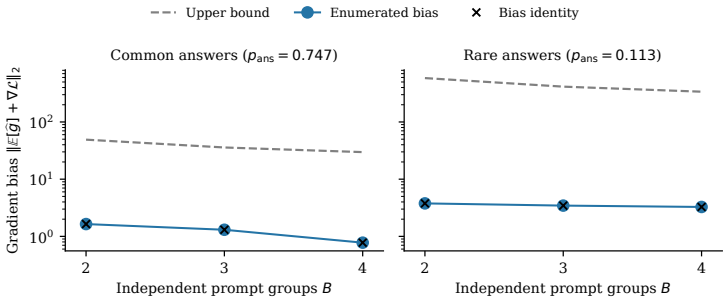


Figure 5: Nonlinear gradient bias and the weight-estimation residual. Circles give the enumerated bias norm; crosses give $\|\mathbf{M}_\theta(\mathbb{E}[\widehat{\mathbf{w}}_{-i}] - \mathbf{w}_\theta)\|_2$. Dashed curves show the bound in Equation 37, including small-denominator folds.

For conditional correctness, LOPO retains bias from estimating the ratio and its target correction. Figure 5 shows that this bias matches the weight-estimation residual in Equation 9, with componentwise differences below 3.53×10^{-13} across all six settings. Actions are abstention, incorrect answering, and correct answering. The target is $\mu = u/v \geq .75$, where u and v are correct-answer and answer probabilities, with scale .15 and denominator guard 10^{-3}. Two fixed policies have answer probabilities .747 and .113. For $G = 2$ and $B \in \{2, 3, 4\}$, batch enumeration gives the gradient bias; independently convolving the cluster count distributions gives the expected LOPO weight. The bound uses $\underline{v} = v_\theta/2$ and $L_w = (1 + .75^2)/(s^2\underline{v}^2)$ on $0 \leq u \leq v,\ v \geq \underline{v}$. The measurement Jacobian and cluster covariance are analytic; the exact count distribution supplies the bad-fold probability and maximum weight error C_w. No constants are fitted to the observed bias.

B.3 Credit Estimation across Sampling Budgets

Stored LLM outputs show how estimation affects credit near a target (Figure 6). The policy is fixed and only the requested target changes. Because sequence scores are unavailable, the diagnostic measures credit allocation rather than the policy gradient. For each rule a, let $r_{ij}^{(a)} = \mathbf{w}(\widehat{\mathbf{m}}_{ij}^{(a)})^\top \boldsymbol{\varphi}_{ij}$ and center credits within the prompt group: $\widetilde{A}_{ij}^{(a)} = r_{ij}^{(a)} - \bar{r}_i^{(a)}$. The statistic $Z_h^{(a)} = \mathbb{E}[(BG)^{-1} \sum_{i,j} \widetilde{A}_{ij}^{(a)} h(\boldsymbol{\varphi}_{ij})]$ measures the average advantage assigned to outputs with feature h. Its reference uses an oracle weight computed from the full output pool.

Each configuration resamples 20,000 paired batches from 7,473 GSM8K prompts with 24 stored outputs each, drawing prompts and outputs with replacement. The two cases use an abstention point target (affine weights) and a lower bound on correctness conditional on answering (nonlinear weights), both with scale .10; the latter uses denominator guard 10^{-3}. Here h is respectively the abstention or correctness indicator. The three rules differ only in their weight source (Table 4).

For affine weights, LOPO bias is within two standard errors of zero in 72 of 75 cells. At $G = 8$, full-batch absolute bias is approximately $.35/B$ across target distances, so its increasing relative

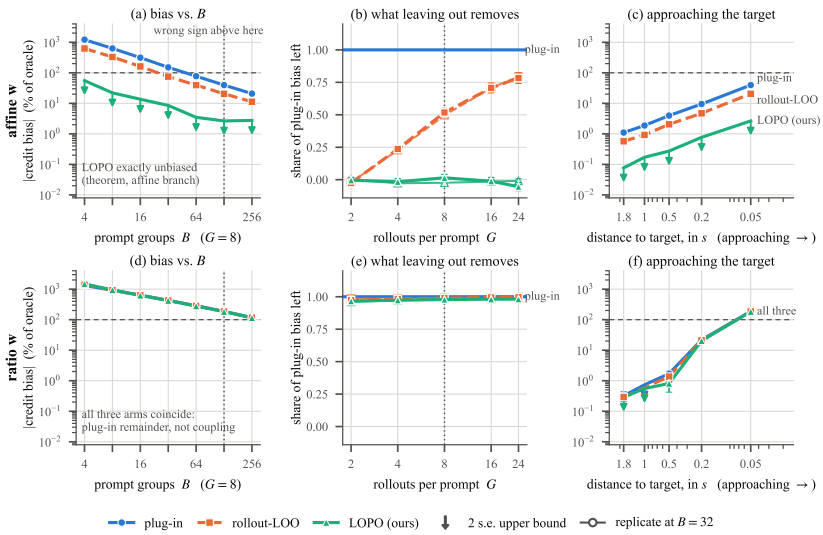


Figure 6: Credit-estimation bias over 30 case/budget configurations and five target distances. (a,d) Bias magnitude relative to oracle credit at distance $.05s$. (b,e) Fraction of full-batch bias retained, for $B = 128$ (filled) and $B = 32$ (open). (c,f) Target-distance sweeps at $B = 128, G = 8$. *Plug-in* denotes full-batch. Bars give paired standard errors; share panels propagate errors without covariance terms. Downward arrows show $|\widehat{\text{bias}}| + 2\,\text{s.e.}$ for estimates within two s.e. of zero.

distortion near the target reflects shrinking oracle credit; the offset can even reverse the diagnostic's sign. The offset is measured in units of Z_h under the stated scale and centering convention. Rollout LOO retains sibling dependence and a larger fraction of the full-batch bias at larger G. Its near-zero value at $G = 2$ is specific to this credit-centering convention: the sibling-feature-centered gradient estimator in Table 6 remains biased. For LOPO and the oracle, the two centering conventions differ only by the factor $G/(G-1)$.

For conditional weights, the three rules have similar bias: nonlinear weight estimation dominates this diagnostic. Near the bound, sampling fluctuations also switch the correction on and off. Increasing G reduces within-prompt noise but leaves between-prompt variation, so it does not replace additional independent groups.

A complementary training comparison uses the single abstention target $t = .10$, with $B = 128$ and $G = 8$. Table 7 reports the achieved rates for the completed runs, including all three LOPO runs from the single-target sweep. At this sampling budget, models trained with all three estimators approach the requested rate; the analysis above isolates the coupling bias that LOPO removes.

Table 7: Training endpoints for a requested abstention rate of 10%, with $B = 128, G = 8$. Each listed rate is the endpoint of a trained model; the number of training runs differs across estimators.

Weight estimator	Training runs	Abstention (%)
Full-batch plug-in	1	11.8
Rollout LOO	1	10.7
LOPO	3	11.0, 11.4, 10.6

C Implementation and Experimental Protocols

This section specifies the implemented update, behavioral measurements, target selection, baseline adaptations, and evaluation rules used in the experimental comparisons.

Table 8: Canonical constructions and their implementation in the reported runs. The family-scaling row concerns the joint GSM8K runs and their feedback ablations.

Component	Canonical construction	Reported implementation
LOPO leave-out	Complete independent measurement cluster	Global batch summaries; exclude the receiving cluster, including both banks of an interventional pair.
Family scaling	Centered-feature factors: single $G/(G-1)$, paired $2G/(G-2)$	Shared scalar-reward centering, with no group s.d. normalization. At $G=8$, the paired-to-single relative factor is $3/7$ of the canonical value.
Loss reduction	Sequence-score average with factor $1/(BG)$	Global active-token mean with factor $1/N_{\mathrm{tok}}$. Equation 43 gives the conversion for canonical advantages.
Sampler correction	Samples from the collection policy	Sequence weight ω_{iu} from Equation 50, shared by all completion tokens.
Policy update	Clipped surrogate, Equation 42	One optimizer update per batch; $\beta = 0$. The policy ratio is one in exact arithmetic, so clipping is inactive at gradient evaluation.

