# MDP-GRPO: Stabilized Group Relative Policy Optimization for Multi-Constraint Instruction Following

**Mohammad Mahdi Salmani-Zarchi[1]*, Zahra Rahimi[2], Heshaam Faili[1], Mohammad Javad Dousti[1,*]**

[1]Department of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran
[2]Department of Statistics, Mathematics and Computer Science, Allameh Tabataba'i University, Tehran, Iran

m.salmani78@ut.ac.ir za.rah@atu.ac.ir hfaili@ut.ac.ir mjdousti@ut.ac.ir

## Abstract

Reinforcement learning with verifiable rewards is ideal for multi-constraint instruction following, yet standard group-relative policy optimization (GRPO) becomes unstable under discrete, low-dispersion rewards, where within-group reward distributions are frequently homogeneous. We identify and formalize three pathologies of z-score group normalization in this regime: low-variance amplification, mean-centering blindness, and zero-variance collapse. To address them, we propose MDP-GRPO, which stabilizes learning through (1) multi-temperature sampling to increase reward dispersion, (2) dual-anchor advantages to restore gradients in homogeneous groups and stop mean-centering blindness, (3) prospect-theoretic shaping to bound updates and penalize violations based on Kahneman & Tversky's theory, and (4) asymmetric KL regularization. Evaluated on FollowBench, IFEval, and a curated multi-constraint dataset, MDP-GRPO outperforms standard GRPO, improving strict constraint satisfaction by up to 5.0% on Llama-3.2-3B. Our method also enables stable convergence with small group sizes while preserving general capabilities on MMLU and ARC [1].

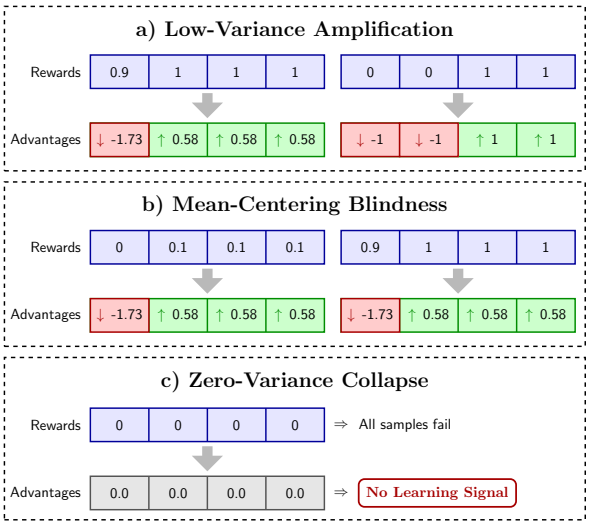



Figure 1: Illustration of group-normalized advantage pathologies in GRPO.

## 1 Introduction

Large language models (LLMs) can follow many natural-language instructions (Ouyang et al., 2022; Chung et al., 2024). Yet, they remain brittle when a request bundles multiple explicit constraints, such as asking the LLM to respond in a particular structure with an exact ending phrase, while adhering to strict lexical constraints and casing rules (Jiang et al., 2024; Geng et al., 2023; Park et al., 2025). In real deployments, these *multi-constraint* prompts are common: product and legal templates demand fixed formats (Westermann, 2024; Narendra et al., 2024), developer tools require machine-readable outputs (Schick et al., 2023; Yao et al., 2022; Shen et al., 2025), and safety constraints impose hard exclusions (Chen et al., 2025). In such settings, a response that is mostly correct but violates a single constraint is often unusable (Zhou et al., 2023).

Recently, reinforcement learning with *verifiable* rewards (RLVR) has emerged as a promising direction, where each constraint is checked deterministically and the model is optimized to satisfy as many constraints as possible (Wen et al., 2025; Guo et al., 2025). Relying on rule-based checkers is a deliberate design choice for domains requiring strict compliance; it provides a highly reliable learning signal, avoids expensive preference labels, and completely eliminates the bias and hallucinations inherent in learned reward models or LLM-as-a-judge evaluators (Chen et al., 2024; Ye et al., 2025). However, the resulting rewards are discrete and frequently low-variance early in training, making stable policy-gradient

* Corresponding authors.

[1]Our codes are available at https://github.com/m-salmani78/MDP-GRPO



*Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 42784–42797

learning notoriously challenging (Wen et al., 2025; Li et al., 2025).

In particular, when policy updates rely on group-relative normalization (as in GRPO-style methods) (Shao et al., 2024), multi-constraint reward structures induce three recurring pathologies (Figure 1). First, **low-variance amplification**: when within-group reward variance is small but nonzero, z-score normalization can inflate minor reward differences into disproportionately large advantages, yielding brittle updates. Second, **mean-centering blindness**: because z-score normalization discards absolute reward level, semantically distinct groups (e.g., consistently easy vs. consistently hard prompts) can receive nearly identical normalized advantage patterns, obscuring which cases truly require correction. Third, **zero-variance collapse**: homogeneous groups arise frequently early in training, and if all samples for a prompt satisfy or violate the same constraints, rewards will be equal and thereby group-normalized advantages collapse to zero, providing no learning signal. The resulting advantages can destabilize training and even trigger regressions in general capabilities (Lin et al., 2024; Kotha et al., 2024).

This paper introduces three independent and composable modules to make group-based RL reliable in these failure modes. We target both prevention and treatment. To prevent homogeneous groups, we employ *multi-temperature sampling*, mixing high-quality and exploratory completions to ensure within-group reward dispersion. To treat signal collapse and absolute reward blindness, we introduce *dual-anchor advantages* which interpolate between group-relative and goal-aware advantages. We then apply a bounded, asymmetric shaping function inspired by *Prospect Theory* (Kahneman and Tversky, 1979) to yield more human-aligned update magnitudes. While our primary focus is on deterministic and verifiable constraints, the core components of our approach are fundamentally agnostic to the reward source. They can be readily adapted to reward models or human-feedback settings.

We evaluate on our multi-constraint test set as well as established benchmarks (IFEVAL and FOLLOWBENCH). We additionally track MMLU and ARC to verify that instruction-following gains do not come at the expense of degrading general knowledge and reasoning capabilities (Hendrycks et al., 2020; Clark et al., 2018).

Overall, our contributions are: (1) We identify three failure modes of group-relative policy updates: low-variance amplification, mean-centering blindness, and zero-variance collapse; and we show how these arise naturally from discrete, low-dispersion reward structures. (2) We propose multi-temperature group sampling as a prevention mechanism that improves reward diversity in small-batch sampling (e.g., $G = 4$) without altering the underlying objective. (3) We introduce dual-anchor advantages to restore the learning signal in zero-variance and mean-centering blindness groups. (4) We apply bounded, asymmetric advantage shaping inspired by Prospect Theory to cap update magnitudes and emphasize violations, thereby improving robustness. (5) We evaluate the method on a custom multi-constraint test set, IFEVAL, and FOLLOWBENCH, demonstrating that our approach improves constraint satisfaction and stability with negligible overhead, while retaining general capabilities.

## 2 Related Work

### 2.1 Multi-constraint instruction data

Multi-constraint instruction following has been studied via synthetic constraint injection and large-scale data construction. He et al. (2024) show that training on compositional constraints improves complex instruction adherence and transfers to novel constraint combinations. RECAST further expands coverage with large, automatically verifiable multi-constraint data and validators that support both supervised training and RL with verifiable signals (Liu et al., 2025). We evaluate on established instruction-following benchmarks alongside a constraint-heavy test set (Zhou et al., 2023; Jiang et al., 2024).

