REGULATING WHAT POLICIES LEARN FROM: GEOMETRY-AWARE CALIBRATED POLICY OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Critic-free, group-based methods such as Group Relative Policy Optimization (GRPO) enable scalable LLM post-training from sampled rollouts, but typically do not account for differences in learning signal quality across rollout groups. Uncertainty provides a natural internal signal for regulating their contributions, yet existing rollout-level methods largely treat disagreement as a scalar proxy for reliability. We analyze rollout-level uncertainty from an optimization perspective and identify two key limitations through theoretical and statistical analysis: the anisotropic gap, where probability dispersion fails to capture the geometric magnitude of semantic disagreement, and the calibration gap, where uncertainty is decoupled from reward informativeness and therefore cannot distinguish noisy variation from useful reward-differentiated variation. To address these gaps, we propose **Geometric-aware Calibrated Policy Optimization (GCPO)**, which combines geometry-aware uncertainty measures with reward-based calibration to regulate policy updates according to both disagreement structure and learning informativeness. Across QA and mathematical reasoning benchmarks, GCPO delivers stronger GRPO-style optimization than entropy-based uncertainty baselines, while statistical and ablation analyses support the underlying design assumptions. Our results suggest a broader principle for uncertainty-aware policy optimization: uncertainty should characterize not only how much rollouts disagree, but also whether that disagreement provides useful learning signals.

1 INTRODUCTION

Critic-free, group-based methods such as Group Relative Policy Optimization (GRPO) (Shao et al., 2024) enable scalable LLM post-training by learning from relative rewards over sampled rollouts (Liu et al., 2025; Ma et al., 2026; Ahmadian et al., 2024). However, the standard formulation does not explicitly account for differences in learning signal quality across queries (Dong et al., 2025b;a). Consequently, groups with less reliable learning signals may receive undue weight, introducing noise into policy updates and motivating more selective regulation of their contributions. To address this issue, a promising direction is to use uncertainty to assess the reliability of model-generated outputs and regulate their contributions to learning. Unlike external supervision or learned critics, uncertainty can be derived directly from the model's own predictions, providing a scalable signal for distinguishing reliable from potentially noisy outputs. Early approaches rely on token-level entropy to quantify uncertainty in the model distribution (Fadeeva et al., 2023; 2024; Wang et al., 2024). More recent work explores *rollout-level uncertainty*, including semantic entropy (Chen et al., 2025a; Zhang et al., 2025b), predictive entropy (Duan et al., 2025), kernel-based variants Nikitin et al. (2024), and self-consistency scores (Chen et al., 2025b; Wen et al., 2025). By characterizing disagreement across sampled responses, these measures are particularly well aligned with group-based optimization such as GRPO, where learning itself is driven by variation within rollout groups.

Despite promising improvements in reasoning and training stability, existing methods largely treat uncertainty as a scalar proxy for reliability, without explicitly considering whether the measured disagreement reflects optimization-relevant variation. This limitation is twofold: an uncertainty signal must first *faithfully characterize* the disagreement that drives gradient variation, and then be *properly calibrated* to determine whether such variation is detrimental noise or informative for reward-driven learning. Through theoretical and empirical statistical analyses, to summarize, we identify two corresponding gaps in existing uncertainty methods. **1) The anisotropic gap:** existing methods captures the dispersion of probability mass but ignores the geometric magnitude of semantic disagreement,

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.

such that responses with similar entropy can induce substantially different gradient variation; And **2) The calibration gap:** even accurately measured disagreement is not necessarily undesirable; because entropy is decoupled from reward informativeness, it cannot distinguish uninformative uncertainty from reward-differentiated variation that provides useful learning signals.

To address these two gaps, we propose Geometric-aware Calibrated Policy Optimization **(GCPO)**, which regulates policy updates by jointly considering how strongly rollouts disagree and whether that disagreement carries useful reward information. To address **the anisotropic gap**, GCPO introduces geometry-aware measures, including *Cosine Dispersion (CD)* and *Barycentric Transport (BoT)*, that characterize the magnitude and structure of semantic disagreement beyond probability dispersion. Crucially, semantic disagreement alone does not determine whether the resulting variation should be suppressed until it's calibrated. To address **the calibration gap**, we therefore calibrate these geometric uncertainty estimates with *Reward Dispersion (RD)*, using within-group reward variation to distinguish uninformative disagreement from variation that exposes meaningful reward contrast. Together, these signals allow GCPO to regulate update strength according to both the structure of gradient-relevant variation and its informativeness for policy learning. We support this design through theoretical analysis connecting semantic geometry and reward dispersion to complementary aspects of optimization-relevant gradient variation, together with empirical statistical and ablation studies that provide support for these underlying assumptions. Across various QA and math reasoning benchmarks, GCPO further yields consistent improvements over GRPO and entropy-based uncertainty baselines, supporting the broader principle that uncertainty signals for policy optimization should capture not only disagreement, but also if that disagreement provides useful learning signals that preserve informative gradients while suppressing spurious variance.

- We analyze rollout-level uncertainty from an optimization perspective and identify, through theoretical and statistical analysis, two key limitations of existing entropy-based methods: ***the anisotropic gap*** and ***the calibration gap***.
- We propose **G**eometric-aware reward-**C**alibrated **P**olicy **O**ptimization **(GCPO)**, which combines geometry-aware semantic disagreement with reward-based calibration to preserve informative variation while suppressing noisy updates.
- Extensive experiments across QA and math reasoning tasks show that GCPO delivers stronger optimization than existing entropy-based uncertainty baselines, providing a practical guideline for designing uncertainty signals that better preserve useful learning signals.

2 Efficacy and Limitations of Rollout-level Uncertainty

2.1 Preliminary

GRPO. Critic-free, group-based methods such as Group Relative Policy Optimization (GRPO) have become an effective and scalable approach to LLM post-training, as they avoid learning an additional value model and instead optimize directly from relative rewards among sampled responses (Shao et al., 2024). Take GRPO for example, for each input x, GRPO samples a group of responses $Y = \{y_i\}_{i=1}^{G}$ and estimates the policy gradient using normalized rewards:

$$\hat{g}(x) = \frac{1}{G}\sum_{i=1}^{G} \hat{A}_i \, \nabla_\theta \log \pi_\theta(y_i \mid x), \quad \hat{A}_i = \frac{R(y_i) - \mu_R}{\sigma_R + \epsilon}. \tag{1}$$

Rollout-level Uncertainty Measures. Early uncertainty methods primarily rely on token-level entropy, which measures uncertainty over individual token predictions and can be sensitive to surface-form variation (Chen et al., 2025a; Nikitin et al., 2024). More recent methods such as semantic entropy (Kuhn et al., 2023) instead estimate uncertainty across sampled rollouts, capturing disagreement at the response or semantic level. We broadly refer to these approaches as *rollout-level uncertainty measures*, denoted by $U_{\mathrm{roll}}(x)$. These measures are particularly relevant to group-based optimization such as GRPO, since both uncertainty estimation and policy updates depend on variation among sampled rollouts. For instance, in semantic entropy, $H_{\mathrm{sem}}(x) = -\sum_{k=1}^{K} C_k \log C_k$, where C_k denotes the probability mass of semantic cluster k. In predictive entropy, $H_{\mathrm{pred}}(x) = -\sum_{k=1}^{K} q_k \log q_k$, where q_k represents the mass of prediction groups. We use semantic entropy as the main representative when developing the analysis below, while the the identified limitations also apply to other rollout-level uncertainty measures and token-level methods.

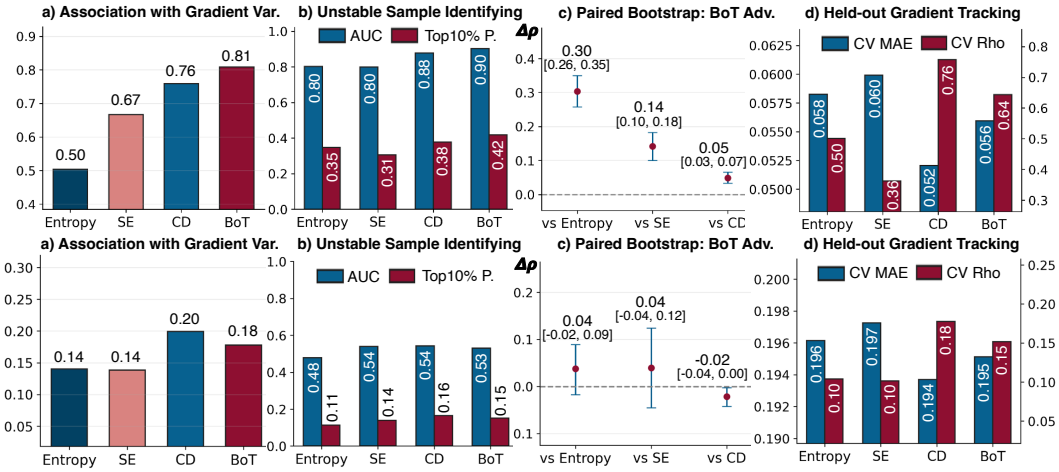

Figure 1: Statistical analysis of rollout-level uncertainty against optimization-relevant gradient variability. Top row: NarrativeQA; bottom row: Qasper. Geometry-aware measures (CD, BoT) improve alignment over existing rollout-level uncertainty measures, supporting the anisotropic gap; however, the remaining and task-dependent mismatch shows that disagreement alone is insufficient, motivating reward-based calibration. Full statistical analysis and experimental details are in Appendix D.