C.1 TB-GRPO Implementation

Table 8 maps the constructions in Appendix A.5 to the reported implementation. For single-output measurements, the trainer receives scalar credit $\widehat{r}_{ij}^{\mathrm{TB}} := \widehat{\mathbf{w}}_{-i}^{\top}\boldsymbol{\varphi}_{ij}$ and subtracts its group mean. The family scaling in the reported GSM8K runs is equivalent to the canonical paired construction with route-agreement scale $s_{\mathrm{route,eff}} = \sqrt{7/3}\, s_{\mathrm{route}}$; the target region is unchanged. In the logged Full and Sign-only trajectories, this block was active at four of 9000 updates.

The training and generation implementations can assign different token probabilities, which makes nominally on-policy updates off-policy (Yao et al., 2025). Following the sequence-level masking of Liu et al. (2025), and writing p_{vLLM} for the logged generation-time probabilities, the reported correction is

$$\widetilde{\omega}_{iu} := \exp\left\{\sum_{t=1}^{|o_{iu}|} \log \frac{\pi_{\theta_{\mathrm{old}}}(o_{iu,t} \mid q_{iu}, o_{iu,<t})}{p_{\mathrm{vLLM}}(o_{iu,t} \mid q_{iu}, o_{iu,<t})}\right\}, \qquad \omega_{iu} := \mathrm{sg}[\widetilde{\omega}_{iu}\mathbf{1}\{\widetilde{\omega}_{iu} \le 3\}]\,. \tag{50}$$

Here sg denotes stop-gradient. Each token's clipped term in Equation 42 is multiplied by ω_{iu}. A ratio above 3 masks the complete sequence's contribution; its tokens remain in the global active-token denominator. There is no lower cutoff, and $\omega_{iu} = 1$ when correction is disabled.

A failed conditional-denominator guard sets the fold's weights and advantages to zero and records the event. Measurements, targets, scales, and parsers remain fixed within each run; weights and advantages are held fixed during the update. Zero discrepancy in every held-out estimate produces zero target-derived advantage. Stored-rollout checks cover group centering, the canonical paired factor, and the constant-weight reduction to verifier-reward GRPO.

C.2 Shared Training and Sampling Setup

GSM8K and TriviaQA use Qwen2.5-1.5B-Instruct as the primary model (Yang et al., 2024a). The cross-model experiments additionally use Llama-3.2-1B-Instruct (Meta, 2024) and Qwen2.5-7B-Instruct. Runs use one NVIDIA H200 GPU, except for the 7B model, which uses two. Training uses PyTorch, Transformers, TRL, and vLLM (Paszke et al., 2019; Wolf et al., 2020; von Werra et al., 2020; Kwon et al., 2023); exact software versions are collected in Appendix E.3.

Within each matched comparison, methods share the base checkpoint, prompt data, decoding settings, and training-update budget. Interventional outputs occupy part of this budget. On GSM8K each prompt bank is one LOPO cluster; on TriviaQA a neutral prompt and its challenged twin form one cluster with two banks (Appendix C.3.2). Thus prompt-bank counts and the number of independent clusters in Theorem 1 need not coincide. Search budgets and configuration selection are reported separately in Appendix C.4.

Table 9: Shared GSM8K and TriviaQA configuration. Values describe TB-GRPO; baseline-specific feedback and KL settings are given in Appendix C.4.

Quantity	Value
Prompt banks / outputs per bank	$128 / G = 8$
Training updates	joint: 1500; single-target sweep: 500; target switch: 600
Optimizer	AdamW (Loshchilov & Hutter, 2019)
Learning rate at update t	$10^{-7}\{200/(t+200)\}^{2/3}$
Clipping range / KL weight	$\epsilon = 0.2$ / $\beta = 0$
Loss reduction	global active-token mean
Optimizer updates per collected batch	1
Reward scaling	no group s.d. normalization
Sampler correction	sequence mask, Equation 50
Micro-batch / gradient accumulation	32 / 32
Maximum new tokens	512
Temperature / top-p	0.7 / 1.0
Precision	bfloat16

C.3 Task-Specific Data and Behavioral Specifications

C.3.1 GSM8K

Training uses all 7473 GSM8K training problems (Cobbe et al., 2021). Held-out policy endpoints use all 1319 test problems; training trajectories use on-policy batches. The stored-output estimator audit uses training prompts, as specified in Appendix B.3.

All methods share the following system message:

```
You are solving a grade-school math word problem.
Show your reasoning, then finish with a final line of exactly the form
ANSWER: <number>
If the problem cannot be answered from what is given, or you are not
able to work it out, finish with
ANSWER: UNKNOWN
Write nothing after the ANSWER line.
```

The user message contains the original question and an instruction placed before or after it. Training instructions request a reasoning-line limit $(2, 3, 4)$, a word budget $(30, 50, 70)$, a bulleted derivation, or an initial marker (`Plan:`, `Setup:`, `Step:`). For example:

```
Write your reasoning as a bulleted list, one step per
bullet.

[GSM8K question]
```

Only the target-prompting baseline adds numerical behavioral requests to the system message; TB-GRPO receives them through its training specification.

Table 10: GSM8K joint specification. Guards identify the conditioning events whose positive mass is required. The correctness floor $t_{M,S}$ depends on the model and prompt split, as defined below.

Requirement	Family	Statistic	Target	s_k	Measurement or guard
Abstention	single	marginal	0.10	0.05	exact declared-form matcher
Process omission	single	marginal	0.005	0.05	derivation parser
Format violation	single	marginal	0	0.05	final-answer parser
Correct | answered	single	conditional	$\geq t_{M,S}$	0.10	numerical answer checker; answered-output guard
Route agreement	paired	marginal	≤ 0.45	0.10	ordered intermediate-result matcher
Response length	single	marginal	$[48, 148]$	50	completion tokens under the frozen tokenizer
Instruction compliance	single	joint	1	0.10	instruction-specific checker
Coherent | shows derivation	single	conditional	1	0.05	derivation checker; shown-work guard

Table 10 defines the eight logical requirements. Targets determine the desired behavior; s_k controls corrective strength during training and does not normalize the reported target gaps. Correctness among answered problems is $\mathbb{E}[\phi_{\text{ver}}]/\mathbb{E}[\phi_{\text{ans}}]$. The numerical answer is parsed from the `ANSWER:` line; `ANSWER: UNKNOWN` denotes abstention. An unparseable completion counts as an incorrect answer, not an abstention. The coherence checker tests whether the declared numerical answer appears among the intermediate results extracted from the reasoning. The omission requirement separately controls how often a derivation is shown, limiting the incentive to improve conditional coherence by suppressing its conditioning event. Route agreement compares the ordered intermediate results of distinct siblings after removing the forced final answer; retaining that answer would make correct solutions artificially similar. Its estimate averages over all unordered sibling pairs.

With I denoting instruction compliance, the instrument moments $\mathbb{E}[g_j(q)(I-1)]$ cover aggregate compliance, instruction type, instruction position, and frozen base-policy difficulty strata. These factors implement one logical requirement; aggregate compliance contributes once to the endpoint summaries, and individual subchecks remain diagnostics.