### 2.2 RLVR and group-relative optimization

Reinforcement learning with verifiable rewards replaces preference modeling with deterministic checks and has recently enabled strong gains on reasoning-style tasks (Wen et al., 2025; Guo et al., 2025). Group-based optimizers such as GRPO estimate advantages from multiple samples per prompt via within-group normalization, avoiding an explicit critic (Shao et al., 2024). In discrete and partially sparse reward regimes, however, group normalization can become fragile under low-dispersion or homogeneous groups, motivating modifications to sampling and advantage estima-

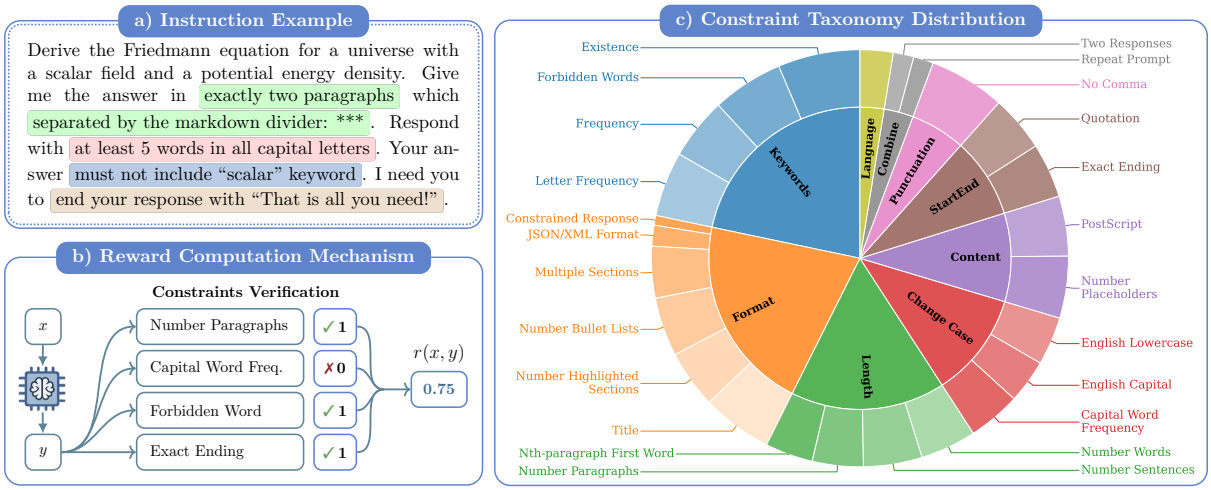



Figure 2: Characteristics of the dataset. (a) Example of a multi-constraint instruction. (b) Reward computation mechanism via rule-based constraint verification. (c) Taxonomy and relative frequency of constraint types.

tion.

### 2.3 Robust group advantages

Recent work addresses pathologies of group-relative advantages. MAPO proposes mixed, sample-adaptive advantage schemes to mitigate advantage allocation failures (Huang et al., 2025). NGRPO targets all-negative (homogeneously incorrect) groups and recovers learning signal via advantage calibration and asymmetric clipping (Nan et al., 2025). Our approach is complementary: we reduce homogeneous groups through temperature-diverse sampling, restore goal-aware signal with dual-anchor advantages, and apply bounded, loss-averse shaping to stabilize updates under multi-constraint RLVR.

### 2.4 Prospect-style, risk-aware objectives

Prospect-theoretic objectives have been explored for alignment, notably in Kahneman-Tversky Optimization (KTO) (Ethayarajh et al., 2024), which incorporates loss aversion via a human-centric utility function (Kahneman and Tversky, 1979). KTO applies this framework at the *objective level* to offline, un-paired preference data (e.g., binary thumbs-up/down labels). In contrast, our approach applies prospect-inspired shaping strictly at the *advantage* level for group-based RL, yielding bounded, asymmetric updates while keeping the underlying reward specification unchanged.

## 3 Problem Setup

### 3.1 Task Definition

Given an instruction $x$, a model generates a completion $y$. Each instruction specifies $C(x)$ explicit constraints, each of which is validated by a deterministic rule-based checker.

The score of a completion, $r(x, y) \in [0, 1]$, is defined by the fraction of satisfied constraints. Let $c_t(x, y) \in \{0, 1\}$ be the binary verification result of the $t$-th constraint. The resulting reward is:

$$r(x, y) = \frac{1}{C(x)} \sum_{t=1}^{C(x)} c_t(x, y). \quad (1)$$

This reward is used for RL training. We focus on *critic-free* group-based objectives, specifically Group Relative Policy Optimization (GRPO), which avoids the computational overhead of training a separate value network.

### 3.2 Multi-Constraint Instruction-Following Dataset

To study instruction following under explicit compositional constraints, we compile a dataset of 3,000 training prompts.

#### 3.2.1 Seed Prompts

We begin with a collection of seed instructions sourced through a hybrid approach. Approximately one-third of the seed prompts are sampled from the dataset introduced by He et al. (2024). The remaining prompts are newly curated to ensure a balanced coverage across three broad intent types: general Q&A, creative writing, and material assistance. All seed prompts are in English and represent the base request without explicit constraint markup, resembling natural user queries.

#### 3.2.2 Constraint Taxonomy

We adopt the constraint taxonomy from prior work on complex instruction synthesis (He et al., 2024).

We utilize 26 constraint types organized into 9 high-level categories (see Figure 2).

Each constraint type is instantiated through parameterized templates (e.g., varying forbidden words or numerical limits). Crucially, all constraints are designed to be automatically verifiable using deterministic validators (regular expressions, parsers), enabling fast, reproducible reward computation.

#### 3.2.3 Constraint Injection Procedure

Given a seed instruction $x$, we inject between 1 and 6 constraints. To ensure validity, we enforce strict compatibility rules (e.g., preventing contradictory casing constraints). We filter constraints that conflict with those already selected, then render the final instruction by appending natural-language descriptions of the chosen constraints to the seed prompt. The underlying formal constraint set $\{c_t\}_{t=1}^{C(x)}$ is stored for evaluation. Constraint counts follow a unimodal distribution peaking at 3–5 constraints, emphasizing moderately complex compositions.

### 3.3 Baseline: Group Relative Policy Optimization

We adopt GRPO (Shao et al., 2024) as our baseline. For each query $x$, GRPO samples a group of $G$ outputs $\{y_1, y_2, \ldots, y_G\}$ from the old policy $\pi_{\theta_{old}}$. The policy is optimized to maximize the surrogate objective:

$$\mathcal{J}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\{y_i\}\sim\pi_{\theta_{old}}}\Big[ \frac{1}{G}\sum_{i=1}^{G}\min\left(\frac{\pi_\theta(y_i|x)}{\pi_{\theta_{old}}(y_i|x)}A_i, \text{clip}(\ldots)A_i\right) - \beta_{KL}\mathbb{D}_{KL}\Big], \quad (2)$$

where $\beta_{KL}$ controls the KL-divergence penalty. Crucially, GRPO computes the advantage $A_i$ for the $i$-th completion using group-relative z-score normalization:

$$A_i = \frac{r_i - \mu_{\text{group}}}{\sigma_{\text{group}} + \epsilon}, \quad (3)$$

where $r_i = r(x, y_i)$ is the reward of completion $y_i$, while $\mu_{\text{group}}$ and $\sigma_{\text{group}}$ are the mean and standard deviation of rewards within the sampled group.

## 4 Method

We introduce MDP-GRPO (visualized in Figure 3), a method designed to stabilize policy gradients for multi-constraint tasks. Our approach addresses the pathologies of standard group-relative advantages through three complementary mechanisms: (1) Temperature-diverse sampling to prevent zero-variance groups; (2) Dual-anchor advantages to maintain signal when group variance collapses; and (3) Prospect-theoretic shaping to bound updates and enforce loss aversion.

### 4.1 Motivation: Pathologies of Group Normalization

As defined in Section 3.3, standard GRPO relies on the z-scored advantage $z_i = (r_i - \mu_{\text{group}})/(\sigma_{\text{group}} + \epsilon)$. In the multi-constraint regime, this formulation suffers from three specific failure modes:

1. **Low-Variance Amplification:** When $\sigma_{\text{group}}$ is small but non-zero, minor reward fluctuations are inflated into disproportionately large advantages, causing brittle updates.

2. **Zero-Variance Collapse:** When all completions in a group satisfy the exact same subset of constraints (a common occurrence with deterministic checkers), $\sigma_{\text{group}} \to 0$. The advantage becomes undefined or noise-dominated (determined solely by $\epsilon$), providing no valid learning signal.

3. **Mean-Centering Blindness:** A group where the model fails no constraints and a group where it fails all constraints can produce identical normalized advantages, discarding critical absolute performance information.