2.2 The Anisotropic Gap

Prior research has established that the stability and quality of policy optimization are strongly influenced by the signal-to-noise ratio of gradient estimates, making variance control important for stable and effective reinforcement learning (Cui et al., 2025; Roberts & Tedrake, 2008). To understand how rollout-level uncertainty interacts with optimization from this perspective, we first analyze the variability of the score-gradient component induced by sampled responses. Let $Y \sim \pi_\theta(\cdot \mid x)$ denote a sampled response, and let $Z = c(Y) \in \{1, \dots, K\}$ denote its semantic cluster label. Since GRPO operates on sampled rollouts, different semantic responses may induce substantially different update directions, making their geometric variation directly relevant to optimization. Let $g(Y) = \nabla_\theta \log \pi_\theta(Y \mid x)$. By the law of total covariance (Casella & Berger, 2024),

$$\mathrm{Cov}(g \mid x) = \mathbb{E}[\mathrm{Cov}(g \mid Z, x) \mid x] + \mathrm{Cov}(\mathbb{E}[g \mid Z, x] \mid x)\,. \tag{2}$$

Taking the trace yields:

$$\mathcal{V}(x) = \underbrace{\sum_{k=1}^{K} p_k(x)\,\mathrm{Tr}\big(\mathrm{Cov}(g \mid Z = k, x)\big)}_{\mathcal{V}_{\mathrm{intra}}(x)} + \underbrace{\mathrm{Tr}\big(\mathrm{Cov}(\mu_Z \mid x)\big)}_{\mathcal{V}_{\mathrm{inter}}(x)}, \tag{3}$$

where $p_k(x) := \mathbb{P}(Z = k \mid x)$ and $\mu_k := \mathbb{E}[g \mid Z = k, x]$. This decomposition separates variability within semantic modes from variability across modes. Mass-based rollout-level uncertainty measures such as H_{sem} depend only on the distribution $\{p_k(x)\}$, and therefore capture how probability mass is dispersed across modes, but not the geometric distances between their induced gradient directions.

This decomposition also reveals a key limitation of mass-based rollout-level uncertainty. Specifically, two rollout groups can have the same semantic-mode distribution $\{p_k(x)\}$, and therefore the same H_{sem}, while their induced gradient directions are separated by very different distances. In other words, probability dispersion captures *how responses are distributed*, but not *how different their optimization effects are*. **We refer to this mismatch as the anisotropic gap**. Intuitively, reasoning errors can differ substantially in their optimization effects. Under the semantic-gradient alignment assumption, a near-miss may induce an update direction close to that of a correct solution, whereas a fundamentally different reasoning path may induce a much more distant direction. A mass-based uncertainty score cannot distinguish these cases when their probability distributions are similar. This limitation can also be seen from a pairwise view of gradient variance:

$$\mathcal{V}(x) \approx \frac{1}{2} \sum_{i=1}^{K} \sum_{j=1}^{K} \pi_i \pi_j \|\mu_i - \mu_j\|^2 \;\le\; \frac{\Delta_{\max}^2}{2} H(x), \tag{4}$$

where the variance depends on both the *probability dispersion* ($\pi_i\pi_j$) and the *disagreement magnitude* ($\|\mu_i - \mu_j\|^2$). Entropy-based bounds replace the heterogeneous pairwise distances with a single worst-case constant $\Delta^2_{\max}$ (i.e., $\|\mu_i - \mu_j\|^2 \leq \Delta^2_{\max}$), effectively treating all disagreements as equally severe. As a result, they can provide only a loose surrogate for gradient variance when disagreement is strongly anisotropic. Detailed derivations are provided in Appendix B. Note that this analysis relies on the assumption that semantic geometry reflects differences in induced gradient directions; To further validate this assumption, Figure 1 (a, c) shows that geometry-aware measures exhibit stronger alignment with sample-level gradient variance than mass-based rollout-level uncertainty measures, providing statistical evidence consistent with the anisotropic gap (Appendix D).

2.3 The Calibration Gap

The anisotropic gap shows that rollout disagreement must be characterized faithfully, but accurate characterization alone is insufficient: substantial variation may represent either noise or useful learning signals. In GRPO, each response contributes through $\hat{A}(Y)g(Y)$, where $g(Y) = \nabla_\theta \log \pi_\theta(Y \mid x)$ captures its induced update direction, while $\hat{A}(Y)$ reflects relative reward contrast. Thus, semantic disagreement determines how responses differ in optimization, whereas reward contrast determines whether these differences are informative. Two rollout groups with similar uncertainty can therefore have very different learning value, which response-only uncertainty cannot distinguish. This argument implies that rollout-level uncertainty alone should be calibrated to explain reward-weighted gradient variability. Consistent with this expectation, Figure 1(b,d) shows weaker identification and held-out tracking of high-variance updates for existing rollout-level uncertainty measures (Appendix D). Therefore, when existing methods solely suppress the advantage of high uncertainty tasks (Zhang et al., 2025b; Chen et al., 2025b), they implicitly suppress the variety of responses and potentially reward variance, further suppressing the tasks which provides rich learning signals:

$$\mathbb{E}_x\Big[\mathrm{Var}\Big(\hat{A}(Y)g(Y) \mid x\Big)\Big] \downarrow \implies \mathbb{E}_x[\mathrm{Var}(\hat{g}(x) \mid x)] \downarrow . \tag{5}$$

potentially resulting in weaker updates and diminished learning signals. More broadly, this induces a biased training distribution. Let $\mathcal{D}_H := \{x : H_{\mathrm{sem}}(x) \leq \tau\}$ denote the subset retained after entropy filtering. Since high-entropy inputs are more likely to contain diverse and reward-differentiated responses, removing them reduces the overall gradient variability: $\mathbb{E}_{x\sim\mathcal{D}_H}\big[\mathrm{Var}(\hat{g}(x))\big] \ll \mathbb{E}_{x\sim\mathcal{D}}\big[\mathrm{Var}(\hat{g}(x))\big]$, thereby suppressing the primary source of corrective learning signals in reasoning tasks. **We refer to this mismatch as the calibration gap**: rollout-level uncertainty characterizes how strongly responses disagree, but not whether that disagreement should be suppressed or preserved. Effective regulation therefore requires calibrating disagreement with reward informativeness, rather than treating uncertainty alone as a measure of learning signal quality.

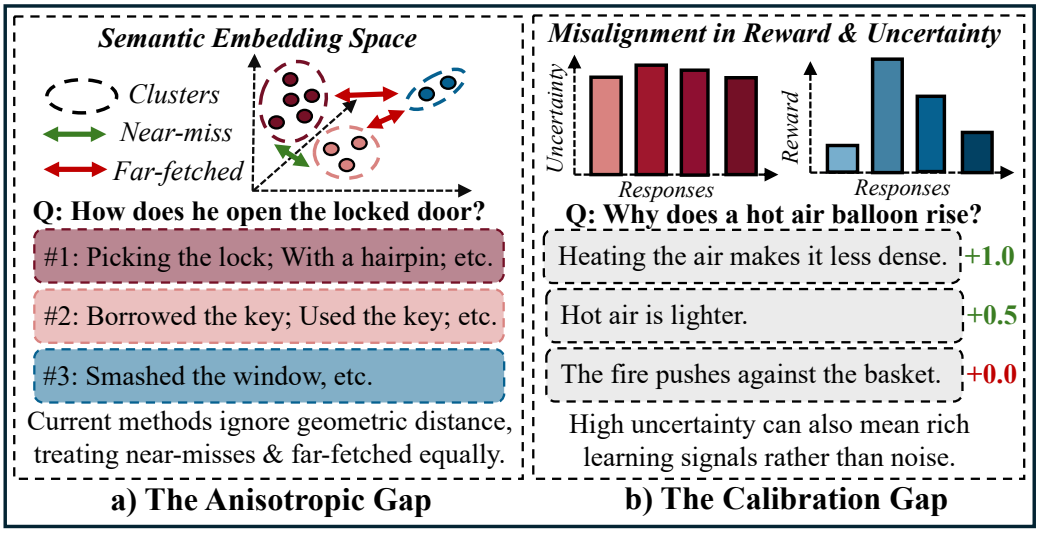

Figure 2: **The Two Core Gaps: An Intuitive View.** (a) The anisotropic gap captures missing geometry in disagreement; (b) the calibration gap captures its mismatch with learning value. Together, they reveal complementary factors for tracking reward-weighted gradient variability.

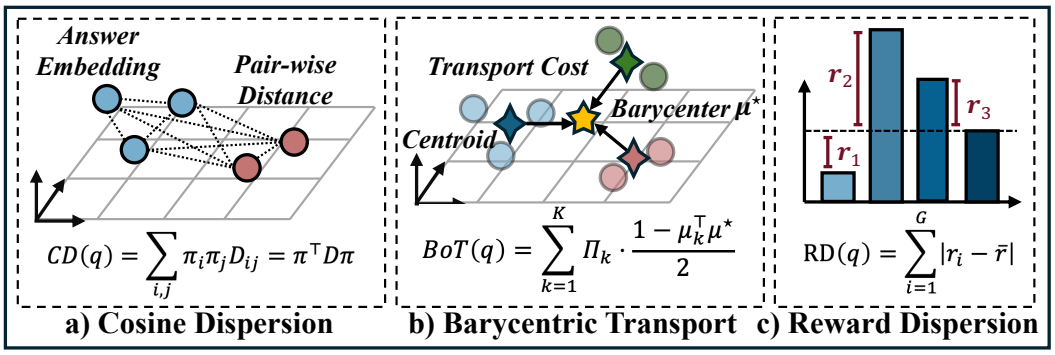

Figure 3: **Illustration of the key components of GCPO.** The formulas above highlight the core principles rather than the complete method; full details are provided in Section 3.