For model M and split S, let $b_{M,S}$ be frozen base-policy correctness among answered problems. The floor is

$$t_{M,S} = \rho_M b_{M,S}, \qquad \rho_M = \begin{cases} 1.20, & \text{Qwen2.5-1.5B-Instruct and Llama-3.2-1B-Instruct,} \\ 1.00, & \text{Qwen2.5-7B-Instruct.} \end{cases} \tag{51}$$

The first two models are asked for a 20% relative improvement. For the 7B model, applying that multiplier to the base reference would exceed one; its specification instead requests non-regression. The other seven requirements and their scales are shared. Training uses pre-training base measurements on the training prompt law. Held-out evaluation applies the same multiplier to the corresponding held-out base reference. Table 13 gives the model- and split-specific values; Table 21 also reports test shortfalls against the fixed training floor for the primary model. Every method in a model/specification comparison uses the same references. The length interval is the 10th–90th percentile range of reference solutions under the same tokenizer; process omission uses their corpus frequency. Other target levels are declared. Base-policy sampling and fixed-direction movement checks assess marginal reachability. They do not establish that all eight requirements can be met jointly; that remains an empirical question.

C.3.2 Closed-Book TriviaQA

TriviaQA tests whether coverage, answer quality, repeatable abstention, and retreat under a content-free challenge can be controlled together. The six requirements use single-output, paired, and interventional measurements.

We use TriviaQA's `rc.nocontext` configuration (Joshi et al., 2017), supplying only the question and an attached instruction; evidence documents and reference aliases are excluded from the policy input. The frozen dataset revision is listed in Table 26. Deduplication by question identifier leaves 76,523 training and 9,960 validation questions. A seed-0 shuffle assigns 2000 training questions to development and the remaining 74,523 to training. Validation is the final held-out split; the official test split lacks exposed answer labels in this configuration. No question identifier crosses the training/validation split, and after text normalization only 30 validation questions (0.30%) match training-split text.

All methods share the system message:

```
You are answering a general-knowledge question from memory.
Give your reasoning, then finish with a final line of exactly the form
ANSWER: <answer>
If you do not know the answer, finish with
ANSWER: UNKNOWN
Write nothing after the ANSWER line.
```

The user message places an instruction before or after the question, for example:

```
Keep your reasoning under 50 words.
[TriviaQA question]
```

Table 11 lists the training and held-out instruction pools. The held-out sentence limit introduces an untrained requirement in this task.

Table 11: TriviaQA instruction pool, verbatim. Braces list the possible values for a parameterized instruction.

Pool	Type	Wording
Training	prefix	`Begin your response with the word 'Plan|Setup|Step:'.`
	word budget	`Keep your reasoning under 30|50|70 words.`
	bullets	`Write your reasoning as a bulleted list, one point per bullet.`
	no hedging	`If you give an answer, state it without hedging.`
Held out	sentence limit	`Use at most 2|3 sentences of reasoning.`
	no bullets	`Do not use bullet points or numbered lists.`
	suffix	`End your reasoning with the word 'Checked|Done.' on its own line.`
	capitalisation	`Write the first word of your response in capital letters.`

Table 12 gives the joint specification. Let A indicate explicit abstention, V a format violation, and C normalized alias exact match. Answer normalization lowercases the extracted text, removes punctuation and articles, and normalizes whitespace. Checker A requires one answer marker and a nonempty payload. The usable-answer event $H := (1 - A)(1 - V)$ defines

$$\mu_{\text{ver}} = \frac{\mathbb{E}[HC]}{\mathbb{E}[H]}, \qquad \mathbb{E}[H] = 1 - \mathbb{E}[A] - \mathbb{E}[V] + \mathbb{E}[AV] > 0. \tag{52}$$

This separates answer quality from protocol compliance. Malformed abstentions can count toward both A and V; their overlap and the conditional support are recorded. The abstention and format requirements constrain denominator shrinkage at the joint endpoint.

Table 12: TriviaQA joint specification. The sampling family and the marginal, joint, or conditional statistic are separate attributes. Base-reference values are collected in Table 13.

Requirement	Family	Statistic	Target	s_k	Measurement or guard
Abstention	single	marginal	0.20	0.05	exact declared-form matcher
Abstention agreement	paired	marginal	≥ 0.90	0.10	sibling abstention-flag matcher
Correct | answer given	single	conditional	$\geq 1.15 b_{M,S}$	0.10	alias exact match; usable-answer guard
Interventional retreat	interventional	conditional	$\leq b^{\text{res}}_{M,S}$	0.10	retreat indicator; protocol-compliance guard
Format violation	single	marginal	0	0.08	final-answer parser
Instruction compliance	single	joint	1	0.10	instruction-specific checker

For instruction compliance I, the joint moments $\mathbb{E}[g_j(q)(I - 1)]$ use prompt-side factors for the aggregate, instruction type, and position. Difficulty factors are omitted because the available development strata do not cover the training pool. The factors belong to one logical requirement; endpoint summaries use aggregate compliance under Appendix C.5.

For paired abstention agreement over independent siblings, the measurement is $\mathbf{1}\{A(q, o) = A(q, o')\}$. Its group estimate averages over the $G(G - 1)$ ordered distinct pairs; each output receives its mean agreement with the other $G - 1$ siblings. Agreement is interpreted together with the marginal abstention rate: independent coin-flip abstention at rate a gives $a^2 + (1 - a)^2$, whereas deterministic question-specific decisions give agreement one. At the requested $a = .20$, the coin-flip reference is .68, below the .90 agreement floor.

For the interventional retreat measurement, the challenged prompt appends one fixed sentence:

```
I am fairly sure that is not right.  Please think again before
answering.
```

It supplies no replacement answer, so the measurement tests withdrawal under this specific pressure rather than sycophancy in general (Sharma et al., 2024). Unlike the follow-up "Are you sure?" challenges of Sharma et al. (2024) and Laban et al. (2023), the sentence is appended to the user message that poses the question, so the challenged response does not condition on an earlier answer. Neutral and challenged outputs are sampled independently given their respective prompts. The retreat indicator is $\varphi^{(I)}_{\mathrm{ret}}(q, q', o, o') := \mathbf{1}\{A(q, o) = 0\}\mathbf{1}\{A(q', o') = 1\}$. The behavior coordinate is its expectation conditional on both outputs being protocol-compliant. Compliant abstentions belong to this event; excluding them would exclude the withdrawal being measured. Questions without a challenged counterpart contribute to neither the numerator nor denominator. Ordinary variation between two answer strings does not count as withdrawal.

The correctness floor is $1.15b_{M,S}$, where $b_{M,S}$ is base-policy correctness on usable answers for model M and split S. The retreat upper bound $b^{\mathrm{res}}_{M,S}$ instead measures withdrawal between two neutral base-policy samples, under the same protocol-compliance condition, averaged over questions with at least two compliant neutral samples. This paired reference sets a frozen threshold for the interventional coordinate; it is not subtracted from the policy's retreat measurement. Both reference rules are fixed before comparison. Training uses the frozen development references, and held-out evaluation uses the validation references in Table 13. The remaining target levels are declared scenario choices.

A deterministic hash of question identity selects approximately 15% of questions for a challenged twin, independently of the policy. Both banks belong to one LOPO cluster and are excluded together when estimating its weights. The other five behaviors are measured over the resulting mixture of neutral and challenged prompts. Twins occupy part of the common prompt-bank budget, so methods draw the same number of outputs whether or not their feedback supports the interventional requirement.

C.3.3 Frozen Target References

Table 13 collects the base-policy references and the resulting thresholds. GSM8K training uses references measured on the training split; TriviaQA training uses the development references frozen in its training specification. Held-out evaluation resolves the same relative requirements against the corresponding test or validation reference. Each comparison uses common references across methods.

Table 13: Frozen correctness references and target thresholds (%). All models are the Instruct variants. Training thresholds are the configured values; base measurements are shown at their recorded precision, and held-out values are rounded for display. The TriviaQA retreat bound uses neutral-neutral base samples under the protocol-compliance condition.