### 4.2 Temperature-Diverse Group Sampling

To mitigate the frequency of homogeneous groups, we modify the sampling strategy. Instead of sampling all $G$ outputs with a fixed temperature, we utilize a temperature schedule $\boldsymbol{T} = [\tau_1, \ldots, \tau_G]$.

For the $i$-th element of the group, we sample $y_i \sim \pi_{\theta_{\text{old}}}(\cdot \mid x; \tau_i)$. We configure $\boldsymbol{T}$ to purposefully mix exploitation and exploration, e.g., $\boldsymbol{T} = [0.1, 0.4, 0.7, 1.0]$. Lower temperatures encourage exploitation by stabilizing a high-quality, "best-effort" baseline, while higher temperatures induce exploration, increasing the likelihood of within-group reward dispersion without altering the underlying group-based RL objective.

### 4.3 Dual-Anchor Advantages

To address mean-centering blindness and zero-variance collapse, we introduce a goal-aware

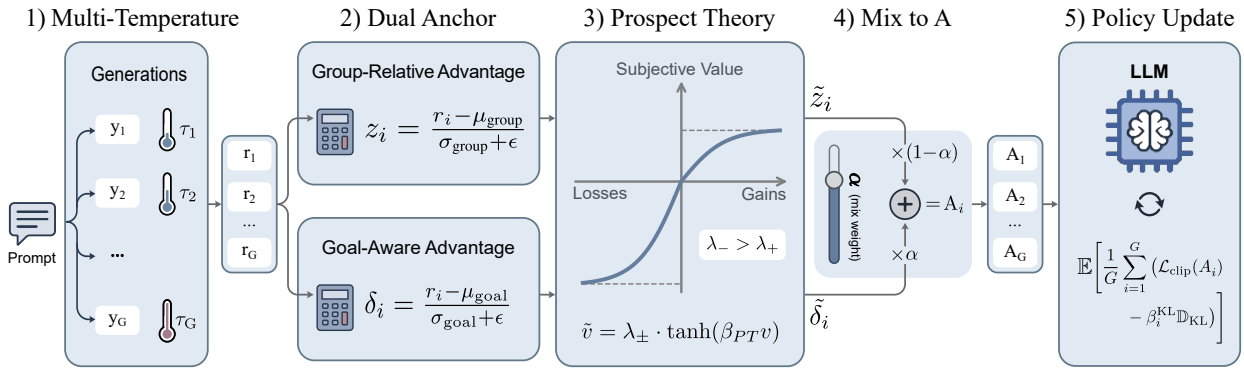



Figure 3: MDP-GRPO pipeline. (1) Multi-temperature sampling generates a group of $G$ completions per prompt. (2) Compute group-relative $z_i$ and goal-aware $\delta_i$ advantage signals. (3) Prospect-theoretic shaping yields $\tilde{z}_i$ and $\tilde{\delta}_i$. (4) Mixing $\tilde{z}_i$ and $\tilde{\delta}_i$ gives $A_i$. (5) Apply a standard GRPO policy update using $A_i$.

anchor. We model a *neutral* baseline as a binomial process where each of the $C(x)$ constraints is satisfied independently with probability $p = 0.5$. For a reward normalized to $[0, 1]$, this neutral baseline has mean $\mu_{\text{goal}} = 0.5$ and standard deviation $\sigma_{\text{goal}} = \frac{1}{2\sqrt{C(x)}}$. We define the absolute, goal-aware advantage $\delta_i$ as:

$$\delta_i = 2\sqrt{C(x)}(r_i - 0.5). \tag{4}$$

This scales the reward such that a completion satisfying all constraints yields $\delta_i = +\sqrt{C(x)}$, and one satisfying none yields $\delta_i = -\sqrt{C(x)}$. Crucially, $\delta_i$ remains well-defined and informative even when the group variance is zero, providing our second anchor alongside the standard group-relative advantage $z_i$.

### 4.4 Prospect-Theoretic Advantage Shaping

Prospect Theory (Kahneman and Tversky, 1979) is a foundational model in behavioral economics that describes how humans make decisions involving risk. It demonstrates a distinct, non-linear relationship between objective outcomes and their perceived subjective value. A core tenet of this theory is *loss aversion*: the psychological pain of losing 50 dollars is experienced much more intensely than the satisfaction of winning 50 dollars. Additionally, human perception exhibits diminishing sensitivity; as the absolute magnitude of gains or losses increases, marginal changes are felt less strongly.

In the context of GRPO, the advantage signals act as these objective outcomes. Standard policy gradient updates treat advantages linearly. However, raw standard advantages can grow unboundedly (especially when the batch variance $\sigma$ is extremely small). Furthermore, treating positive and negative completions symmetrically fails to adequately penalize the model when it regresses on strict constraints.

Inspired by Prospect Theory, we frame the raw advantage as an objective outcome that must be transformed into a subjective, "human-aligned" value signal. We apply a bounded, asymmetric shaping function to our raw advantage signals before mixing them.

Let $v_i \in \{z_i, \delta_i\}$ represent either of the raw advantage signals. We approximate the prospect-theoretic value function using a scaled 'tanh' transformation to compute the shaped advantage $\tilde{v}_i$:

$$\tilde{v}_i = \begin{cases} \lambda_+ \tanh(\beta_{\text{PT}} v_i) & \text{if } v_i \geq 0 \\ \lambda_- \tanh(\beta_{\text{PT}} v_i) & \text{if } v_i < 0 \end{cases} \tag{5}$$

By setting $\lambda_- > \lambda_+ > 0$, we enforce loss aversion: negative outcomes (constraint violations) result in larger magnitude gradient updates than equivalent positive outcomes, heavily penalizing regression. Concurrently, the $\tanh$ function captures diminishing sensitivity and strictly bounds the advantages to $[-\lambda_-, \lambda_+]$, preventing unbounded gradients and stabilizing training. Applying this function yields the shaped group-relative advantage $\tilde{z}_i$ and the shaped goal-aware advantage $\tilde{\delta}_i$.

Finally, we compute the final combined advantage $A_i$ by interpolating between the two shaped signals:

$$A_i = (1 - \alpha)\tilde{z}_i + \alpha\tilde{\delta}_i, \tag{6}$$

where $\alpha \in [0, 1]$ controls the reliance on the absolute anchor.

**Theoretical Validity.** Crucially, because the mixed baseline relies only on the group mean

and a fixed goal-aware anchor, it remains action-independent for any individual sample within the group. Therefore, its incorporation is theoretically valid under the policy gradient theorem (Sutton et al., 1999). While this introduces a controlled bias relative to the standard zero-mean GRPO estimator, this goal-aware bias is essential for stabilizing gradients in low-dispersion regimes where group normalization becomes ill-conditioned (i.e., as $\sigma_{\text{group}} \to 0$). We discuss and empirically validate this bias-variance tradeoff, alongside the sensitivity of the mixing weight $\alpha$, in Appendix B.

### 4.5 MDP-GRPO Objective and Algorithm

We integrate these components into the GRPO framework. We also optionally employ an *asymmetric KL penalty* to permit larger deviations when the model is improving (positive advantage) while strictly constraining it when performance drops.

The final objective is:

$$\mathcal{J}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\left(\mathcal{L}_{\text{clip}}(A_i) - \beta_i^{\text{KL}}\mathbb{D}_{\text{KL}}\right)\right], \quad (7)$$

where $\mathcal{L}_{\text{clip}}$ is the standard GRPO clipping term using $A_i$, and the asymmetric KL coefficient is:

$$\beta_i^{\text{KL}} = \begin{cases} \beta_{\text{low}} & \text{if } A_i \geq 0 \\ \beta_{\text{high}} & \text{if } A_i < 0 \end{cases} \quad (8)$$

The full training procedure is detailed in Algorithm 1 in Section A.

## 5 Experiments

### 5.1 Benchmarks

We evaluate instruction-following on two standard benchmarks: IFEVAL (Zhou et al., 2023), which uses strict code-based verification, and FOLLOWBENCH (Jiang et al., 2024), which utilizes a hybrid evaluation framework (rules and LLM judges) to assess fine-grained constraints. To complement these public benchmarks, we additionally report results on our custom test set of 500 multi-constraint prompts, created using a workflow similar to the one described in Section 3.2.