3 Geometric-aware Calibrated Policy Optimization

Building on the anisotropic and calibration gaps identified above, we propose Geometric-aware reward-Calibrated Policy Optimization **(GCPO)**, organized along the lifecycle of an uncertainty signal: quantification, calibration, and utilization. Specifically, for quantification, we purpose *Cosine Dispersion* (Section 3.1) and its alternative *Barycentric Transport* (Section 3.2) to address the anisotropic gap by characterizing the magnitude and structure of semantic disagreement; For calibration, we purpose *Reward Dispersion* (Section 3.3) to addresses the calibration gap by relating this disagreement to within-group reward informativeness. Finally, for utilization, we introduce how the resulting calibrated signal modulates the GRPO update. These stages are coupled: faithful quantification determines what variation is represented, proper calibration determines what that variation means for learning, and effective utilization determines how the calibrated signal ultimately shapes the policy update. Figure 3 illustrates our key components. Throughout this section, q denotes the input query, corresponding to x in Section 2, and G denotes the number of sampled rollouts.

3.1 Quantification: Cosine Dispersion for Geometry-aware Disagreement

As discussed in Section 2.2, mass-based rollout-level uncertainty measures such as semantic entropy depend primarily on probability mass and therefore do not capture the *magnitude* of disagreement between responses. To address this anisotropic gap, we first introduce *Cosine Dispersion (CD)*, which measures semantic disagreement directly in embedding space. Specifically, given the sampled rollouts $\{y_i\}_{i=1}^G$, let s_i denote the text representation of rollout i used for uncertainty estimation, and let $f_{\text{emb}}(\cdot)$ denote the semantic embedding function. We compute the normalized embedding $\mu_i = f_{\text{emb}}(s_i)/\|f_{\text{emb}}(s_i)\|_2$. The pairwise cosine distance is defined as $D_{ij} = \text{clip}(1-\mu_i^\top \mu_j, 0, 1)$, where $\text{clip}(v, l, u) = \min(\max(v, l), u)$. With uniform rollout weights $\pi_i = 1/G$, let $\pi = (\pi_1, \ldots, \pi_G)^\top$ and $D = [D_{ij}]_{i,j=1}^G$. CD is formally defined as

$$\text{CD}(q) = \sum_{i,j} \pi_i \pi_j D_{ij} = \pi^\top D \pi. \tag{6}$$

Unlike mass-based uncertainty, CD reflects not only whether responses differ, but also *how far apart* they are in semantic space. When responses are aligned, CD remains small; when disagreement becomes geometrically significant, CD increases accordingly. We use a group-size-adjusted modulation coefficient $\alpha_G = \alpha_{\text{base}}/\log G$, where α_{base} is tunable, and translates CD into a reliability weight:

$$\omega_{\text{CD}}(q) = \text{clip}\left(1 - \alpha_G \,\text{CD}(q)^2, 0, 1\right). \tag{7}$$

This formulation provides a direct geometric correction to mass-based uncertainty by penalizing large semantic deviations while remaining tolerant to minor variations. CD captures fine-grained pairwise disagreement and is designed for settings where semantic variation is smooth and continuous.

3.2 Quantification: Barycentric Transport for Semantic Structure

As an alternative, while CD captures pairwise geometric disagreement, it treats all samples independently and does not explicitly account for how probability mass is distributed across distinct semantic modes. This limitation becomes pronounced when rollouts form multiple coherent semantic clusters, where within-cluster variation may be benign while across-cluster divergence is

substantial. To capture this structure, we draw inspiration from *optimal transport* (OT), and in particular the idea of representing multiple distributions relative to a shared barycentric consensus. Concretely, given distributions $\{\nu_k\}_{k=1}^K$ with weights $\{\Pi_k\}_{k=1}^K$, a Wasserstein barycenter is defined as $\nu^\star = \arg\min_\nu \sum_{k=1}^K \Pi_k \mathcal{W}_c(\nu_k, \nu)$, where $\mathcal{W}_c(\cdot,\cdot)$ denotes the optimal transport cost under ground metric $c(\cdot,\cdot)$. We do not solve this OT problem directly. Instead, we use this barycentric perspective to construct a tractable geometry-aware surrogate, ***Barycentric Transport (BoT)***, which measures the deviation of semantic modes from a shared consensus direction in embedding space. Specifically, we group the G rollout representations into K semantic clusters using an entailment-based procedure. Let $z_i \in \{1,\dots,K\}$ denote the cluster assignment of rollout i. The empirical probability mass of cluster k is $\Pi_k = \sum_{i:z_i=k} \pi_i$, where $\pi_i = 1/G$, and μ_k denotes the normalized embedding centroid of cluster k. We define the shared barycentric direction as

$$\mu^\star = \frac{\sum_{k=1}^K \Pi_k \mu_k}{\left\| \sum_{k=1}^K \Pi_k \mu_k \right\|_2}. \tag{8}$$

We then define the Barycentric Transport score as

$$\mathrm{BoT}(q) = \sum_{k=1}^K \Pi_k \cdot \frac{1 - \mu_k^\top \mu^\star}{2}. \tag{9}$$

Here, $(1 - \mu_k^\top \mu^\star)/2$ measures the cosine-based deviation between semantic mode k and the shared barycentric direction. BoT therefore serves as a geometry-aware proxy for the dispersion of semantic modes around a shared consensus. We further discuss the OT interpretation, approximation, and associated limitations of BoT in Appendix F.6. Unlike CD, BoT collapses intra-cluster variation and focuses on inter-mode structure, making it less sensitive to paraphrastic redundancy while remaining sensitive to semantic divergence (Detailed discussion in Appendix F.1). The corresponding modulation weight is

$$\omega_{\mathrm{BoT}}(q) = \mathrm{clip}\left(1 - \alpha_G \,\mathrm{BoT}(q)^2, 0, 1\right). \tag{10}$$

CD and BoT therefore represent two alternative instantiations of geometry-aware uncertainty quantification. CD captures fine-grained pairwise disagreement, while BoT captures structured dispersion across semantic modes. We further discuss their theoretical connections with gradient variance in Appendix C. This connection relies on a semantic-gradient alignment assumption; we discuss its implications and provide empirical evidence supporting this assumption in Appendix F.4.

3.3 Calibration: Reward Dispersion via Informative Signal

While geometry-aware measures address the anisotropic gap, they do not resolve the calibration gap: semantic disagreement alone does not indicate whether a rollout group provides useful learning signals. As discussed in Section 2, GRPO updates are driven by relative rewards, suggesting that geometric disagreement should be calibrated with respect to reward variation.

Let $r_i = R(y_i)$ denote the reward assigned to rollout y_i, and let $\bar{r} = \frac{1}{G}\sum_{i=1}^G r_i$ denote the mean reward within the rollout group. We define ***Reward Dispersion (RD)*** as a proxy for reward informativeness:

$$\mathrm{RD}_{\mathrm{raw}}(q) = \sum_{i=1}^G |r_i - \bar{r}|, \quad \mathrm{RD}(q) = \mathrm{clip}\left(\frac{\mathrm{RD}_{\mathrm{raw}}(q)}{\mathrm{RD}_{\max}(G)}, 0, 1\right), \quad \omega_{\mathrm{RD}}(q) = 1 + \alpha_G \,\mathrm{RD}(q). \tag{11}$$

Here, $\mathrm{RD}_{\max}(G)$ denotes the maximum achievable raw reward dispersion for a group of size G when rewards are bounded within the configured range $[r_{\min}, r_{\max}]$:

$$\mathrm{RD}_{\max}(G) = \frac{2}{G}\left\lfloor \frac{G}{2} \right\rfloor \left\lceil \frac{G}{2} \right\rceil (r_{\max} - r_{\min}). \tag{12}$$

In contrast to the geometric terms, RD amplifies updates when reward contrast is strong. The intuition is that rollout groups with greater reward differentiation can provide more informative relative learning signals, whereas uniformly scored rollouts provide limited reward contrast. In this way, RD calibrates semantic disagreement using reward information relevant to optimization. We discuss the interaction between uncertainty and reward dispersion, including safeguards against noisy reward variation, in Appendix F.1.

3.4 UTILIZATION: UNIFIED MODULATION OF POLICY UPDATES

After quantification and calibration, the remaining question is how the resulting signals should affect learning. We incorporate them as query-level modifiers of the normalized GRPO advantage:

$$\widetilde{A}_i = \hat{A}_i \, \omega_{\mathrm{geo}}(q) \, \omega_{\mathrm{RD}}(q), \qquad \hat{g}_{\mathrm{GCPO}}(q) = \frac{1}{G} \sum_{i=1}^{G} \widetilde{A}_i \nabla_\theta \log \pi_\theta(y_i \mid q), \tag{13}$$

where ω_{geo} is instantiated as either ω_{CD} or ω_{BoT}. Viewed jointly, the weighting functions defined above share a common neutral point at 1, recovering the original GRPO update when both signals are neutral. The geometric term moves below 1 as disagreement grows, with the squared form remaining conservative for small deviations while becoming more responsive to substantial ones; reward dispersion moves above 1 as within-group learning contrast increases. Their multiplicative composition therefore combines two complementary considerations within a single query-level factor, while preserving the relative advantages among rollouts in the group. The shared coefficient $\alpha_G = \alpha_{\mathrm{base}} / \log G$ controls the strength of these deviations from the neutral update. Further discussion of these weighting choices and their interaction is provided in Appendix F.1. Because this modulation changes the relative contribution of rollout groups, it also changes the effective weighting of the original GRPO objective; we discuss this effect in Appendix F.3. Its relation to objective-level KL-constrained optimization and its computational overhead are discussed in Appendices F.2 and F.5.