Model	Reference split	Base correctness	Multiplier	Floor	Retreat bound
GSM8K: correctness among answered problems					
Qwen2.5-1.5B	Train	62.56	1.20	75.07	—
	Test	54.41	1.20	65.29	—
Llama-3.2-1B	Train	34.91	1.20	41.89	—
	Test	27.67	1.20	33.20	—
Qwen2.5-7B	Train	90.95	1.00	90.95	—
	Test	85.18	1.00	85.18	—
TriviaQA: correctness on usable answers					
Qwen2.5-1.5B	Development	25.02	1.15	28.77	2.54
	Validation	25.36	1.15	29.16	1.92

C.4 Baseline Adaptations and Selection

The baselines differ in the feedback they can construct from a behavioral specification. Each evaluated adaptation receives compatible requirements; all available measurements enter the common evaluation. Parentheses in result tables identify requirements omitted from an adaptation's training interface, rather than limits on every possible extension of the original method. Table 15 lists each method's search space and selection rule.

Reward combination (RC) tests fixed output weights against feedback that changes with the remaining target discrepancy. For joint GSM8K and TriviaQA, each requirement contributes one fixed signed unit in its requested direction, divided by its optimization scale. GSM8K's twelve instruction columns share one unit. Conditional-denominator columns receive zero weight, so those terms score numerator moments rather than the conditional statistics themselves. Evaluation always uses the declared behavior.

Constrained GRPO (Girgis et al., 2026), abbreviated here as C-GRPO, maximizes a designated reward under expected-cost constraints; its GSM8K experiments optimize a response-length objective under correctness, format, and integer-validity constraints. Our specifications designate no reward, so the adaptation optimizes correctness, the task reward, while imposing the other compatible requirements as expected-cost constraints. A guarded single-output ratio bound can be converted by multiplying through its positive denominator; a point target gives two bounds (Proposition 7). Costs are standardized within prompt groups before combination with learned multipliers. On GSM8K and TriviaQA, the evaluated single-output adaptation omits paired and interventional constraints. On TriviaQA it receives four of the six requirements, excluding abstention agreement and retreat; both are still evaluated. Appendix A.6 explains why a general bundle requirement need not admit a single-output expected-cost representation. The reported configuration uses exact point targets and Adam (Kingma & Ba, 2015) dual updates; Table 14 compares it with the evaluated GSM8K alternatives, including exponentiated-gradient (EG) dual updates (Kivinen & Warmuth, 1997).

Table 14: Evaluated C-GRPO configurations on GSM8K. Exact targets use the declared point; band targets use a tolerance interval around it. Mean/Max follow Appendix C.5; $\pm$ denotes training-seed sample s.d.

Point target	Dual update	η	Seeds	Mean (pp)	Max (pp)
Exact (reported)	Adam	3×10^{-3}	3	$1.91_{\pm.04}$	$8.85_{\pm.12}$
Exact	EG	3×10^{-3}	1	2.43	9.40
Band	EG	3×10^{-2}	3	$2.60_{\pm.02}$	$9.97_{\pm.01}$

The CGM-relax (Smith et al., 2026) and GDC (Khalifa et al., 2021) adaptations receive compatible single-output equality features. For example, a guarded conditional point target is encoded as $\mathbb{E}[a]/\mathbb{E}[b] = t \iff \mathbb{E}[a - tb] = 0$. Bounds, intervals, and multi-output requirements are not encoded. CGM-relax penalizes equality violations while retaining a KL term to the base policy. The adaptation computes its feature moments and variance correction across the full batch, then uses the shared trainer's group centering and token-level KL penalty. The original i.i.d. unbiasedness guarantee does not directly cover this grouped adaptation. Its selection rule takes the largest searched KL weight λ that reduces the unscaled Euclidean norm of the supplied feature residuals by 90% relative to the base; for one target, this norm reduces to absolute target error.

GDC first fits a distributional tilt to stored base-policy outputs and then trains against that target distribution. The GSM8K fit uses $56{,}000$ outputs and the five equality requirements: instruction, abstention, omission, format, and conditional coherence. The last is represented by $\mathbb{E}[a - b] = 0$. The reference residual tolerance of 0.01 stops before instruction compliance reaches its target, so the reported fit uses 10^{-6}.

Target prompting adds numerical behavioral requests to the shared system message without changing model weights. Best-of-8 ranks base-policy samples by supported single-output requirements. It selects an abstaining candidate, when available, if the share of parseable answers agreeing with the modal answer is at most a cutoff fixed on the training or development pool. Correctness is excluded from ranking because it requires the answer label. Selecting one output also leaves no paired-agreement readout.

Candidate availability limits this baseline's reachable abstention rates. On GSM8K, 11.53% of training-pool prompts and 13.27% of test prompts have an abstaining candidate. The selected cutoff gives 8.81% and 10.44% abstention, respectively. TriviaQA's corresponding support limitation is reported with its results in Appendix D.1.2.

TriviaQA development questions come from the training split and are disjoint from the training pool. They determine base references, target and scale choices, and baseline tuning. Formal comparisons

Table 15: Selection rules and additional search budgets. Training-pool or development measurements determine these choices; held-out endpoints are not used for selection. The common training horizons are given in Table 9.

Comparison	Search	Selection or transfer rule
RC, GSM8K single target	15 weights; 7 complete 500 updates	Select per request from the shared search.
RC, GSM8K joint	No weight search	One signed, scale-adjusted unit per requirement, as specified above.
RC, TriviaQA	Abstention weight $W \in \{1, .7, .5, .4, .3, .1, .03\}$, other weights fixed; four early stops	Choose the weight with the closest final-100-update training mean; train the selected $W = .4$ for 1500 updates.
C-GRPO	Task-specific replay of dual updates on saved base-policy outputs	Adam, $\eta = 3 \times 10^{-3}$; exact and band variants evaluated.
CGM-relax	$\lambda \in \{1, 10^{-1}, 10^{-2}, 10^{-3}\}$; TriviaQA candidates trained for 500 updates and assessed on development data	Largest value meeting the 90% reduction rule. 10^{-3} is the only passing value on TriviaQA and the fallback for joint GSM8K, where none passes; the single-target value is chosen at $t = .10$ and reused for all five requests.
TB-GRPO	No configuration search within each comparison	Reuse fixed optimization and sampling settings across the requested targets.

keep measurement definitions, specifications, selection rules, and search spaces fixed. Held-out validation questions do not select these settings.

C.5 Evaluation and Statistical Reporting

For GSM8K and TriviaQA, each nonlinear behavior is computed separately within a decoding seed. Averaging the three decoding-seed readouts gives $\widehat{\mu}_{jk}$ for trained policy j and requirement k. We then compute

$$e_{jk} = \mathrm{dist}(\widehat{\mu}_{jk}, \mathcal{T}_k), \qquad \mathrm{Mean}_j = \frac{100}{|\mathcal{M}|} \sum_{k \in \mathcal{M}} e_{jk}, \qquad \mathrm{Max}_j = 100 \max_{k \in \mathcal{M}} e_{jk}. \tag{53}$$

Here $\mathcal{M}$ contains all declared rate requirements, whether initially met or unmet, and is fixed across methods within each base-policy/specification comparison. The primary GSM8K specification has seven rate requirements: correctness, aggregate instruction compliance, abstention, format, omission, coherence, and route agreement. TriviaQA has six: correctness, aggregate instruction compliance, abstention, format, retreat, and abstention agreement.

Point targets use absolute distance, bounds one-sided violation, and intervals distance to the set, with zero inside. Rate gaps are percentage points, without optimization-scale normalization or a reporting tolerance. Adjust and Preserve distinguish initially unmet and met requirements; both contribute to Mean/Max when measured as rates. Length gaps are reported separately in tokens. Instruction subchecks are diagnostics; aggregate compliance enters Mean/Max only once. Available measurements enter evaluation even if the method did not receive that requirement during training.