### 5.2 Evaluation Metrics

Following established protocols (Zhou et al., 2023; Jiang et al., 2024), we report two complementary accuracy metrics. Let $N$ be the number of evaluation prompts. For the $i$-th prompt with $C_i$ constraints, let $c_{i,j} \in \{0, 1\}$ denote the satisfaction status of the $j$-th constraint.

**Constraint Accuracy (SSR).** We define *Soft Success Rate* (SSR) as the mean constraint satisfaction rate across all instructions, analogous to $\text{acc}_{\text{con}}$ in prior work. Formally, $\text{SSR} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{C_i}\sum_{j=1}^{C_i} c_{i,j}\right)$. This metric captures partial progress under sparse multi-constraint rewards.

**Prompt Accuracy (HSR).** We define *Hard Success Rate* (HSR) as the strict all-constraints success rate, analogous to $\text{acc}_{\text{ins}}$. It requires satisfying every constraint within a prompt simultaneously: $\text{HSR} = \frac{1}{N}\sum_{i=1}^{N}\prod_{j=1}^{C_i} c_{i,j}$. This reflects whether an output is fully usable in strict template-like settings.

### 5.3 Models

We conduct experiments on two instruction-tuned models: Gemma-2-2B-Instruct and Llama-3.2-3B-Instruct. Unless otherwise stated, the policy is initialized from the corresponding instruct model, and the reference policy $\pi_{\text{ref}}$ used for KL regularization is this frozen initialization checkpoint.

### 5.4 Compared Methods

To isolate the contribution of each module, we compare six training variants. We evaluate the zero-shot Baseline (the initial instruct model) and standard GRPO against our individual ablations: MT-GRPO (Multi-Temperature sampling), DA-GRPO (Dual-Anchor advantages), PT-GRPO (Prospect-Theoretic shaping), and DA-PT-GRPO (combining DA and PT). Finally, we evaluate MDP-GRPO, our full pipeline combining all three mechanisms to address stability and performance in multi-constraint regimes.

### 5.5 Training Setup

We primarily train with a group size of $G = 8$ completions per prompt. We also perform ablations with $G = 4$ to demonstrate the efficacy of multi-temperature sampling in small-batch regimes. We train all models on a single NVIDIA A100 GPU.

**Decoding.** We use stochastic decoding with `do_sample=True`. For MT variants, the group is sampled using a fixed temperature schedule $\boldsymbol{T} = [\tau_1, \ldots, \tau_G]$. For standard methods, we use a fixed temperature. We use `top_p` nucleus sampling

| Models (G=8) | Gemma-2-2B-Instruct | | | | | | Llama-3.2-3B-Instruct | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Benchmarks | IFEVAL | | Custom Test Set | | FollowBench | | IFEVAL | | Custom Test Set | | FollowBench | |
| | SSR | HSR | SSR | HSR | SSR | HSR | SSR | HSR | SSR | HSR | SSR | HSR |
| Baseline | 56.7 | 45.1 | 54.8 | 18.8 | 63.7 | 52.9 | 54.2 | 46.8 | 60.3 | 20.8 | 69.7 | 59.8 |
| GRPO | 73.7 | 62.4 | 68.4 | 29.0 | 64.0 | 53.2 | 66.1 | 58.5 | 65.1 | 24.8 | 68.4 | 58.9 |
| MT-GRPO | 73.9 | 63.7 | 68.6 | 29.6 | 64.3 | 53.7 | 68.1 | 59.1 | 65.4 | 24.6 | 70.0 | 59.9 |
| DA-GRPO | 75.2 | 64.7 | 70.7 | 32.6 | 65.4 | 56.1 | 71.2 | **62.5** | 66.1 | 24.9 | 69.6 | 58.9 |
| PT-GRPO | **75.7** | **65.8** | 71.4 | 30.6 | 66.5 | 56.1 | 70.1 | 61.3 | 64.7 | 23.6 | 69.3 | 57.6 |
| DA-PT-GRPO | 74.8 | 64.3 | **70.8** | **33.4** | **68.2** | **59.7** | **71.5** | 60.4 | **66.3** | **25.4** | **70.2** | 59.2 |
| MDP-GRPO | 75.3 | 64.1 | 70.3 | 32.8 | 66.9 | 57.4 | 71.3 | 59.8 | 65.8 | 25.2 | 69.4 | 59.1 |

Table 1: Instruction-following performance of GRPO variants using a standard group size ($G = 8$). Results are reported using SSR and HSR across three benchmarks.

| Models (G=4) | Gemma-2-2B-Instruct | | | | | | Llama-3.2-3B-Instruct | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Benchmarks | IFEVAL | | Custom Test Set | | FollowBench | | IFEVAL | | Custom Test Set | | FollowBench | |
| | SSR | HSR | SSR | HSR | SSR | HSR | SSR | HSR | SSR | HSR | SSR | HSR |
| Baseline | 56.7 | 45.1 | 54.8 | 18.8 | 63.7 | 52.9 | 54.2 | 46.8 | 60.3 | 20.8 | 69.7 | 59.8 |
| GRPO | 69.7 | 58.2 | 67.3 | 28.6 | 64.5 | 53.5 | 67.2 | 55.0 | 63.3 | 21.6 | 69.4 | 60.5 |
| MT-GRPO | 71.1 | 59.4 | 68.4 | **30.6** | 64.2 | 53.5 | **70.5** | **58.4** | 63.4 | **22.8** | 69.9 | **60.7** |
| DA-GRPO | 70.5 | 58.9 | 67.3 | 28.5 | 64.6 | 53.6 | 69.3 | 56.5 | 63.5 | 22.6 | 68.8 | 59.6 |
| PT-GRPO | 69.2 | 56.6 | **68.7** | 29.9 | **66.4** | **55.7** | 69.7 | 58.2 | 63.1 | 22.0 | **71.1** | 61.6 |
| DA-PT-GRPO | 70.1 | 58.9 | 67.2 | 29.4 | 63.9 | 53.4 | 68.5 | 56.9 | **64.0** | 22.6 | 69.4 | 59.2 |
| MDP-GRPO | **71.2** | **59.5** | 67.8 | 30.4 | 65.8 | 55.3 | 68.5 | 57.0 | 63.7 | **22.8** | 70.2 | 60.7 |

Table 2: Instruction-following performance of GRPO variants using a reduced group size ($G = 8$). Results are reported using SSR and HSR across three benchmarks.

with $p = 0.9$ and a maximum generation length of 1024 tokens.

**Optimization.** We use a learning rate of $1 \times 10^{-5}$, batch size 32, PPO clip $\epsilon_{\text{clip}} = 0.2$, and a base KL coefficient $\beta_{\text{KL}} = 0.01$. For runs utilizing asymmetric KL, we set $\beta_{\text{KL}}^{\text{high}} = 0.025$ (applied when advantages are negative) and $\beta_{\text{KL}}^{\text{low}} = 0.01$. Models are trained for one epoch.

**Method Hyperparameters.** For dual-anchor variants, we evaluate static mixing weights $\alpha \in \{0.1, 0.2, 0.4\}$. Unless otherwise noted, we use $\alpha = 0.2$ and set the anchor mean $\mu_{\text{goal}} = 0.5$ (neutral baseline). For Prospect Shaping variants, we set $\beta_{\text{PT}} = 0.8$ to approximate empirical findings in behavioral economics, and $(\lambda_+, \lambda_-) = (1.25, 2.0)$. This configuration ensures a slope of $\approx 1.0$ for positive advantages.

**Ablations and Sensitivity.** We report additional sensitivity analyses for the dual-anchor design (mixing weight $\alpha$ and the goal-aware center in $\delta_i$) in Section B. Ablations for prospect-theoretic shaping and asymmetric KL regularization are reported in Section C.

### 5.6 General Capability Evaluation

To assess alignment tax, we evaluate all trained policies on MMLU (Hendrycks et al., 2020) and ARC (Easy and Challenge) (Clark et al., 2018) in Section D.