4 EXPERIMENTS

4.1 EXPERIMENT SETUP

Benchmarks. We evaluate GCPO on a diverse set of benchmarks spanning open-ended QA, fact-driven QA, and mathematical reasoning. We first focus on open-ended question answering datasets, including NarrativeQA Kočiskỳ et al. (2018) and Qasper Dasigi et al. (2021), where multiple semantically distinct yet partially correct responses naturally arise. Such settings induce high response-level variability, making them particularly suitable for studying uncertainty signals and their interaction with optimization dynamics. We report BLEU, token-level F1, and accuracy (F1 > 0.5) for evaluation. We further include fact-driven QA via HotpotQA Yang et al. (2018), which emphasizes multi-hop reasoning with more constrained answer spaces. Evaluation follows standard token-level F1 and Exact Match (EM). Finally, We additionally evaluate on mathematical reasoning benchmarks, including MATH500 Hendrycks et al. (2021), AIME24, and OlympiadBench Huang et al. (2024), to examine generalization to structured reasoning tasks with deterministic answers. Detailed dataset descriptions and configurations are provided in Appendix E.1.

Baselines. We focus on baselines directly aligned with our scope, including: (i) **GRPO**, without uncertainty modulation (Shao et al., 2024); (ii) **semantic entropy-based methods**, such as Seed-GRPO (Chen et al., 2025a) and EMPO (Zhang et al., 2025b), which regulate updates using response-level entropy; (iii) **kernel-based entropy**, which refines uncertainty via similarity-aware measures (Nikitin et al., 2024); (iv) **consistency-based methods**, e.g., GRPO-CARE (Chen et al., 2025b), which use agreement across rollouts as a proxy for uncertainty; and finally (v) **most recent baselines** for a broader comparison (Xie et al., 2026; Hu et al., 2026; Luo et al., 2026). All methods are evaluated under a unified protocol with the same backbone and configuration on 5 consecutive seeds. Implementation details and adaptation procedures are provided in Appendix E.2.

4.2 MAIN EXPERIMENT

We present the main results in Table 1. Overall, GCPO shows strong and consistent gains over GRPO and remains competitive with both classical uncertainty-based methods and recent policy-optimization baselines. In particular, GCPO-BoT achieves the best results on all NarrativeQA metrics, Qasper F1, and both HotpotQA metrics, improving over GRPO by 10.15%–27.55%. The gains are also substantial over recent methods such as Q-Hawkeye, EG-SPO, and R2VPO: GCPO achieves the best performance on six of eight reported metrics.

The results further support our central motivation that rollout-level uncertainty is useful only when it captures optimization-relevant variation. Classical uncertainty methods generally improve over

Method	NarrativeQA			Qasper			HotpotQA	
	Acc	BLEU	F1	Acc	BLEU	F1	F1	EM
GRPO	66.23±1.53	48.71±1.70	67.02±0.95	11.59±1.63	9.40±0.95	20.51±1.36	50.00±2.92	37.06±2.69
Classical Uncertainty Methods								
Seed-GRPO	71.76±0.47	52.20±1.80	72.24±0.35	13.29±0.40	10.80±1.37	22.23±0.47	57.33±1.29	44.18±1.13
KLE	72.03±0.50	52.11±1.41	72.39±0.30	13.46±0.63	11.62±0.35	22.26±0.30	57.62±0.90	44.31±0.83
EMPO	57.38±0.25	41.32±0.28	60.96±0.20	12.00±0.27	9.58±0.17	18.05±0.23	n/a	n/a
GRPO-Care	71.40±0.60	51.84±0.65	71.65±0.48	13.32±0.75	10.83±0.30	22.07±0.46	57.03±1.99	43.75±2.14
Most Recent SOTA Methods								
Q-Hawkeye	71.13±0.40	52.44±0.40	71.63±0.26	12.98±1.08	10.94±0.44	23.50±1.01	58.61±1.49	45.10±1.81
EG-SPO	71.51±1.12	52.16±0.88	71.84±0.78	**14.29±1.22**	**12.40±1.03**	23.91±0.83	57.75±0.89	44.28±0.90
R2VPO	72.32±0.24	52.12±1.18	72.52±0.23	13.02±0.28	9.67±0.72	23.45±0.46	57.60±1.46	44.46±1.28
Our Methods								
GRPO + CD	72.59±0.48	53.72±0.89	73.25±0.27	13.67±0.33	11.28±0.83	24.25±0.17	59.97±0.67	46.80±0.77
Δ *vs. GRPO*	↑9.60%	↑10.29%	↑9.30%	↑17.95%	↑20.00%	↑18.24%	↑19.94%	↑26.28%
GRPO + BoT	**73.77±0.16**	**54.85±0.74**	**73.82±0.07**	13.91±0.63	10.94±1.49	**24.37±0.64**	**60.78±1.06**	**47.27±1.08**
Δ *vs. GRPO*	↑11.38%	↑12.61%	↑10.15%	↑20.02%	↑16.38%	↑18.82%	↑21.56%	↑27.55%

Table 1: Performance comparison on NarrativeQA, Qasper, and HotpotQA. We report mean and standard deviation over five runs. Best results are bolded and second-best underlined.

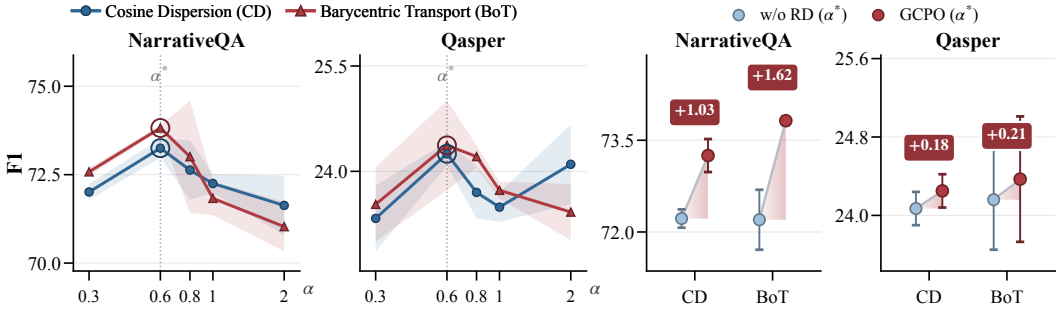

Figure 4: **Ablations of modulation strength and reward calibration. Left:** F1 on NarrativeQA and Qasper as α varies (logarithmic axis; mean $\pm$ sd over runs), with the operating point $\alpha^* = 0.6$ circled. **Right:** removing the RD term (w/o RD) versus full GCPO, both at α^*, for CD and BoT.

GRPO, but the gains are uneven across methods and tasks, indicating that uncertainty regulation alone is insufficient. GCPO instead combines geometry-aware quantification with reward-based calibration, yielding particularly clear gains on NarrativeQA and HotpotQA and the strongest Qasper F1. This pattern is consistent with our earlier analysis: the benefit of uncertainty depends not only on the amount of response disagreement, but on whether that disagreement corresponds to informative learning variation. Finally, GCPO also exhibits relatively low variance across runs on most metrics, suggesting that the proposed modulation improves optimization stability while preserving useful variation. Together, these results support viewing uncertainty as a mechanism for regulating the quality of rollout-level learning signals, rather than simply suppressing or amplifying dispersion.

4.3 Understanding GCPO: Modulation Strength and Reward Calibration

We further examine how the proposed modulation behaves through the ablations in Figure 4. Varying α reveals a clear intermediate operating regime: performance first improves as uncertainty begins to meaningfully modulate the GRPO update, peaks around $\alpha^* = 0.6$, and then declines when the modulation becomes too strong. This behavior is consistent across CD and BoT and on both datasets. Importantly, this is more than a hyperparameter effect: under the weighting form $1 - \alpha_G U^2_{\text{geo}}$, small geometric deviations are intentionally treated conservatively, while larger disagreement receives progressively stronger attenuation. The observed optimum therefore reflects a balance between leaving informative variation largely intact and suppressing rollout groups whose disagreement becomes increasingly unreliable; excessive α distorts this balance by allowing the uncertainty signal to dominate the underlying GRPO update. Having established how strongly the geometric signal

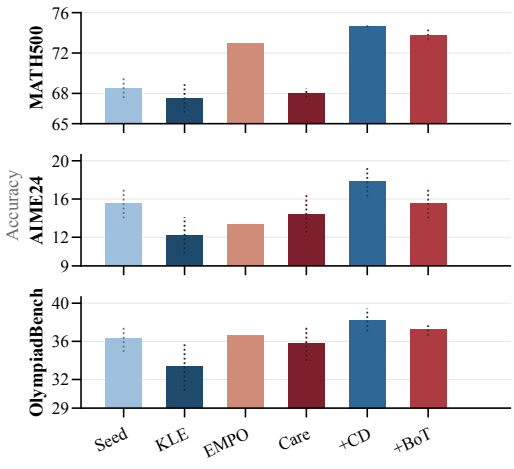

Figure 5: Accuracy on MATH500, AIME24, and OlympiadBench with Qwen2.5-MATH-1.5B (mean ± sd; EMPO is reported without sd). Y-axes are per panel and start above zero.