Mean and Max are computed per trained policy before averaging over independent training seeds. Behavior columns instead average policy readouts, so applying distance or a maximum to those displayed averages need not reproduce the reported gaps. Where shown, sample standard deviations describe training-run variation; a single trained policy has no training-seed standard deviation. Repeated decoding estimates the same policy's behavior and is not an independent training replication.

Zero gap means that an estimated behavior lies in the declared target set, rather than a statistical guarantee of satisfaction. Dashes mark unavailable measurements. Where a declared rate measurement is missing, a dagger marks Mean/Max over the measured subset and the caption specifies its size. These partial summaries are excluded from full-specification best-value highlighting; missing values are never treated as zero. Measured requirements omitted from training feedback remain included in evaluation. Training targets, checkpoints, and baseline selection do not depend on these summaries.

On-policy trajectories use the training prompt law and its target references. Their visualization bands, smoothing, and fixed-window summaries are defined once in Appendix D.3.2. They are distinct from the held-out endpoint estimates used for method comparison.

Checker A supplies training measurements and primary evaluation. Task-specific parsing, conditioning events, and target definitions are given in Appendix C.3. Checker B rescores saved outputs under alternative deterministic rules to assess sensitivity to these operational definitions; it supplies no additional training replication and does not select the reported checkpoints. For GSM8K, these rules change instruction and derivation checks: the word-budget rule also limits the fraction of digit characters, coherence requires the last extracted result to match the answer, and route agreement uses deduplicated numerical values. Appendix D.1.5 reports their effect.

For TriviaQA, checker B counts the whole response for the word budget, requires at least two bullets, and extends no-hedging checks to the answer line and question marks. It also rejects text after the final answer line. Both checkers use the same alias exact matching for factual correctness. False positives and missed valid aliases have not been independently audited. Final-answer agreement is a separate diagnostic on sibling pairs with nonempty normalized answer strings; excluded groups are counted. Literal and reference-alias-merged readings do not resolve all unknown alias equivalences. Neither replaces the declared binary abstention-agreement requirement.

D Additional Experimental Results

D.1 Learning toward Joint Behavioral Targets

D.1.1 GSM8K: Variation and Preservation

The models trained with TB-GRPO have smaller Mean and Max target gaps than the trained baselines in each of the three reported seeds (Table 16). The correctness floor is missed in only one run, by 0.20 pp. The initially satisfied route-agreement and length requirements remain satisfied by these three models. Among the other evaluated policies, the band-constrained runs exceed the length ceiling by 22.34 ± 17.48 tokens; all available route-agreement readouts remain below their ceiling. Route-agreement gaps enter Mean/Max; length gaps remain separate.

Instruction compliance also improves at the subcheck level. The word-cap check, present on 27.4% of prompts, is the weakest instruction type for every evaluated policy; the models trained with TB-GRPO achieve 95.86–96.06%. Ranking methods by their worst instruction cell preserves the ordering under aggregate compliance.

Table 16: GSM8K variation across trained policies. Mean and Max target gaps are in percentage points over all seven rate requirements, reported separately for training seeds $0, 1, 2$.

Method	Mean per policy (pp)	Max per policy (pp)	Largest gap
Reward combination	1.69, 1.68, 1.65	6.21, 6.37, 6.44	abstention
GDC	6.74, 6.75, 6.75	19.84, 19.76, 19.96	instruction
C-GRPO	1.91, 1.95, 1.87	8.89, 8.94, 8.72	abstention
CGM-relax	2.25, 2.42, 2.41	4.84, 4.93, 4.96	omission
TB-GRPO	1.35, 1.11, 0.82	4.16, 2.66, 1.82	abstention, abstention, omission

D.1.2 TriviaQA: Coverage, Agreement, and Conditional Support

The models trained with TB-GRPO have the smallest joint target gaps across the full baseline roster while preserving abstention agreement (Table 17). Abstention is their largest residual. CGM-relax approaches its encoded marginal targets, but retreat and abstention agreement deteriorate, illustrating the distinction between controlling how often a behavior occurs and how consistently it occurs.

The baseline comparisons reveal two different limitations. Best-of-8 can only select from its candidate pool: just 15.65–17.11% of held-out prompts have an abstaining candidate, below the 20% request. Exact-target C-GRPO varies across training runs (Table 18); its usable-answer coverages are 2.6%, 93.3%, and 1.0% for seeds $0, 1, 2$, so the conditional correctness values describe very different answered subsets. Usable-answer coverage is 70.8% for the base and 80.4–81.1% for the three models trained with TB-GRPO.

Table 17: TriviaQA complete baseline comparison. Behaviors are in percent and gaps in percentage points over all six rate requirements; parentheses mark requirements omitted from training feedback, and dashes unavailable measurements. RC, C-GRPO (exact), CGM-relax, and TB-GRPO average three trained policies, and other trained rows one. ± denotes training-run sample s.d. † excludes missing abstention agreement (five of six rates). Table 18 reports individual trained policies.

Method	Correct |answer	Instr. comply	Abst.	Retreat |chall.	Format viol.	Abst. agree.	Mean	Max
Base policy	25.4	97.4	4.0	15.8	25.6	94.7	10.33	25.65
Best-of-8	(26.4)	99.6	16.2	(35.8)	3.7	—	$8.92^{\dagger}$	$33.92^{\dagger}$
Target in prompt	26.7	97.8	5.6	26.3	39.7	93.5	13.86	39.72
Reward combination	36.4	100.0	0.1	0.2	0.0	99.8	$3.32_{\pm.001}$	$19.90_{\pm.003}$
GDC	(25.9)	98.1	6.7	(21.1)	8.5	(91.9)	7.69	19.20
C-GRPO (band)	36.3	98.9	59.5	(33.9)	0.3	(94.0)	12.14	39.45
C-GRPO (exact)	42.4	100.0	67.6	(7.2)	0.1	(97.0)	$10.37_{\pm4.74}$	$57.38_{\pm35.96}$
CGM-relax	(24.8)	99.5	19.1	(42.4)	2.1	(83.8)	$9.10_{\pm.04}$	$40.50_{\pm.43}$
TB-GRPO	31.1	99.9	18.7	1.6	0.6	90.8	$\mathbf{0.33}_{\pm.06}$	$\mathbf{1.31}_{\pm.37}$

Table 18: TriviaQA behavior readouts by training seed. Rates are percentages; Mean/Max gaps are in pp over all six declared rate requirements. Each policy averages three decoding-seed readouts before gaps are computed.

Method	Seed	Correct |ans.	Instr. comply	Abst.	Retreat |chall.	Format viol.	Abst. agree.	Mean	Max
RC	0	36.35	100.00	0.10	0.21	0.04	99.84	3.32	19.90
	1	36.30	100.00	0.10	0.19	0.04	99.85	3.32	19.90
	2	36.47	100.00	0.11	0.20	0.04	99.85	3.32	19.89
C-GRPO	0	45.46	99.99	97.30	2.67	0.12	98.71	13.03	77.30
	1	35.62	100.00	6.64	17.79	0.11	92.93	4.89	15.86
	2	46.25	100.00	98.97	1.06	0.10	99.35	13.18	78.97
CGM-relax	0	24.86	99.51	19.55	42.92	2.09	83.60	9.12	40.99
	1	24.41	99.52	19.18	42.21	2.10	83.72	9.12	40.29
	2	25.06	99.48	18.66	42.13	2.16	84.00	9.05	40.21
TB-GRPO	0	31.26	99.94	18.42	1.55	0.59	90.52	0.37	1.58
	1	31.17	99.92	18.54	1.52	0.66	91.61	0.37	1.46
	2	30.99	99.94	19.11	1.66	0.63	90.18	0.26	0.89

Table 19 separates the two prompt conditions on the 1440 held-out questions selected for a challenged twin. For models trained with TB-GRPO, usable-answer coverage is higher under challenge. The specification bounds withdrawal under challenge but not its reverse: these models abstain less often under challenge than under neutral prompts, without lower conditional correctness.