## 6 Results

### 6.1 Main Results on Instruction Following

Table 1 reports instruction-following performance under our standard group size ($G = 8$). Across Gemma-2-2B-Instruct and Llama-3.2-3B-Instruct, MDP-GRPO variants generally improve over both the baseline model and standard GRPO on IFEval, FollowBench, and our Custom Test Set, with the largest gains appearing on harder multi-constraint prompts.

On IFEval, the individual components exhibit mechanistic complementarity rather than strict uniform dominance. This aligns with their targeted failure modes (see Section 6.3): stability diagnostics show DA mitigates homogeneous-

group collapse, PT maintains competitive rewards with controlled KL drift, and MT increases within-group diversity. Consequently, rather than a single ablation dominating everywhere, the best-performing variant depends on the model and benchmark. For example, on Gemma-2-2B, PT-GRPO yields the strongest strict success (HSR 65.8% vs. 62.4% for GRPO). On Llama-3.2-3B, DA-PT-GRPO achieves the peak soft success rate (SSR 71.5%), slightly edging out the full MDP-GRPO pipeline (71.3%). The full pipeline yields the most consistent stability-performance profile across all benchmarks, even if it does not always achieve the peak score on any single metric.

On the Custom Test Set, which emphasizes difficult constraint compositions, anchoring proves most consistently beneficial. For example, DA-GRPO improves Gemma-2-2B strict success from 29.0% (GRPO) to 32.6%, consistent with mitigating zero-variance collapse.

**Impact of Group Size and Compute Trade-offs.** To further analyze the role of multi-temperature sampling, we evaluated both models using a restricted group size of $G = 4$ (Table 2). In this low-diversity regime, the gains from MT are significantly magnified. For Gemma-2-2B, MT-GRPO improves over standard GRPO by $+1.4\%/+1.3\%$ (SSR/HSR) on IFEval, and $+1.1\%/+\mathbf{2.0}\%$ on the Custom Test Set. Notably, MT-GRPO at $G = 4$ approaches the performance of GRPO at $G = 8$ (e.g., attaining an HSR of 30.6% vs. 29.0% for $G = 8$ GRPO on the Custom set). Furthermore, when averaging across benchmarks under $G = 4$, the full MDP-GRPO pipeline achieves the highest overall performance, whereas DA-PT-GRPO yielded the best average under $G = 8$. This regime dependence confirms that MT is crucial for recovering performance under tight compute budgets by synthetically injecting within-group diversity.

### 6.2 Performance by Difficulty and Category

Figure 4 stratifies strict success by the number of constraints. All methods degrade with difficulty, but stabilized variants degrade more slowly in the high-complexity regime: the unaligned baseline drops below 10% HSR by Difficulty 4, while DA-PT-GRPO retains roughly 20% HSR at Difficulty 5 (vs. ≈12% for GRPO). This matches the intended role of anchoring and shaping in preventing near-zero learning signals when feasible solutions are rare.

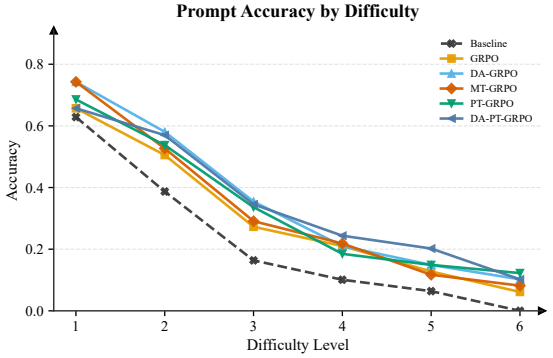



Figure 4: Prompt accuracy (HSR) by difficulty, defined by the number of constraints per prompt (1–6).

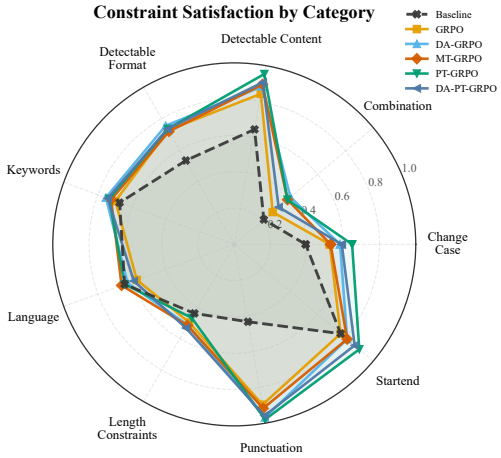
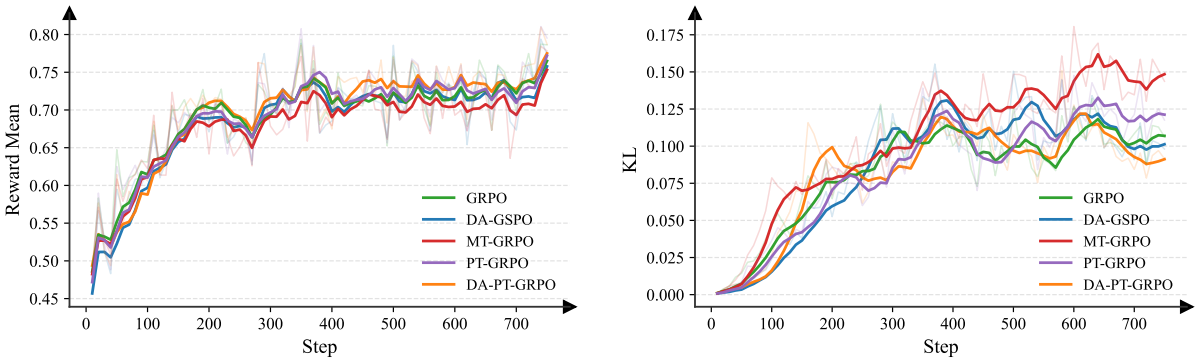



Figure 5: Constraint satisfaction by category. Each spoke reports instruction-level satisfaction (SSR).

Figure 5 shows that improvements are concentrated on deterministic, verifiable compliance constraints (e.g., format/content, punctuation, length and boundary constraints), where RLVR drives satisfaction toward near-ceiling levels. In contrast, gains are limited for constraints that depend primarily on semantic knowledge (e.g., language), suggesting that verifiable rewards are most effective at correcting compliance failures rather than adding new knowledge.

### 6.3 Training Dynamics and Failure-Mode Diagnostics

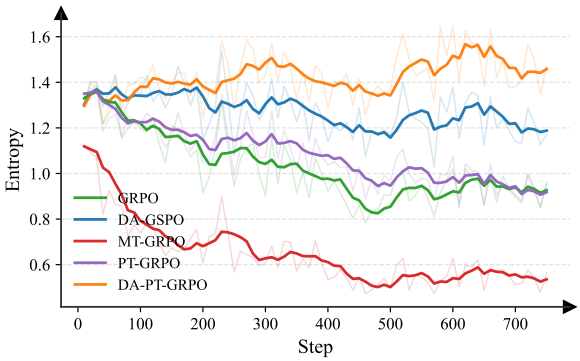
Figure 6 connects training behavior to our targeted failure modes. DA-GRPO reduces `frac_reward_zero_std`, indicating fewer effectively homogeneous groups and supporting its role in mitigating collapse under discrete multi-constraint rewards. PT-GRPO achieves comparable reward while maintaining a controlled KL divergence trajectory (and DA-PT-GRPO further suppresses KL drift), consistent with bounded, loss-averse shaping stabilizing updates. MT-GRPO can



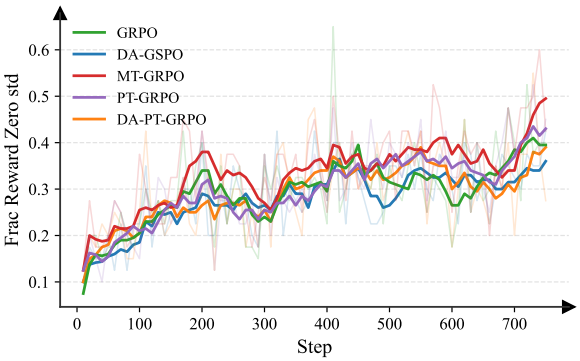

(a) Training reward mean




(b) KL divergence during training




(c) Policy entropy over training




(d) Homogeneous-group rate

Figure 6: Training dynamics and stability diagnostics (G=8). (a) Mean verifiable reward, (b) KL divergence to the reference policy, (c) policy entropy, and (d) fraction of groups with zero within-group reward standard deviation.

increase KL and reduce entropy under our current schedule, motivating careful decoding control and temperature tuning.