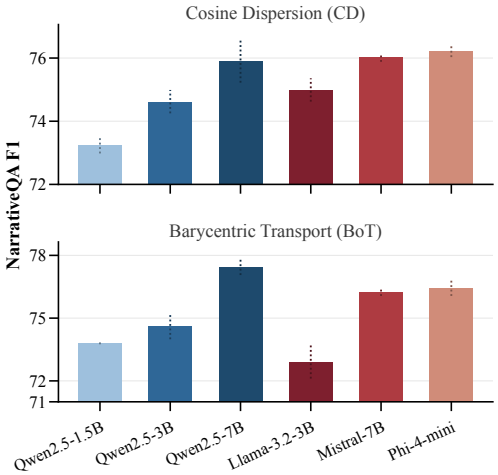

Figure 6: NarrativeQA F1 of CD (top) and BoT (bottom) across six instruction-tuned backbones (mean ± sd over three runs). Y-axes are per panel and start above zero.

should modulate the GRPO update, we next ask whether geometry alone is sufficient. As Figure 4 shows, removing RD consistently degrades both CD and BoT, with F1 drops of 1.03 and 1.62 on NarrativeQA and 0.18 and 0.21 on Qasper, respectively. Geometry alone can identify the structure and magnitude of disagreement, but cannot determine whether that disagreement carries useful reward contrast. RD supplies this missing calibration, recovering substantially larger gains on NarrativeQA, where reward differentiation is more informative, and smaller but consistent gains on Qasper. Together, these results support the coupled design of GCPO: geometry controls which variation should be attenuated, while RD determines when that variation remains valuable for learning.

4.4 Generalization Across Task and Representation Geometry

Beyond the main QA setting, we further examine GCPO under two complementary changes: a different task geometry and a different model-induced representation space. Figure 5 evaluates mathematical reasoning, where uncertainty is computed over full reasoning trajectories rather than short-form answers. GCPO transfers strongly to this setting: CD achieves the best performance on all three benchmarks, while BoT also remains competitive with recent uncertainty-aware baselines. This extension is non-trivial because mathematical solutions may follow semantically different valid paths, and correctness can hinge on a small intermediate deviation. The results therefore suggest that geometry-aware uncertainty can remain useful when the object being compared changes from answers to structured reasoning trajectories. Figure 6 complements this analysis by changing the representation space itself. Across six Qwen, Llama, Mistral, and Phi backbones, both CD and BoT maintain strong NarrativeQA performance, indicating that GCPO is not tied to a particular model family. Within the Qwen-2.5 family, performance generally improves with model size, while the cross-family results reveal a more informative distinction: CD varies relatively smoothly across backbones, whereas BoT is more sensitive to the underlying model, most visibly on Llama-3.2-3B. Together with Figure 5, this suggests that pairwise geometric dispersion provides a more stable uncertainty signal across changing task and representation geometries, while mode-level structure can offer stronger gains when the induced semantic clusters are well formed.

5 Conclusion

In this paper, we study rollout-level uncertainty for group-based RL methods from an optimization perspective and show that effective regulation requires both faithful quantification of disagreement and proper calibration to its learning value. Our analysis reveals the anisotropic and calibration gaps, motivating GCPO, which couples geometry-aware uncertainty with reward-based calibration and utilizes the resulting signal to modulate policy updates. Across QA and mathematical reasoning benchmarks, GCPO consistently improves GRPO-style optimization, while statistical, ablation, and generalization analyses support the underlying design principles. More broadly, our results suggest that uncertainty for policy optimization should characterize not only how much rollouts disagree, but also whether that disagreement provides useful learning signals.

AI USE STATEMENT

In this work, we used generative AI tools to assist with experiment design and feedback; support implementation of the proposed method; clean and reformat datasets; assist with qualitative and thematic analysis; Formulate mathematical claims; assist with proof writing and support interpretation of experimental results. We did not use generative AI tools to develop theoretical models or conceptual frameworks, propose or refine hypotheses, design the overall research agenda, assist with translation, or provide key ingredients for mathematical proofs. Additionally, we used generative AI tools to create research artifacts, draft portions of the manuscript, summarize and analyze existing literature, search for and identify relevant information and references, improve the readability of the manuscript, and format references. All AI-assisted outputs were reviewed and verified by the authors. LLM-generated code was manually inspected and tested, synthetic data and experimental outputs were checked for correctness and consistency, literature-related suggestions were verified against the original sources, and all AI-assisted writing and analysis were reviewed and revised by the authors. We take full responsibility for the final content of this work, including all text, claims, code, data, analyses, and artifacts produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

We provide detailed materials to support reproducibility of both the theoretical and empirical results. The assumptions and complete derivations underlying the anisotropic-gap analysis and the geometry-gradient connection are given in Appendices B and C, with additional discussion of their scope and limitations in Appendix F.4. Appendix D describes the statistical analyses used to examine these assumptions empirically. Dataset descriptions, baseline adaptations, training configurations, reward definitions, semantic embedding and clustering procedures, and the end-to-end GCPO pipeline are documented in Appendix E, together with representative prompt templates used for QA and mathematical reasoning. Additional discussion of the weighting design, query-level reweighting, computational overhead, and the interpretation of BoT is provided in Appendix F. We further provide the experimental configurations and implementation materials in the supplementary material to facilitate reproduction of the reported results.

REFERENCES

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, 2024.

George Casella and Roger Berger. *Statistical inference*. Chapman and Hall/CRC, 2024.

Minghan Chen, Guikun Chen, Wenguan Wang, and Yi Yang. Seed-grpo: Semantic entropy enhanced grpo for uncertainty-aware policy optimization. *arXiv preprint arXiv:2505.12346*, 2025a.

Yi Chen, Yuying Ge, Rui Wang, Yixiao Ge, Junhao Cheng, Ying Shan, and Xihui Liu. Grpo-care: Consistency-aware reinforcement learning for multimodal reasoning. *arXiv preprint arXiv:2506.16141*, 2025b.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, et al. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4599–4610, 2021.

Guanting Dong, Licheng Bao, Zhongyuan Wang, Kangzhi Zhao, Xiaoxi Li, Jiajie Jin, Jinghan Yang, Hangyu Mao, Fuzheng Zhang, Kun Gai, et al. Agentic entropy-balanced policy optimization. *arXiv preprint arXiv:2510.14545*, 2025a.

Guanting Dong, Hangyu Mao, Kai Ma, Licheng Bao, Yifei Chen, Zhongyuan Wang, Zhongxia Chen, Jiazhen Du, Huiyang Wang, Fuzheng Zhang, et al. Agentic reinforced policy optimization. *arXiv preprint arXiv:2507.19849*, 2025b.

Jinhao Duan, James Diffenderfer, Sandeep Madireddy, Tianlong Chen, Bhavya Kailkhura, and Kaidi Xu. Uprop: Investigating the uncertainty propagation of llms in multi-step agentic decision-making. *arXiv preprint arXiv:2506.17419*, 2025.

Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, et al. Lm-polygraph: Uncertainty estimation for language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 446–461, 2023.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, et al. Fact-checking the output of large language models via token-level uncertainty quantification. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 9367–9385, 2024.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Yuelin Hu, Zhengxue Cheng, Wei Liu, and Li Song. Entropy-gated selective policy optimization: Token-level gradient allocation for hybrid training of large language models. *arXiv preprint arXiv:2602.03309*, 2026.

Zhen Huang, Zengzhi Wang, Shijie Xia, Xuefeng Li, Haoyang Zou, Ruijie Xu, Run-Ze Fan, Lyumanshan Ye, Ethan Chern, Yixin Ye, et al. Olympicarena: Benchmarking multi-discipline cognitive reasoning for superintelligent ai. *Advances in Neural Information Processing Systems*, 37:19209–19253, 2024.

Mingjian Jiang, Yangjun Ruan, Prasanna Sattigeri, Salim Roukos, and Tatsunori B. Hashimoto. Graph-based uncertainty metrics for long-form language model outputs. *ArXiv*, abs/2410.20783, 2024. URL https://api.semanticscholar.org/CorpusID:273654396.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328, 2018.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *ICLR*, 2023.

Zhenwen Liang, Sidi Lu, Wenhao Yu, Kishan Panaganti, Yujun Zhou, Haitao Mi, and Dong Yu. Can llms guide their own exploration? gradient-guided reinforcement learning for llm reasoning. *arXiv preprint arXiv:2512.15687*, 2025.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.

Yu Luo, Shuo Han, Yihan Hu, Dong Li, and Jianye Hao. Ratio-variance regularized policy optimization for efficient llm fine-tuning. *arXiv preprint arXiv:2601.03320*, 2026.

Tianyi Ma, Yiyue Qian, Yiyang Li, Zehong Wang, Yifang Ding, Zheyuan Zhang, Yan Liang, Chuxu Zhang, and Yanfang Ye. Non-monotonic autoregressive sequence model. In *ICML*, 2026.

Potsawee Manakul, Adian Liusie, and Mark John Francis Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *ArXiv*, abs/2303.08896, 2023. URL https://api.semanticscholar.org/CorpusID:257557820.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities. *Advances in Neural Information Processing Systems*, 37:8901–8929, 2024.

Prasanna Parthasarathi, Mathieu Reymond, Boxing Chen, Yufei Cui, and Sarath Chandar. Grpo-lambda: Credit assignment improves llm reasoning. *arXiv preprint arXiv:2510.00194*, 2025.

Wenwen Qiang, Ziyin Gu, Jiahuan Zhou, Jie Hu, Jingyao Wang, Changwen Zheng, and Hui Xiong. On the plasticity and stability for post-training large language models. *arXiv preprint arXiv:2602.06453*, 2026.

John Roberts and Russ Tedrake. Signal-to-noise ratio analysis of policy gradient algorithms. *NeurIPS*, 2008.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Zehong Wang, Zheyuan Zhang, Nitesh V Chawla, Chuxu Zhang, and Yanfang Ye. Gft: Graph foundation model with transferable tree vocabulary. *Advances in neural information processing systems*, 37:107403–107443, 2024.