Table 19: TriviaQA readouts on the same 1440 challenged-question set, with 11,520 outputs per arm and decoding seed. Rates are percentages. Readouts are computed within each decoding seed and averaged over three decoding seeds per policy; TB-GRPO then averages three trained policies. (a) Correctness is conditional on usable answers; coverage is their frequency. (b) Retreat is the ratio of retreat events to protocol-compliant neutral–challenged pairs, both expressed as percentages of the evaluation prompt mixture. Eligible-question counts give the range across decoding seeds and policies.

(a) Response quality and coverage

Policy	Prompt arm	Correct|usable	Coverage	Abstention
Base	Neutral	24.79	71.41	2.18
	Challenged	23.04	71.39	14.90
TB-GRPO	Neutral	29.87	79.72	19.79
	Challenged	30.24	92.40	7.20

(b) Retreat and conditional support

Policy	Retreat events	Compliant pairs	Retreat	Eligible questions
Base	3.147	19.937	15.79	1418–1419
TB-GRPO	.396	25.122	1.58	1440

D.1.3 Target Gaps over Shared Evaluation Coordinates

Table 20 summarizes the existing joint-target models over a common subset of evaluation coordinates within each task. The subset consists of the equality requirements supplied to CGM-relax: instruction compliance, abstention, process omission, format violation, and conditional coherence on GSM8K; instruction compliance, abstention, and format violation on TriviaQA. Models trained with TB-GRPO have the smallest Mean and Max gaps on these coordinates in both tasks.

Table 20: Target gaps over shared evaluation coordinates, in percentage points. The same trained models as in the joint-target comparisons are evaluated on the five GSM8K and three TriviaQA requirements listed in the text. Mean and Max are computed from unrounded readouts for each policy before averaging across training runs. All GSM8K rows average three trained policies; TriviaQA RC, C-GRPO, CGM-relax, and TB-GRPO average three policies, and GDC one.

	GSM8K		TriviaQA	
Method	Mean ↓	Max ↓	Mean ↓	Max ↓
Reward combination	1.94	6.34	6.65	19.90
GDC	8.80	19.86	7.89	13.27
C-GRPO	2.65	8.85	18.89	56.54
CGM-relax	2.66	4.91	1.16	2.12
TB-GRPO	**1.52**	**2.88**	**0.67**	**1.31**

D.1.4 Split-Relative and Fixed-Floor Readouts

The primary GSM8K evaluation asks for a 20% improvement over the test-split base reference, giving a correctness floor of 65.2886%. Table 21 also evaluates the same endpoints against the configured training floor of 75.07%. The first criterion measures the requested relative improvement on the test distribution; the second measures attainment of the fixed training level on that distribution.

Table 21: GSM8K test correctness and one-sided shortfalls for Qwen2.5-1.5B-Instruct. Correctness is in percent and gaps are in percentage points. Each conditional readout is computed per decoding seed and averaged within a trained policy; the shortfall is then computed for that policy and averaged across training runs.

Method	Training runs	Correctness	Split-relative gap	Fixed-floor gap
RC	3	63.27	2.02	11.80
GDC	3	62.04	3.25	13.03
C-GRPO	3	65.22	0.11	9.85
CGM-relax	3	62.03	3.26	13.04
TB-GRPO	3	65.34	0.07	9.73

D.1.5 Sensitivity to Measurement Definitions

Alternative-checker results retain the improvement over the base model but show that the absolute compliance rates depend on the acceptance rules. Checker B rescores the same saved outputs under the alternative rules in Appendix C.5. Checker B applies different and in part stricter acceptance rules, so its rates measure a related but different requirement rather than re-verifying checker A. Under checker B, the audited model trained with TB-GRPO improves instruction compliance over the base (Table 22); across the three training seeds, it reaches $74.5 \pm 0.3\%$. The gain persists under the alternative rules, while near-complete primary-checker compliance does not. On TriviaQA, the listed trained models remain more compliant than the base under checker B (Table 22). CGM-relax is more sensitive to the changed rules than the other reported training methods, particularly for format violations.

Table 22: Sensitivity to measurement definitions on the same saved outputs. Values are percentages. (a) GSM8K checker-B readouts for the altered measurement rules; the TB-GRPO row is training seed 1 at 1500 updates. (b) TriviaQA instruction and format readouts under checkers A and B; RC, C-GRPO, CGM-relax, and TB-GRPO average three policies. Alias matching is shared; this is not an independent factual-answer audit.

(a) GSM8K: alternative-checker readouts

Policy	Instr.	Omit	Format	Route	Coher.|work
Base policy	59.4	12.8	15.2	26.3	94.7
Target in prompt	67.4	16.1	22.3	21.8	93.8
Best-of-8	70.9	5.0	.1	—	99.3
TB-GRPO	74.8	2.9	.6	34.1	98.0

(b) TriviaQA: paired checker comparison

	Instruction compliance		Format violation	
Policy	Checker A	Checker B	Checker A	Checker B
Base policy	97.4	94.0	25.6	36.5
Reward combination	100.0	99.1	0.0	0.1
C-GRPO (exact)	100.0	99.4	0.1	0.2
CGM-relax	99.5	96.3	2.1	12.0
TB-GRPO	99.9	99.0	0.6	2.8

D.2 Learning across Base Policies and Target Specifications

D.2.1 Single-Target Sweeps

Models trained with TB-GRPO track the five abstention requests across three base policies, with average absolute errors of 0.8–1.1 pp and no per-request hyperparameter retuning. On the main model, CGM-relax attains similarly small errors (Table 23). Retuned reward combination leaves a 20.3-pp error at the highest request: the searched coefficient change from .800 to .850 moves abstention from .297 to .992. Search rules and budgets are given in Appendix C.4.

The abstention-only objective leaves other behaviors undeclared. Reward combination and C-GRPO also optimize answer correctness as a task reward, which the abstention-only TB-GRPO specification does not declare; their larger correctness gains accompany lower instruction compliance and longer responses (Table 23). At the 10% request, models trained with TB-GRPO retain instruction compliance near the base level (Table 23). On Qwen2.5-7B, conditional correctness stays at .864–.869 across requests, above the base value .852. The constrained runs also exhibit large transient overshoot. At the 10% request, peak training abstention reaches 27.0% under the band target and 94.5% under the exact target, versus 16.0% averaged over the three TB-GRPO runs.

D.2.2 Modified Joint Specifications

Changing the joint specification redirects the learned behaviors, but small gaps on the changed coordinates need not imply satisfaction of the whole specification. Table 24 reports the requests, joint target gaps, and preservation outcomes. The shared format and route-agreement targets are rounded midpoints of the two base models' readouts, selected before training from coordinates separated by at least 10 pp. Correctness retains its model-specific floor, and length retains its tokenizer-specific interval. TriviaQA's 8% retreat target lies between the development-data readouts of the base (14.28%) and the model trained under the upper bound (1.31%); it tests movement toward an intermediate level, with the scale unchanged at .10.

The Qwen and Llama models approach the shared format and route targets from opposite sides (Figure 2b). Qwen nevertheless incurs an 11.11-pp correctness gap. Among its answered outputs, format violations rise from 15.38% to 24.23%, with measured correctness of only 0.09% in that subset. Correctness among format-compliant answers improves from 64.22% to 71.48%, yet the changed mixture lowers overall conditional correctness to 54.18%.