### 6.4 Summary of Findings

Our findings indicate that (i) standard GRPO substantially improves verifiable instruction following over the zero-shot baseline; (ii) dual anchoring mitigates homogeneous-group collapse, yielding higher strict success on complex prompts; and (iii) prospect-theoretic shaping improves training stability by controlling KL drift while preserving reward gains, with multi-temperature sampling providing additional benefits in low-diversity regimes.

## 7 Conclusion

We investigated reinforcement learning with verifiable rewards for multi-constraint instruction following, where rewards are discrete and sometimes low-variance. We identify three failure modes of standard group-relative advantages and propose MDP-GRPO, which combines temperature-diverse group sampling, dual-anchor advantages, and prospect-theoretic shaping. Across IFEval, FollowBench, and a custom multi-constraint evaluation set, MDP-GRPO variants consistently improved both soft and hard success rates, achieving the most significant gains on more complex instructions. Training diagnostics confirm that our modifications successfully reduce homogeneous-group collapse and control KL drift, ultimately enhancing alignment without severely degrading general capabilities.

## Limitations

Our methodology relies on explicit constraints verified by deterministic automatic checkers. While this ensures reliable and computationally inexpensive rewards, it does not naturally extend to subjective, stylistic, or underspecified constraints. Applying MDP-GRPO to such open-ended scenarios would likely necessitate learned reward models or preference feedback, thereby reintroducing potential reward misspecification and judge bias.

Furthermore, MDP-GRPO introduces several hyperparameters, including anchor mixing weights, shaping parameters, and temperature schedules. Although we provide sensitivity analyses for the core components, transferring this approach to

vastly different domains or reward scales may require recalibration and careful monitoring of KL divergence.

Finally, our empirical validation focuses on standard instruction-following benchmarks, a custom complex constraint set, and two specific instruction-tuned model sizes (2B and 3B). Broader validation across larger model families, multilingual contexts, and highly structured domains (e.g., complex tool use or code generation) remains an area for future work to fully characterize the generalizability of our approach.

## Ethical Considerations

The primary objective in this work is improving reliability on *explicit, automatically checkable* constraints (e.g., formatting and structural requirements). Our method does not, by itself, ensure broader safety properties such as harmlessness or factuality beyond what is encoded in the reward specification.

## Acknowledgments

We thank MCILab for supporting this research and providing the computational resources used in this study.

## References

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or LLMs as the Judge? A Study on Judgement Bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Xin Chen, Yarden As, and Andreas Krause. 2025. Learning Safety Constraints for Large Language Models. *arXiv preprint arXiv:2505.24445*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, and 1 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv preprint arXiv:1803.05457*.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. KTO: Model Alignment as Prospect Theoretic Optimization. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*. PMLR.

Saibo Geng, Martin Josifoski, Maxime Peyrard, and Robert West. 2023. Grammar-Constrained Decoding for Structured NLP Tasks without Finetuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. DeepSeek-R1 incentivizes reasoning in llms through reinforcement learning. *Nature*.

Qianyu He, Jie Zeng, Qianxi He, Jiaqing Liang, and Yanghua Xiao. 2024. From Complex to Simple: Enhancing Multi-Constraint Complex Instruction Following Ability of Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Miami, Florida, USA. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring Massive Multitask Language Understanding. *arXiv preprint arXiv:2009.03300*.

Wenke Huang, Quan Zhang, Yiyang Fang, Jian Liang, Xuankun Rong, Huanjin Yao, Guancheng Wan, Ke Liang, Wenwen He, Mingjun Li, Leszek Rutkowski, Mang Ye, Bo Du, and Dacheng Tao. 2025. MAPO: Mixed Advantage Policy Optimization. *arXiv preprint arXiv:2509.18849*.

Yuxin Jiang, Yufei Wang, Xingshan Zeng, Wanjun Zhong, Liangyou Li, Fei Mi, Lifeng Shang, Xin Jiang, Qun Liu, and Wei Wang. 2024. FollowBench: A multi-level fine-grained constraints following benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Daniel Kahneman and Amos Tversky. 1979. Prospect Theory: An Analysis of Decision under Risk. *Econometrica*.

Suhas Kotha, Jacob Mitchell Springer, and Aditi Raghunathan. 2024. Understanding Catastrophic Forgetting in Language Model Fine-tuning. In *NeurIPS 2023 Workshop on Distribution Shifts: New Frontiers with Foundation Models*.

Wenyun Li, Wenjie Huang, and Chen Sun. 2025. Shaping Sparse Rewards in Reinforcement Learning. *arXiv preprint arXiv:2501.19128*.

Yong Lin, Hangyu Lin, Wei Xiong, Shizhe Diao, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang Wang, Wenbin Hu, Hanning Zhang, Hanze Dong, Renjie Pi, Han Zhao, Nan Jiang, Heng Ji, Yuan Yao, and Tong Zhang. 2024. Mitigating the Alignment Tax of RLHF. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Wenhao Liu, Zhengkang Guo, Mingchen Xie, Jingwen Xu, Zisu Huang, Muzhao Tian, Jianhan Xu, Muling Wu, Xiaohua Wang, Changze Lv, He-Da Wang, Hu Yao, Xiaoqing Zheng, and Xuanjing Huang. 2025. RECAST: Expanding the boundaries of LLMs' complex instruction following with multi-constraint data. In *Submitted to The Fourteenth International Conference on Learning Representations*. Under review.

Gongrui Nan, Siye Chen, Jing Huang, Mengyu Lu, Dexun Wang, Chunmei Xie, Weiqi Xiong, Xianzhou Zeng, Qixuan Zhou, Yadong Li, and 1 others. 2025. NGRPO: Negative-enhanced group relative policy optimization. *arXiv preprint arXiv:2509.18851*.

Savinay Narendra, Kaushal Shetty, and Adwait Ratnaparkhi. 2024. Enhancing Contract Negotiations with LLM-Based Legal Document Comparison. In *Natural Legal Language Processing Workshop*. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*.

Kanghee Park, Timothy Zhou, and Loris D'Antoni. 2025. Flexible and Efficient Grammar-Constrained Decoding. *arXiv preprint arXiv:2502.05111*.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. *Advances in Neural Information Processing Systems*.

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

Zhengyuan Shen, Darren Yow-Bang Wang, Soumya Smruti Mishra, Zhichao Xu, Yifei Teng, and Haibo Ding. 2025. SLOT: Structuring the Output of Large Language Models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*. Association for Computational Linguistics.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. 1999. Policy Gradient Methods for Reinforcement Learning with Function Approximation. In *Advances in Neural Information Processing Systems*, volume 12. MIT Press.

Xumeng Wen, Zihan Liu, Shun Zheng, Zhijian Xu, Shengyu Ye, and 1 others. 2025. Reinforcement Learning with Verifiable Rewards Implicitly Incentivizes Correct Reasoning in Base LLMs. *arXiv preprint arXiv:2506.14245*.

Hannes Westermann. 2024. Dallma: Semi-Structured Legal Reasoning and Drafting with Document Automation and LLM Assistance. In *ICML Workshop (GenLaw)*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. ReAct: Synergizing Reasoning and Acting in Language Models. In *The eleventh international conference on learning representations*.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. 2025. Justice or Prejudice? Quantifying Biases in LLM-as-a-Judge. In *The Thirteenth International Conference on Learning Representations*.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-Following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

## A MDP-GRPO Algorithm

Algorithm 1 details the full training loop for MDP-GRPO. The procedure differs from standard GRPO through four key additions: (1) sampling is performed with a heterogeneous temperature schedule $\mathbf{T}$ to ensure group diversity; (2) absolute goal-aware anchors ($\delta_i$) are computed alongside standard group-relative scores ($z_i$); (3) prospect-theoretic advantage shaping is applied via a bounded, asymmetric function to enforce loss aversion; and (4) the advantages are mixed via a convex combination before being used in the PPO-style update with an asymmetric KL penalty.