Kangda Wei and Ruihong Huang. Mmr-grpo: Accelerating grpo-style training through diversity-aware reward reweighting. *arXiv preprint arXiv:2601.09085*, 2026.

Zhihua Wen, Zhizhao Liu, Zhiliang Tian, Shilong Pan, Zhen Huang, Dongsheng Li, and Minlie Huang. Scenario-independent uncertainty estimation for llm-based question answering via factor analysis. In *Proceedings of the ACM on Web Conference 2025*, pp. 2378–2390, 2025.

Wulin Xie, Rui Dai, Ruidong Ding, Kaikui Liu, Xiangxiang Chu, Xinwen Hou, and Jie Wen. Q-hawkeye: Reliable visual policy optimization for image quality assessment. *arXiv preprint arXiv:2601.22920*, 2026.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*, 2023.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *EMNLP*, 2018.

Caiqi Zhang, Fangyu Liu, Marco Basaldella, and Nigel Collier. Luq: Long-text uncertainty quantification for llms. *ArXiv*, abs/2403.20279, 2024. URL `https://api.semanticscholar.org/CorpusID:268793903`.

Kaichen Zhang, Yuzhong Hong, Junwei Bao, Hongfei Jiang, Yang Song, Dingqian Hong, and Hui Xiong. Gvpo: Group variance policy optimization for large language model post-training. *arXiv preprint arXiv:2504.19599*, 2025a.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Peilin Zhao, and Yatao Bian. Right question is already half the answer: Fully unsupervised llm reasoning incentivization. *arXiv preprint arXiv:2504.05812*, 2025b.

Wenzheng Zhang and Karl Stratos. Rence: Learning to reason by noise contrastive estimation. *arXiv preprint arXiv:2601.22432*, 2026.

APPENDIX CONTENT TABLE

A Additional Related Work **14**

A.1 Uncertainty Estimation for Generation and Reasoning. 14

A.2 Variance Control in Policy Optimization (Complementary Perspectives). 14

B Proof: Relations Between Entropy and Variance **14**

B.1 Derivation of the Variance Upper Bound 15

B.2 Relation Between Gini Impurity and Shannon Entropy 15

B.3 Detailed Analysis of the Anisotropic Gap 15

C Link Between Geometry-aware Uncertainty and Gradient Variance **16**

D Statistical Analysis of Uncertainty Measures **16**

D.1 Data and Experimental Setup 16

D.2 Target Quantities 17

D.3 Uncertainty Measures 17

D.4 Association with Score-Gradient Variability 17

D.5 Identification of High-Update-Variability Queries 17

D.6 Statistical Robustness via Paired Bootstrap 18

D.7 Held-out Update-Variability Prediction 18

D.8 Discussion: Dataset-dependent Effects 18

D.9 Final Takeaway 18

E Implementation Details **18**

E.1 Benchmarks 18

E.2 Baselines 19

E.3 Training Details 19

F Discussion **20**

F.1 Interaction Between Geometric Uncertainty and Reward Dispersion 20

F.2 Relation to KL-constrained Optimization 21

F.3 Effect of Query-level Reweighting 21

F.4 On the Semantic-Gradient Alignment Assumption 22

F.5 Computational Overhead Analysis 22

F.6 On the OT Interpretation of BoT 22

A Additional Related Work

A.1 Uncertainty Estimation for Generation and Reasoning.

A primary line of work estimates uncertainty directly from model predictions, where token-level entropy is used as a proxy for confidence. These methods interpret uncertainty as dispersion over next-token distributions and have been widely used in early studies Fadeeva et al. (2023; 2024). More recent approaches incorporate such signals into training, for example by using entropy to modulate gradient magnitude while preserving update direction Hu et al. (2026). While effective for capturing local prediction uncertainty, token-level measures do not directly characterize higher-level semantic disagreement across complete responses, which motivates complementary rollout-level uncertainty measures for reasoning-intensive tasks.

To address this limitation, a growing body of work shifts to *rollout-level uncertainty*, estimating confidence from agreement across multiple generated samples. Early approaches quantify agreement through exact-match frequency or consistency scores Xiong et al. (2023), while later work extends this idea to semantic and factual consistency using external evaluators or LLM-based judgments Manakul et al. (2023); Zhang et al. (2024); Jiang et al. (2024). Entropy-based formulations provide a more principled probabilistic interpretation by modeling distributions over semantic outcomes, including clustering-based semantic entropy Kuhn et al. (2023), its discrete variant Farquhar et al. (2024), and similarity-based extensions such as kernel entropy Nikitin et al. (2024) and nearest-neighbor entropy (Semantic Nearest Neighbor Entropy).

However, as discussed in Section 2, these methods exhibit two key limitations. First, they characterize the distribution of responses but ignore the magnitude and structure of disagreement, treating heterogeneous semantic differences uniformly (the anisotropic gap). Second, they are largely decoupled from reward signals, failing to distinguish between informative diversity and spurious variability (the calibration gap). These limitations motivate our geometry-aware and reward-calibrated formulation.

A.2 Variance Control in Policy Optimization (Complementary Perspectives).

Beyond uncertainty estimation, a separate line of work studies stability in policy optimization through the lens of variance control. These approaches operate at different levels of the optimization pipeline and are largely orthogonal to our focus. At the objective level, KL-constrained and divergence-based methods derive stable updates through analytical weighting or regularization Zhang et al. (2025a). Other methods address gradient instability via conflict resolution or reformulation, including probabilistic gradient arbitration Qiang et al. (2026) and contrastive group optimization Zhang & Stratos (2026). At finer granularity, sample-level approaches reweight updates using uncertainty or reward variability Xie et al. (2026). Additional work targets complementary sources of variance, such as policy-ratio instability Luo et al. (2026), diversity-driven reward shaping Wei & Huang (2026), gradient-geometry-based exploration Liang et al. (2025), or step-level credit assignment Parthasarathi et al. (2025).

It is important to notice that while these methods provide principled mechanisms for improving optimization stability, and both can be phrased as *uncertainty-aware*, they focus on fundamentally different aspects of policy optimization. These methods focus on policy-level or local variance sources and do not explicitly model how semantic structure and reward differentiation jointly influence gradient variance at the response level. In contrast, our approach targets this intermediate level by treating uncertainty as a structured signal that aligns semantic disagreement with reward informativeness to calibrate gradient contributions across rollout groups.

B Proof: Relations Between Entropy and Variance

In Section 2.2, we showed that mass-based uncertainty can provide a worst-case upper bound on the inter-mode component of score-gradient variability while discarding the heterogeneous geometry between semantic modes. Here, we provide the corresponding derivation. Let $\mathbf{p} = \{p_1, \ldots, p_K\}$ denote the probability distribution over K semantic modes, where $\sum_{k=1}^{K} p_k = 1$ and $p_k \geq 0$, and let $\mu_k^g = \mathbb{E}[g(Y) \mid Z = k, x]$ denote the conditional mean score-gradient of mode k.

B.1 Derivation of the Variance Upper Bound

By the law of total covariance, the score-gradient variability decomposes into within-mode and inter-mode components as discussed in Section 2.2. The inter-mode component admits the exact pairwise form

$$V_{\text{inter}}(x) = \frac{1}{2}\sum_{i=1}^{K}\sum_{j=1}^{K} p_i p_j \left\|\mu_i^g - \mu_j^g\right\|_2^2. \tag{14}$$

When within-mode score-gradient variation is small, $V(x) \approx V_{\text{inter}}(x)$, yielding the approximation used in Eq. 4.

Let $\Delta_{\max}^2 = \max_{i,j} \|\mu_i^g - \mu_j^g\|_2^2$. Since the diagonal terms vanish,

$$V_{\text{inter}}(x) \le \frac{\Delta_{\max}^2}{2}\sum_{i\neq j} p_i p_j \tag{15}$$

$$= \frac{\Delta_{\max}^2}{2}\left(1 - \sum_{k=1}^{K} p_k^2\right). \tag{16}$$

Defining the Gini impurity as

$$\mathcal{G}(\mathbf{p}) = 1 - \sum_{k=1}^{K} p_k^2, \tag{17}$$

we obtain

$$V_{\text{inter}}(x) \le \frac{\Delta_{\max}^2}{2}\mathcal{G}(\mathbf{p}). \tag{18}$$

Thus, probability dispersion provides a worst-case upper bound on inter-mode score-gradient variability, with the geometric information compressed into the single constant $\Delta_{\max}$.

B.2 Relation Between Gini Impurity and Shannon Entropy

For the binary case $K = 2$, let $\mathbf{p} = \{p, 1-p\}$ with $p \in [0, 0.5]$. Shannon entropy and Gini impurity are

$$H(p) = -p\log p - (1-p)\log(1-p), \tag{19}$$

$$\mathcal{G}(p) = 2p(1-p). \tag{20}$$

Their derivatives satisfy

$$\frac{d\mathcal{G}}{dp} = 2 - 4p > 0, \qquad \frac{dH}{dp} = \log\frac{1-p}{p} > 0 \qquad \text{for } p < 0.5, \tag{21}$$

so both quantities increase monotonically on this interval.

More generally, using $p(1-p) \le -p\log p$ for $p \in [0,1]$,

$$\mathcal{G}(\mathbf{p}) = \sum_{k=1}^{K} p_k(1-p_k) \le -\sum_{k=1}^{K} p_k \log p_k = H(\mathbf{p}). \tag{22}$$

Combining the two inequalities gives

$$V_{\text{inter}}(x) \le \frac{\Delta_{\max}^2}{2}\mathcal{G}(\mathbf{p}) \le \frac{\Delta_{\max}^2}{2}H(\mathbf{p}). \tag{23}$$

Therefore, reducing a mass-based entropy score tightens a worst-case upper bound on inter-mode score-gradient variability; it does not imply that entropy uniquely determines the underlying gradient geometry.