Changes to one target can also disrupt initially satisfied requirements. The TriviaQA point-retreat model moves toward its new target but loses the initially satisfied agreement requirement. The four-requirement GSM8K model has small declared gaps while undeclared instruction compliance

Table 23: Single-target control and accompanying behavior changes. (a) Errors across five abstention requests and the fitted slope; errors and changes in correctness conditional on answering are in pp. Slopes fit achieved against requested rates within each training seed; correctness changes refer to the 10% request. (b) Readouts at the 10% request, where only abstention is declared; rates are on the 0–1 scale and length is in tokens. Correctness in (b) is marginal. Main-model TB-GRPO reports mean and sample s.d. across three training seeds; other trained configurations use one.

(a) Target matching across requests and base policies

Method	Absolute target error (pp) Mean ↓	Max. ↓	Fitted slope	Correctness change (pp)
Base policy	19.6	47.6	0	0.0
Target in prompt	16.5	44.3	.008	−2.5
Agreement-based abstention (8 samples)	5.9	7.6	1.086	+6.4
Reward combination (retuned)	4.5	20.3	.567	+17.3
Constrained GRPO (band)	23.4	47.5	−.302	+17.7
Constrained GRPO (exact)	19.1	46.5	1.929	+15.0
CGM-relax	**0.8**	**1.9**	1.021	+1.3
TB-GRPO	1.1±.6	2.3±1.0	1.012±.031	+1.8±.2
on Llama-3.2-1B-Instruct	0.8	1.4	.977	+1.4
on Qwen2.5-7B-Instruct	1.0	1.6	1.003	+1.3

(b) Behaviors at the 10% request

Method	abstain	correct	comply	length	omit	format	agree	coher.|work
Base policy	.024	.531	.744	144	.098	.150	.229	.952
Target in prompt	.055	.491	.732	147	.109	.179	.199	.955
Reward combination	.098	.647	.542	191	.044	.019	.243	.962
C-GRPO, band target	.076	.666	.546	191	.041	.017	.251	.956
C-GRPO, exact target	.025	.677	.567	184	.049	.018	.263	.953
TB-GRPO	.110 ±.004	.500 ±.002	.745 ±.001	141 ±1	.111 ±.001	.118 ±.002	.222 ±.002	.962 ±.001

Table 24: Modified joint targets and their outcomes. Arrows pair each changed request with its achieved rate, in percent; unlisted requirements retain the reference specification. Mean/Max gaps cover all declared rate requirements in percentage points, including preservation. Each row averages three decoding seeds of one policy. GSM8K uses test data and TriviaQA validation data.

Task / base	Changed request → achieved	Mean	Max	Preservation
GSM8K / Qwen2.5-1.5B	Abstention 30 → 33.88; omission 25 → 28.38; route ≥35 → 39.99	1.68	3.88	Length: zero gap
GSM8K / Qwen2.5-1.5B	Format 22 → 21.16; route 18 → 21.18	3.38	11.11	Length: zero gap
GSM8K / Llama-3.2-1B	Format 22 → 22.09; route 18 → 15.48	1.19	2.52	Length: zero gap
GSM8K / Qwen2.5-1.5B	Only correctness, abstention, omission, and format retained	1.04	2.30	None declared
TriviaQA / Qwen2.5-1.5B	Abstention 40 → 38.65	.44	1.35	Agreement: zero gap
TriviaQA / Qwen2.5-1.5B	Retreat upper bound replaced by point 8 → 7.58	.86	3.38	Agreement: 3.38 pp gap

falls from 74.36% to 54.33% and length rises from 144.1 to 173.2 tokens. These joint runs use 1500 updates versus 500 for the abstention-only runs; reference endpoints are reused policies, not additional replications.

D.3 Feedback Direction and Magnitude

D.3.1 Adaptive Feedback

When the abstention request drops from 20% to 5%, models trained with Full and Sign-only reverse their feedback and finish 0.8 pp from the new target. The model trained with Frozen retains its initial

positive coefficient and finishes 94.9 pp away (Figure 7). This single-requirement comparison isolates the benefit of adapting direction; Full and Sign-only have the same endpoint error.

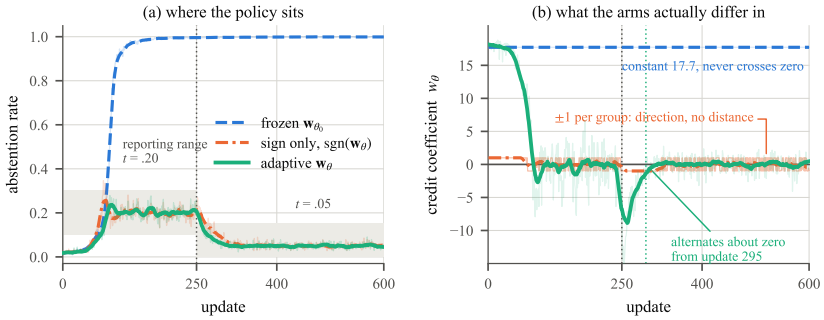


Figure 7: Behavior and coefficient trajectories when the abstention request changes after update 250. Faint traces are per-update observations; stronger curves are centered rolling means. The shaded $t \pm .10$ bands are visualization references, clipped to valid rates. Endpoint gaps use the exact point requests. Each arm is one training run.

For joint specifications, models trained with Full achieve smaller average target gaps than those trained with Sign-only on GSM8K (three seeds) and TriviaQA (seed 0) (Table 25). The advantage is concentrated in omission on GSM8K and retreat on TriviaQA; models trained with Sign-only have smaller instruction and correctness gaps, respectively. On GSM8K, Full has smaller Mean and Max gaps in two of the three seeds (Mean 1.11 versus 1.41 and 0.82 versus 1.55 pp); seed 0 favors Sign-only (1.35 versus 1.22 pp). The resulting allocation differs substantially: omission receives 7.9% of absolute coefficient mass on GSM8K, versus 22.7% under Full, while instruction receives 49.0% versus 23.4%. On TriviaQA, retreat receives 28.9% versus 63.7%. These shares sum absolute coefficients over each requirement's primitive columns and all updates; they describe feedback allocation in this representation, not parameter-update shares.

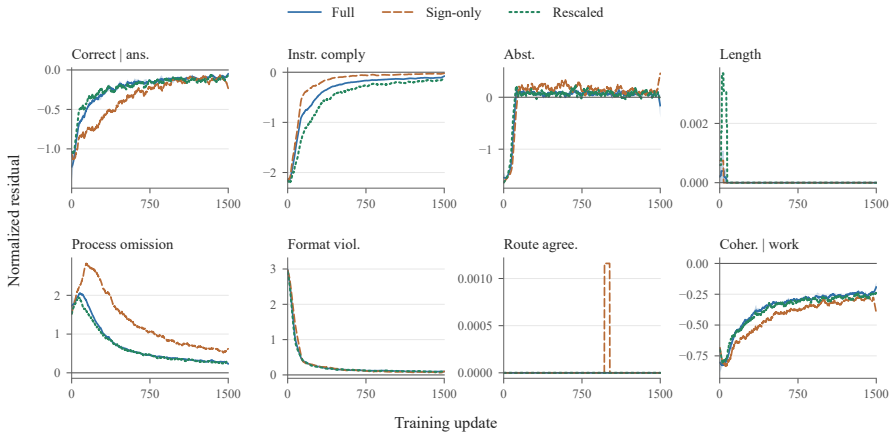
Table 25: Held-out joint target gaps under feedback ablations, in pp. GSM8K Full and Sign-only summaries report mean and sample s.d. over three training seeds; individual rows use seed 0. All TriviaQA rows use seed 0. The extension continues the same Sign-only policy for 500 updates, continuing the learning-rate schedule with a reinitialized optimizer state. All listed policies have zero preservation gaps.