**Algorithm 1** MDP-GRPO for Multi-Constraint Instruction Following

**Require:** Policy $\pi_\theta$, Reference Policy $\pi_{\text{ref}}$, Temperature Schedule $\mathbf{T}$, Mixing Weight $\alpha$, Shaping Params $(\lambda_+, \lambda_-, \beta_{\text{PT}})$.

**for** each training step **do**
  Sample a batch of prompts $x \sim \mathcal{D}$.
  **for** each prompt $x$ **do**
    Sample $G$ completions $y_i \sim \pi_{\theta_{\text{old}}}(\cdot \mid x; \tau_i)$ using schedule $\mathbf{T}$.
    Compute rewards $r_i$ for all $i \in \{1, \ldots, G\}$.
    Compute standard group-relative advantages $z_i$ (Eq. 3).
    Compute absolute goal-aware anchors $\delta_i$ (Eq. 4).
    Apply Prospect-Theoretic shaping to advantages (Eq. 5).
    Compute mixed advantages (Eq. 6).
  **end for**
  Compute the surrogate loss $\mathcal{J}(\theta)$ (Eq. 7) including the Asymmetric KL penalty.
  Update $\theta$ via gradient descent.
**end for**

## B Dual-Anchoring Ablations

### B.1 Effect of the dual-anchor mixing weight $\alpha$

Introducing $\delta_i$ modifies the effective advantage and optimizes a surrogate objective. This introduces a controlled bias relative to standard GRPO, analogous to other widely used modifications such as PPO clipping or advantage normalization that trade strict unbiasedness for drastically reduced variance under finite-sample noise. This bias-variance tradeoff is continuously controlled by $\alpha$: when $\alpha = 0$, we recover standard GRPO exactly. In low-dispersion discrete-reward regimes where group normalization becomes ill-conditioned, the variance reduction from $\delta_i$ dominates.

Figure 7c compares GRPO with dual-anchoring (DA-GRPO) for $\alpha \in \{0.2, 0.4\}$. Across training, the mean verifiable reward rises similarly for all settings and converges to comparable final values, indicating that dual-anchoring does not hinder optimization. However, increasing $\alpha$ noticeably increases the KL divergence: $\alpha = 0.4$ produces substantially larger KL throughout training, suggesting more aggressive policy shifts driven by the absolute (anchor-based) component. Since $\alpha = 0.4$ does not yield commensurate reward gains but incurs higher KL (and thus stronger deviation from the reference policy), we use a moderate value ($\alpha = 0.2$) as a better reward–stability trade-off.

### B.2 Sensitivity to the goal-aware center used in $\delta$

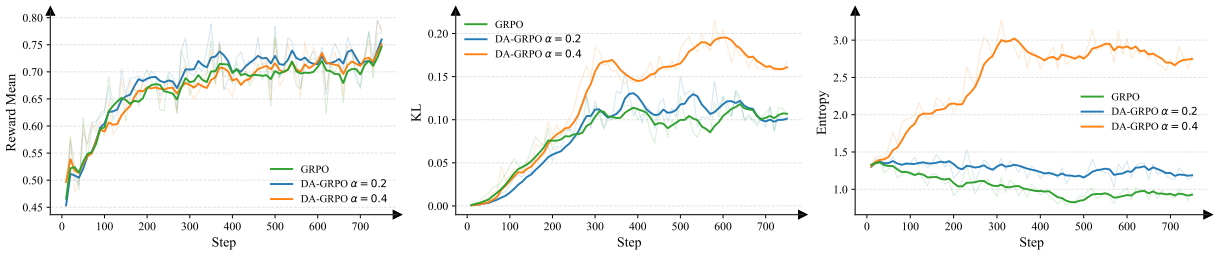
Figure 8c compares two choices for the anchor center in the goal-aware term: (i) $\hat{\mu} = (\mu + C)/2$, which shifts the reference point upward as training progresses, and (ii) $\hat{\mu} = \max(\mu, C/2)$ (for $r \in [0, 1]$, this is $\max(\mu, 0.5)$), which uses the neutral binomial mean unless the group is already above it.

The results show that $\hat{\mu} = (\mu + C)/2$ is unstable: it produces a large and sustained increase in KL divergence and a sharp rise in policy entropy, while achieving substantially *lower* reward than both GRPO and the alternative anchor. This suggests that aggressively moving the center upward can induce overly strong (and effectively noisy) updates that drift far from the reference policy without yielding commensurate improvement in constraint satisfaction.

In contrast, $\hat{\mu} = \max(\mu, C/2)$ tracks GRPO closely in KL and entropy (remaining near the reference policy) while matching or slightly improving the reward trajectory. Overall, using a conservative, piecewise center that falls back to the neutral binomial mean appears to provide a better stability–performance trade-off than continuously increasing the anchor target.

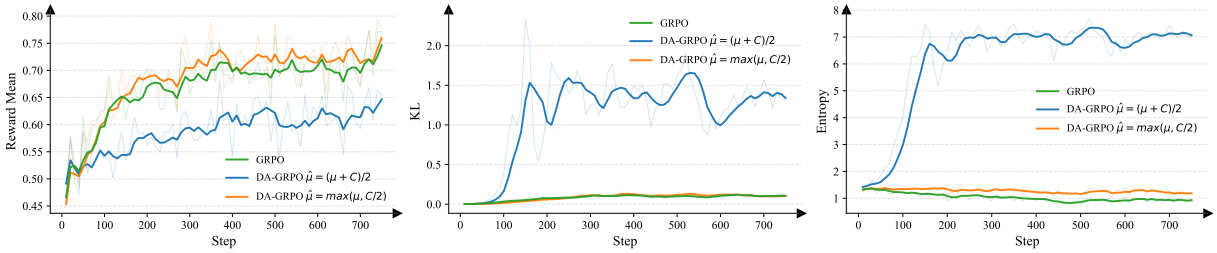
## C Prospect-theoretic Shaping and Asymmetric KL Ablations

Figure 9d studies how loss-averse shaping and its variants affect training dynamics under GRPO.




(a) Training reward mean (b) KL divergence during training (c) Policy entropy over training

Figure 7: Dual-anchor mixing weight $\alpha$ sweep. Training dynamics for GRPO and dual-anchor GRPO (DA-GRPO) with $\alpha \in \{0.2, 0.4\}$. Panels show (a) mean verifiable reward, (b) KL divergence to the reference policy, and (c) policy entropy, highlighting that larger $\alpha$ induces more aggressive policy updates (higher KL/entropy) with limited additional reward gain.




(a) Training reward mean (b) KL divergence during training (c) Policy entropy over training

Figure 8: Sensitivity to the goal-aware center in the anchor term. Comparison of DA-GRPO under two choices of the anchor center used in $\delta_i$: $\hat{\mu} = (\mu_{\text{group}} + C)/2$ versus $\hat{\mu} = \max(\mu_{\text{group}}, C/2)$ (for normalized rewards, $\max(\mu_{\text{group}}, 0.5)$). Panels report (a) mean verifiable reward, (b) KL divergence to the reference policy, and (c) policy entropy, showing that the conservative piecewise center preserves stability while the upward-shifted center induces large drift and elevated entropy without improving reward.