B.3 Detailed Analysis of the Anisotropic Gap

The looseness of the above bound can be characterized by

$$\mathcal{S}(\mathbf{p}) = \frac{\Delta_{\max}^2}{2}\mathcal{G}(\mathbf{p}) - V_{\text{inter}}(x) = \frac{1}{2}\sum_{i\neq j} p_i p_j \left(\Delta_{\max}^2 - \|\mu_i^g - \mu_j^g\|_2^2\right). \tag{24}$$

Tight regime. If the relevant pairwise gradient differences are close to $\Delta_{\max}$, then $\mathcal{S}(\mathbf{p})$ is small and mass dispersion provides a relatively tight bound.

Loose regime. If many pairs have substantially smaller gradient differences than $\Delta_{\max}$, the slack increases even when the semantic-mode probabilities remain unchanged.

Takeaway. The bound therefore explains the anisotropic gap structurally: mass-based uncertainty preserves how probability is distributed across modes, but replaces heterogeneous pairwise geometry with a common worst-case scale. Geometry-aware measures are motivated by retaining this missing pairwise structure rather than treating the entropy bound as a direct estimate of gradient variability.

C Link Between Geometry-aware Uncertainty and Gradient Variance

CD and BoT explicitly retain geometric information that is absent from purely mass-based uncertainty. We provide a structural connection to score-gradient variability under a semantic–gradient alignment assumption. For a rollout y_i, let $g_i = \nabla_\theta \log \pi_\theta(y_i \mid x)$ and let e_i denote its normalized semantic embedding. We assume locally that semantic distance controls score-gradient discrepancy,

$$\|g_i - g_j\|_2^2 \le L_g^2 \, d(e_i, e_j), \tag{25}$$

where $d(\cdot, \cdot)$ denotes the cosine-based semantic distance used by the geometry-aware measure and L_g is task dependent.

For a uniformly sampled rollout group, the empirical pairwise score-gradient variability satisfies

$$V_{\text{score}}(x) = \frac{1}{2} \sum_{i,j} \pi_i \pi_j \|g_i - g_j\|_2^2 \le \frac{L_g^2}{2} \sum_{i,j} \pi_i \pi_j d(e_i, e_j), \tag{26}$$

with $\pi_i = 1/G$. The right-hand side is proportional to CD, showing that CD preserves the pairwise semantic quantities appearing in this bound instead of replacing them by a single worst-case disagreement constant.

A similar argument applies at the semantic-mode level. Let Π_k denote the empirical mass of cluster k, $\bar{g}_k$ its mean score-gradient, and μ_k its normalized embedding centroid as defined in Section 3.2. If cluster-level semantic distance is locally aligned with differences among the corresponding mean gradients, the inter-mode gradient variability can be bounded, up to a task-dependent constant, by the weighted deviations of μ_k from the shared barycentric direction $\mu^\star$. BoT is constructed from precisely these mode-level deviations and therefore provides a structured proxy for inter-mode geometry while discarding benign within-cluster paraphrastic variation.

These relations are intended as structural justification rather than optimization guarantees. The semantic–gradient alignment condition is a modeling assumption, and semantic embeddings need not reproduce gradient geometry exactly. The statistical analysis in Appendix D provides empirical support for this connection on the QA settings considered in our analysis.

D Statistical Analysis of Uncertainty Measures

We conduct statistical experiments to examine how rollout-level uncertainty relates to the two optimization quantities motivating our analysis: score-gradient variability, which isolates the geometric effect studied in the anisotropic gap, and reward-weighted update variability, which additionally reflects the relative reward signal relevant to the calibration gap.

D.1 Data and Experimental Setup

We randomly sample 1,000 queries from each of NarrativeQA and Qasper. For each query q, we generate a rollout group $\mathcal{Y}(q) = \{y_i\}_{i=1}^G$ from $\pi_\theta(\cdot \mid q)$, compute the score-gradient $g_i = \nabla_\theta \log \pi_\theta(y_i \mid q)$, obtain reward r_i, and compute the normalized GRPO advantage $\hat{A}_i$. For the statistical diagnostics only, we remove the 20 queries with the largest target variance to prevent

a small number of extreme values from dominating the analysis; this trimming is not used in any training or downstream evaluation experiment.

D.2 Target Quantities

To isolate response-induced geometry, we define the score-gradient variability

$$\mathcal{V}_{\text{score}}(q) = \frac{1}{G}\sum_{i=1}^{G} \|g_i - \bar{g}\|_2^2, \qquad \bar{g} = \frac{1}{G}\sum_{j=1}^{G} g_j. \tag{27}$$

This quantity measures variation among the score-gradient directions induced by sampled responses and is the target used for the anisotropic-gap analyses in Figure 1(a,c).

To incorporate the learning signal supplied by relative rewards, we further define the reward-weighted update contribution $u_i = \hat{A}_i g_i$ and

$$\mathcal{V}_{\text{update}}(q) = \frac{1}{G}\sum_{i=1}^{G} \|u_i - \bar{u}\|_2^2, \qquad \bar{u} = \frac{1}{G}\sum_{j=1}^{G} u_j. \tag{28}$$

Unlike $\mathcal{V}_{\text{score}}$, this quantity reflects both response-induced gradient geometry and within-group reward contrast, and is therefore used for the calibration-oriented diagnostics in Figure 1(b,d).

D.3 Uncertainty Measures

For each query, we compute token entropy, semantic entropy (SE), Cosine Dispersion (CD), and Barycentric Transport (BoT). Token entropy summarizes predictive uncertainty at the token level; SE summarizes probability mass across semantic clusters; CD measures pairwise semantic geometry; and BoT measures mode-level dispersion around a shared barycentric direction. Each measure produces one scalar uncertainty score $u(q)$.

D.4 Association with Score-Gradient Variability

Figure 1(a) evaluates monotonic association between each uncertainty score and $\mathcal{V}_{\text{score}}(q)$ using Spearman rank correlation,

$$\rho_s = \text{Corr}\left(\text{rank}(u(q)), \text{rank}(\mathcal{V}_{\text{score}}(q))\right). \tag{29}$$

Geometry-aware measures show stronger association on NarrativeQA, while the differences are smaller on Qasper. This dataset dependence is important: the results provide empirical support for the semantic–gradient alignment assumption in the evaluated settings, but do not imply that semantic geometry universally determines gradient geometry.

D.5 Identification of High-Update-Variability Queries

To examine the calibration gap, Figure 1(b) evaluates whether response-only uncertainty can identify queries with large reward-weighted update variability. Let $\mathcal{H}$ denote the top 10% of queries ranked by $\mathcal{V}_{\text{update}}(q)$. We report AUC when $u(q)$ is used as the scoring function for classifying membership in $\mathcal{H}$, together with top-10% precision,

$$\text{Precision@10\%} = \frac{|\mathcal{H} \cap \text{Top-10\%}(u)|}{|\text{Top-10\%}(u)|}. \tag{30}$$

These metrics are descriptive diagnostics rather than direct tests of downstream performance. On NarrativeQA, geometry-aware measures improve identification of high-update-variability queries, whereas the separation is substantially weaker on Qasper. The remaining mismatch is consistent with the calibration gap: response disagreement alone does not determine the reward-weighted learning value of a rollout group.

D.6 Statistical Robustness via Paired Bootstrap

Figure 1(c) evaluates the robustness of the score-gradient association using paired bootstrap resampling. For each bootstrap replicate b, we resample queries with replacement and compute

$$\Delta^{(b)}(u) = \rho_s(\mathrm{BoT}, \mathcal{V}_{\mathrm{score}}) - \rho_s(u, \mathcal{V}_{\mathrm{score}}). \tag{31}$$

On NarrativeQA, the bootstrap intervals show a clear positive advantage of BoT over entropy-based measures and a smaller advantage over CD. On Qasper, the differences are much smaller and the corresponding intervals do not consistently separate from zero. The bootstrap analysis therefore supports a stronger geometry-aware advantage on NarrativeQA without implying universal dominance across datasets.

D.7 Held-out Update-Variability Prediction

Finally, Figure 1(d) evaluates whether the relationship between uncertainty and reward-weighted update variability transfers to held-out queries. For each uncertainty measure $u(q)$, we fit

$$\widehat{\mathcal{V}}_{\mathrm{update}}(q) = \beta_0 + \beta_1 u(q) \tag{32}$$

and evaluate cross-validated mean absolute error and Spearman correlation. Geometry-aware measures provide stronger held-out tracking on NarrativeQA, while differences are again smaller on Qasper. Because we do not conduct paired significance tests across folds, these results are interpreted as descriptive predictive diagnostics rather than statistically significant superiority claims.

D.8 Discussion: Dataset-dependent Effects

The relationship between semantic disagreement and optimization variability is substantially stronger on NarrativeQA than on Qasper. One plausible explanation is that Qasper contains many semantically distinct responses with relatively similar reward outcomes, so response geometry alone is less predictive of the reward-weighted update. NarrativeQA more often produces competing answer hypotheses with clearer differences in reward, creating stronger alignment between response structure and optimization variability. This interpretation is consistent with both the weaker statistical separation on Qasper and the smaller contribution of RD observed in the corresponding ablation.

D.9 Final Takeaway

Across the two QA datasets, geometry-aware measures generally provide a closer description of score-gradient variation than purely mass-based uncertainty, but the strength of this relationship is clearly task dependent. Moreover, the weaker correspondence between response-only uncertainty and reward-weighted update variability shows why accurate quantification alone is insufficient. Together, these results provide empirical support for the two-stage motivation of GCPO: first characterize disagreement more faithfully, and then calibrate that disagreement with information relevant to learning.