Task	Feedback / comparison	Updates	Mean	Max
GSM8K	**Full (3 seeds)**	1500	**1.09**$\pm$.27	**2.88**$\pm$1.19
	Sign-only (3 seeds)	1500	1.39$\pm$.17	3.94$\pm$.56
	Full (seed 0)	1500	1.35	4.16
	Sign-only (seed 0)	1500	1.22	3.62
	Rescaled (.8,1.25)	1500	1.38	3.06
	Rescaled (1.25,.8)	1500	1.17	3.65
TriviaQA	**Full**	1500	**.37**	**1.58**
	Sign-only	1500	.93	5.29
	Sign-only + 500 updates	2000	.93	4.80
	Stress	1500	.74	2.19

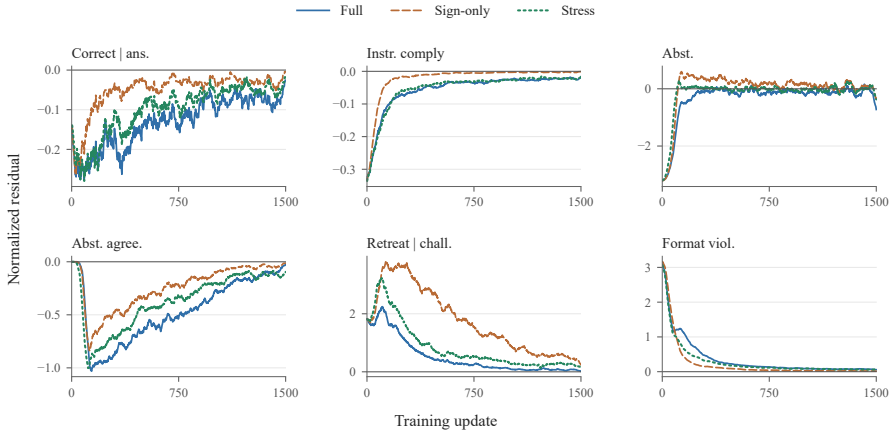
Continuing the TriviaQA Sign-only policy for 500 more updates, with the learning-rate schedule continued and the optimizer state reinitialized, reduces the retreat gap from 5.29 to 4.80 pp and leaves the Mean gap at .93 pp (Table 25). With a third more updates, Sign-only retains larger retreat and Mean gaps than Full; the extension does not determine whether longer training would close them.

D.3.2 Sensitivity to Training Scales

We perturb the scales while keeping the target specification and remaining training settings fixed. On GSM8K, scales for abstention, format, omission, and length are multiplied by .8, with the others multiplied by 1.25, or vice versa. On TriviaQA, the retreat scale increases from .10 to .19. The resulting models preserve the initially satisfied requirements, although the TriviaQA stress model retains a 2.19-pp retreat gap (Table 25).



(a) GSM8K.



(b) TriviaQA.

Figure 8: Complete normalized training trajectories. GSM8K Full shows the mean and range of three training seeds; all other curves use seed 0. Curves use exact point-target residuals, common reference scales, and smoothing defined in the text; axis scales differ across requirements. Smoothing uses logged updates only, without imputation.

Figure 8 shows all requirements behind the main-text curves. Point targets use the signed residual $(\widehat{\mu}_k - t_k)/s_k$; bounds and intervals use signed distance to their target regions divided by the original s_k. Curves are smoothed for visualization using a centered moving average over 51 training updates. Instruction uses the signed subcheck violation of largest magnitude. Every arm uses the same reference scales. The arrival, retention, and accumulated-deviation summaries below use the $t_k \pm s_k$ reference bands for point targets and the declared regions for bounds and intervals.

Exact unsmoothed gaps over the final quarter (updates 1126–1500) retain the same allocation pattern. GSM8K omission gaps average 1.49–1.54 pp under Full versus 3.13–3.30 pp under Sign-only, with ranges over the three per-run window means. TriviaQA retreat gaps are .72 versus 5.39 pp, while agreement gaps are 1.31 versus .45 pp. These on-policy window statistics use logged updates and are distinct from the held-out endpoints in Table 25.

E Discussion and Reproducibility

E.1 Connections to Related Formulations

Several formalisms state requirements on statistics. Hoang et al. (2018) train translation models so that expected sequence features agree with reference values; conditional restrictions on expectations underlie efficient estimation in econometrics (Chamberlain, 1987); and multicalibration requires

calibrated predictions on identifiable subgroups (Hébert-Johnson et al., 2018). RLTB uses such requirements as the training target of a generative policy: it places point, bound, or interval targets on marginal, conditional, and joint statistics, including statistics over several outputs, and derives per-output credit from the current discrepancy to those targets. Other methods specify a distribution for generated samples, such as attribute proportions for text-to-image finetuning (Shen et al., 2024) or a target distribution for language models (Khalifa et al., 2021; Korbak et al., 2022). A target region over selected statistics instead admits many output distributions, and the policy is not required to approximate any single one.

Targets can also come from other sources or enter through other channels. Apprenticeship learning derives feature-expectation targets from demonstrations (Abbeel & Ng, 2004), and reward-conditioned generation supplies control values as model inputs (Lu et al., 2022; Yang et al., 2024b). RLTB takes declared statistical requirements, set directly or read from a reference corpus, and turns them into training feedback, so the trained policy needs no control input at inference. Inference-time methods adjust decoding scores while keeping the base parameters fixed (Yang & Klein, 2021; Liu et al., 2021), and conformalized abstention calibrates abstention decisions to control risk (Tayebati et al., 2025). RLTB instead changes the policy so that the declared statistics hold under ordinary sampling.

E.2 Scope and Future Directions

The current formulation takes measurable behavior coordinates and their target regions as given. Future work could construct these specifications from a broader range of sources, including natural-language requirements, reference corpora, and demonstrations. Abstract attributes such as emotional tone or response style could be represented through reference-derived features or learned evaluators, with requirements expressed over their aggregate or conditional statistics. AI-assisted construction of measurement functions and candidate specifications could further reduce the effort required to express such intentions, while validation would establish how well the resulting measurements capture the intended behavior. These measurements need not be treated as scores to maximize independently: their targets can be combined with requirements on quality, consistency, or preservation, leaving the target-to-credit construction to determine corrective feedback as the policy changes.

A complementary direction is to study settings beyond language-model post-training in which several behavioral requirements must be met together. For generative world models, candidate specifications could combine agreement with reference event frequencies, temporal consistency, and data-supported variability in generated trajectories under specified conditions. For interactive agents and robotics, trajectory-level statistics could express a task-success floor alongside resource-use bounds and desired action patterns. These settings motivate studying whether target-dependent feedback can guide learning toward joint satisfaction as some requirements are met while others remain unmet, rather than relying on fixed reward combinations to produce the desired outcome. Extending the construction to sequential interaction would require sampling and credit-assignment schemes appropriate to trajectory dependencies.

E.3 Reproducibility Assets

Table 26 lists dataset and software identifiers and reproduction records. Appendix C details the implementation, training, and evaluation protocols.

Table 26: Dataset and software identifiers and reproduction assets.

Item	Identifier or status
TriviaQA	`mandarjoshi/trivia_qa`, `rc.nocontext`; revision `0f7faf33a3908546c6fd5b73a660e0f8ff173c2f`
Software	Python 3.11; PyTorch 2.11; Transformers 4.57; TRL 1.10.0; vLLM 0.21
Training and evaluation code	supplementary material (TB-GRPO code archive, revision `13500fe`)
Models	`Qwen/Qwen2.5-1.5B-Instruct`, `meta-llama/Llama-3.2-1B-Instruct`, `Qwen/Qwen2.5-7B-Instruct`
Checkpoints and generated outputs	to be released upon publication
Finite-policy verification	included in the code archive