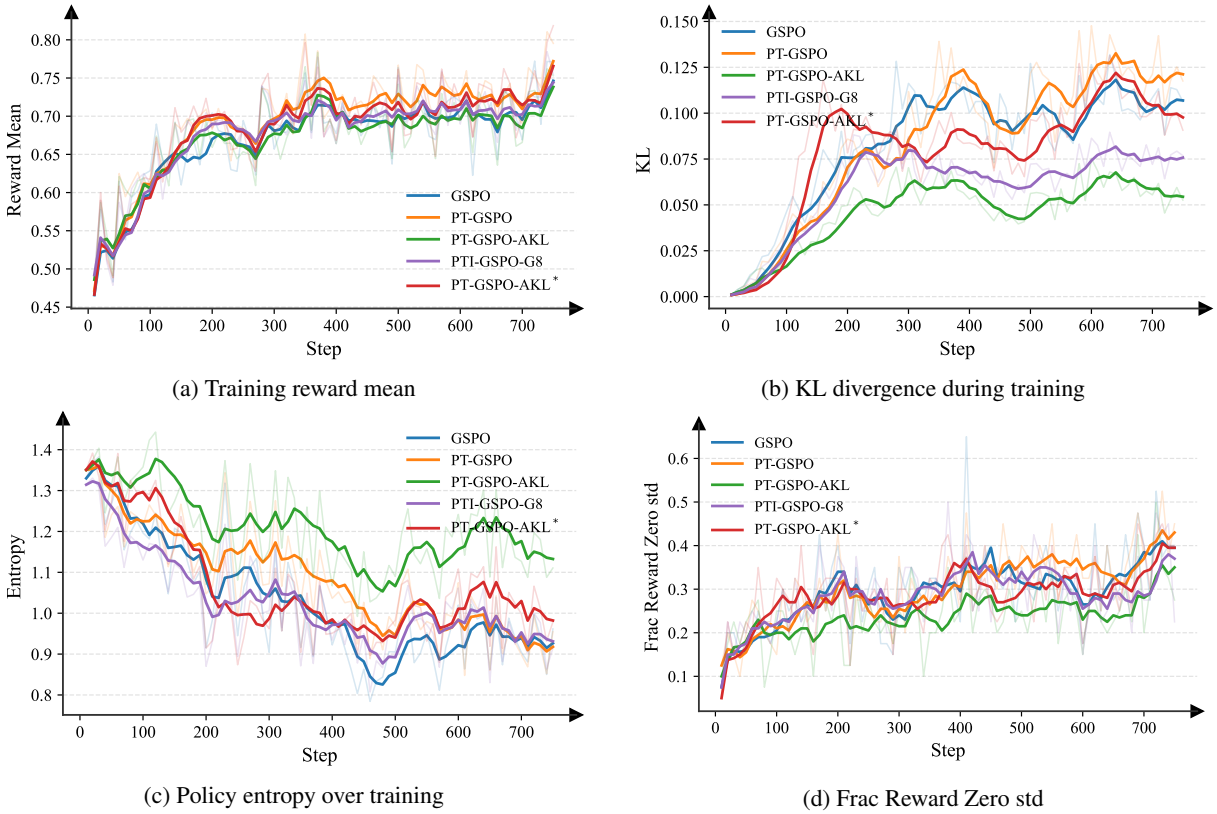
Across settings, the training reward mean is broadly comparable, with PT-GRPO matching or slightly improving over the GRPO baseline, while the *inverse* shaping (PTI-GRPO; $\lambda_+ > \lambda_-$) does not yield consistent gains, suggesting that improvements are driven by *loss aversion* rather than merely applying a bounded nonlinearity.

In terms of stability, PT-GRPO tends to increase KL relative to GRPO, indicating more aggressive policy updates when amplifying negative advantages. Introducing asymmetric KL regularization (PT-GRPO-AKL) substantially reduces KL throughout training while maintaining competitive reward, and also preserves higher policy entropy, consistent with better-controlled updates and reduced premature collapse. The more aggressive variant (PT-GRPO-AKL*) partially recovers the higher-KL behavior without commensurate reward improvements, highlighting a trade-off between correction strength and policy drift.

Finally, PT-GRPO-AKL yields the lowest fraction of steps with near-zero within-group reward variance (*Frac Reward Zero std*), indicating fewer effectively homogeneous groups and a more reliable learning signal over training. Overall, loss-averse shaping is most effective when paired with stronger KL control on negative-advantage updates, yielding a better reward–stability trade-off than either shaping alone or its inverse.

## D General Capability Retention

To test whether instruction-following gains come at the cost of general capabilities, we evaluate the models on MMLU, ARC-EASY, and ARC-CHALLENGE. Table 3 reports accuracy on MMLU under the zero-shot protocol, while Table 4 details performance on the ARC datasets. Overall, our best-performing instruction-following variants preserve (or minimally change) general capability performance. This indicates that the instruction-following improvements are not achieved at the expense of general knowledge or driven by uncontrolled policy drift.




(a) Training reward mean

(b) KL divergence during training

(c) Policy entropy over training

(d) Frac Reward Zero std

Figure 9: Ablation of prospect-theoretic shaping and asymmetric KL under GRPO. We compare GRPO with loss-averse shaping (PT-GRPO), its inverse variant (PTI-GRPO), and two asymmetric-KL versions (PT-GRPO-AKL / PT-GRPO-AKL*; see text for hyperparameters). Panels report (a) mean verifiable reward, (b) KL divergence to the reference policy, (c) policy entropy, and (d) the fraction of steps with near-zero within-group reward standard deviation (homogeneous groups), illustrating the stability–drift trade-offs induced by shaping and KL control.

| **Models** | **Gemma-2-2B-Instruct** | | | | | **Llama-3.2-3B-Instruct** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **MMLU** | | | | | **MMLU** | | | | |
| | **Overall** | **Humanities** | **Social Sci.** | **STEM** | **Other** | **Overall** | **Humanities** | **Social Sci.** | **STEM** | **Other** |
| Base | 56.9 | 50.8 | 67.3 | 48.4 | 64.3 | 60.4 | 59.2 | 67.0 | 50.3 | 65.9 |
| GRPO | 57.1 | 51.0 | 67.4 | 48.8 | 64.7 | 62.3 | 61.1 | 68.4 | 52.0 | 68.5 |
| MT-GRPO | 57.0 | 50.9 | 67.4 | 48.4 | 64.5 | 62.4 | 61.2 | 68.6 | 52.1 | 68.5 |
| DA-GRPO | 57.0 | 51.2 | 67.1 | 48.8 | 64.2 | 62.2 | 61.0 | 68.7 | 51.9 | 68.2 |
| PT-GRPO | 56.9 | 50.8 | 67.3 | 48.4 | 64.2 | 62.4 | 61.4 | 68.7 | 52.0 | 68.4 |
| DA-PT-GRPO | 56.8 | 50.8 | 67.2 | 48.2 | 64.3 | 62.1 | 60.9 | 68.7 | 51.5 | 68.1 |
| MDP-GRPO | 57.0 | 50.9 | 67.4 | 48.7 | 64.5 | 62.3 | 61.0 | 68.8 | 51.8 | 68.3 |

Table 3: MMLU performance across different model configurations. Scores are reported as percentages.

| **Models** | **Gemma-2-2B-Instruct** | | | | **Llama-3.2-3B-Instruct** | | | |
|---|---|---|---|---|---|---|---|---|
| | **ARC-Challenge** | | **ARC-Easy** | | **ARC-Challenge** | | **ARC-Easy** | |
| | **Acc** | **Acc Norm** | **Acc** | **Acc Norm** | **Acc** | **Acc Norm** | **Acc** | **Acc Norm** |
| Base | 50.9 | 52.5 | 81.0 | 78.3 | 43.6 | 45.6 | 74.2 | 68.8 |
| GRPO | 50.4 | 52.0 | 80.3 | 77.0 | 43.5 | 45.6 | 74.3 | 68.3 |
| MT-GRPO | 51.3 | 53.2 | 81.1 | 78.4 | 43.5 | 46.7 | 75.5 | 71.4 |
| DA-GRPO | 50.4 | 54.1 | 80.8 | 77.7 | 43.2 | 46.5 | 74.2 | 68.4 |
| PT-GRPO | 51.5 | 52.4 | 80.7 | 78.4 | 43.6 | 45.7 | 74.0 | 68.3 |
| DA-PT-GRPO | 51.0 | 53.2 | 81.6 | 78.4 | 43.0 | 46.5 | 75.5 | 71.2 |
| MDP-GRPO | 50.9 | 52.5 | 81.1 | 78.2 | 43.9 | 46.5 | 75.4 | 71.8 |

Table 4: ARC-Challenge and ARC-Easy performance across different model configurations. Scores are reported as percentages.