E Implementation Details

E.1 Benchmarks

NarrativeQA (Kočiskỳ et al., 2018) is an open-ended question answering benchmark constructed from books and movie scripts. Unlike extractive QA datasets, answers are not directly grounded in short spans, requiring models to synthesize information from long narratives. Its open-ended answer space naturally produces response-level semantic variation.

Qasper (Dasigi et al., 2021) is a question answering dataset over scientific articles, where questions often require synthesizing evidence across multiple sections. It contains diverse answer types and provides a challenging setting in which semantically distinct responses can receive similar rewards.

HotpotQA (Yang et al., 2018) is a multi-hop QA benchmark requiring information to be combined across supporting passages. Its more constrained answer space complements the open-ended QA settings used in our main analysis.

MATH500 (Hendrycks et al., 2021) is a subset of the MATH benchmark containing competition-level mathematical problems with deterministic final answers. It allows us to evaluate whether the proposed uncertainty mechanism transfers from short-form answer geometry to full reasoning trajectories.

AIME24 contains American Invitational Mathematics Examination problems with deterministic numerical answers and multi-step solutions. Because AIME24 does not provide a training split, we evaluate the policy trained on OlympiadBench, as described below.

OlympiadBench (Huang et al., 2024) contains challenging competition-level reasoning problems and admits diverse solution strategies despite deterministic answer evaluation.

E.2 Baselines

GRPO (Shao et al., 2024) is the base group-relative policy optimization method. It uses normalized within-group rewards but does not apply an additional query-level uncertainty modulation.

Semantic-entropy methods. Seed-GRPO (Chen et al., 2025a) and EMPO (Zhang et al., 2025b) use response-level entropy to regulate optimization based on uncertainty across generated outputs. They form the most direct mass-based uncertainty baselines for our study.

Kernel-based entropy. KLE (Nikitin et al., 2024) incorporates semantic similarity through a kernel construction rather than relying only on discrete semantic-cluster counts, providing a stronger similarity-aware uncertainty baseline.

Consistency-based uncertainty. GRPO-CARE (Chen et al., 2025b) uses agreement among sampled responses as a confidence signal for regulating policy optimization.

Recent policy-optimization baselines. We additionally compare with Q-Hawkeye (Xie et al., 2026), EG-SPO (Hu et al., 2026), and R^{2}VPO (Luo et al., 2026). Because these methods were originally developed under different objectives, modalities, or optimization pipelines, we adapt their central control signals to the same GRPO training framework used by GCPO. The resulting comparisons should therefore be interpreted as controlled GRPO adaptations rather than exact reproductions of every component in the original systems.

Q-Hawkeye adaptation. We retain its uncertainty-aware reweighting principle and define group-level uncertainty from the variance of realized rollout rewards, $u(q) = \mathrm{Var}(r_1, \ldots, r_G)$. After normalization, the corresponding query weight is applied to the group-normalized GRPO advantage.

EG-SPO adaptation. We aggregate token-level predictive entropy into a response-level uncertainty score and use the resulting signal to gate the group-normalized advantage. This preserves its entropy-guided gradient-scaling principle while keeping the underlying GRPO training protocol fixed.

R^{2}VPO adaptation. We retain the idea of damping updates with high policy-ratio variability through a response-level factor of the form $w_i = 1/(1 + \lambda\ \mathrm{Var}_t(\rho_{i,t}))$. This baseline controls a different source of optimization variability and therefore provides a complementary comparison with rollout-level uncertainty regulation.

E.3 Training Details

Optimization configuration. Unless otherwise specified, models are trained for 2,000 steps with learning rate 5×10^{-5}, KL coefficient $\beta = 0.002$, group size $G = 16$, and sampling temperature 0.9. We set $\alpha_{\mathrm{base}} = 0.6$ by default and use $\alpha_G = \alpha_{\mathrm{base}}/\log G$ as defined in Section 3; for brevity, the horizontal axis in Figure 4 denotes α_{base} by α. QA experiments use Qwen2.5-1.5B-Instruct by default, while the mathematical-reasoning experiments in Figure 5 use Qwen2.5-MATH-1.5B. The backbone study follows the six models reported in Figure 6. Training is performed on PSC Bridges-2 using a single H100 GPU per run.

Task-dependent response representation. For QA-style tasks, CD and BoT operate on the extracted free-form answer text. For mathematical and exact-match tasks, we use the full generated solution rather than only the final extracted answer, so that the uncertainty measure reflects variation across reasoning trajectories.

Semantic clustering and embedding details. For CD and BoT, we use `all-MiniLM-L6-v2` as the embedding function f_{emb}, implemented with `SentenceTransformer`. Given normalized

Benchmark	Metric / extraction	Reward range for RD
Qasper	$2.0\times$ token-F1 over references	$[0, 2]$
NarrativeQA	$2.0\times$ token-F1 over references	$[0, 2]$
HotpotQA	$2.0\times$ QA F1/EM-style score	$[0, 2]$
MATH500	2.0 if extracted answer matches, else 0	$[0, 2]$
AIME24	2.0 if extracted answer matches, else 0	$[0, 2]$
OlympiadBench	2.0 if extracted answer matches, else 0	$[0, 2]$

Table 2: Per-task reward definitions used for reward-dispersion normalization.

response embeddings μ_i, CD computes $D_{ij} = \mathrm{clip}(1 - \mu_i^\top \mu_j, 0, 1)$ and then applies Eq. (6). BoT additionally groups responses using `roberta-large-mnli` with entailment threshold 0.35. Clustering is greedy and order dependent: the first response initializes the first cluster, and each subsequent response is compared with the first response assigned to each existing cluster. The response joins the cluster with highest entailment probability when this probability is at least 0.35; otherwise a new cluster is created. Let Π_k denote the empirical cluster mass and μ_k the normalized centroid of cluster k; BoT then follows Eqs. (8)–(10).

Reward definitions and RD normalization. Reward dispersion is computed separately within each query-level rollout group. For reward range $[r_{\min}, r_{\max}]$,

$$\mathrm{RD}_{\max}(G) = \frac{2}{G} \left\lfloor \frac{G}{2} \right\rfloor \left\lceil \frac{G}{2} \right\rceil (r_{\max} - r_{\min}), \tag{33}$$

and the normalized quantity used throughout the paper is

$$\mathrm{RD}(q) = \mathrm{clip}\left(\frac{\mathrm{RD}_{\mathrm{raw}}(q)}{\mathrm{RD}_{\max}(G)}, 0, 1 \right). \tag{34}$$

All experiments use the configured reward range $[0, 2]$; for the default group size $G = 16$, this gives $\mathrm{RD}_{\max}(16) = 16$.

For QA benchmarks, rewards vary continuously with token-level answer overlap, whereas mathematical benchmarks use binary correctness after answer extraction and normalization. The same normalized $\mathrm{RD}(q)$ is then used by $\omega_{\mathrm{RD}}(q) = 1 + \alpha_G \, \mathrm{RD}(q)$.

End-to-end GCPO procedure. For each query q, we sample G rollouts, compute rewards and normalized GRPO advantages $\hat{A}_i$, select the task-dependent response representation, compute CD or BoT from its semantic embeddings, compute normalized RD from the same reward group, form the query-level factor $\omega_{\mathrm{geo}}(q)\omega_{\mathrm{RD}}(q)$, scale every $\hat{A}_i$ in the group by this common factor, and finally apply the standard GRPO update using the modulated advantages.

F Discussion

F.1 Interaction Between Geometric Uncertainty and Reward Dispersion

GCPO separates two complementary roles in rollout-level regulation: geometry characterizes the reliability of the observed response variation, while reward dispersion characterizes the amount of within-group reward contrast available for learning. Both modulation terms are therefore centered at the neutral value 1, so that the original GRPO update is recovered when neither signal calls for additional adjustment.

For the geometric term, since $U_{\mathrm{geo}}(q) \in [0, 1]$, where $U_{\mathrm{geo}} \in \{\mathrm{CD}, \mathrm{BoT}\}$, we define

$$\omega_{\mathrm{geo}}(q) = \mathrm{clip}\left(1 - \alpha_G U_{\mathrm{geo}}(q)^2, 0, 1\right). \tag{35}$$

The offset at 1 makes zero geometric disagreement recover the unmodulated update. The quadratic form keeps the effect conservative for small deviations while increasing sensitivity as disagreement becomes larger, concentrating attenuation on rollout groups with substantial geometric variation rather than penalizing minor response differences. The lower clipping boundary prevents sufficiently

strong modulation from producing a negative weight and reversing the update direction, while the resulting $[0, 1]$ range gives the geometric term the interpretation of a bounded reliability gate.

Reward dispersion plays the complementary role. After normalization to $\mathrm{RD}(q) \in [0, 1]$, we define

$$\omega_{\mathrm{RD}}(q) = 1 + \alpha_G \, \mathrm{RD}(q). \tag{36}$$

Unlike the geometric term, which can only attenuate the update, reward dispersion moves the weight above the same neutral point when the rollout group exhibits stronger reward contrast. This asymmetry reflects the roles of the two signals: geometric disagreement indicates when variation may be unreliable, whereas reward dispersion indicates when the group contains stronger relative learning contrast.

Because both quantities describe properties of the same rollout group, their multiplicative composition forms a single query-level scaling factor,

$$\widetilde{A}_i = \hat{A}_i \, \omega_{\mathrm{geo}}(q) \, \omega_{\mathrm{RD}}(q). \tag{37}$$

This composition preserves the relative normalized GRPO advantages within the group while allowing update strength to depend jointly on reliability and learning value. In particular, neither geometric disagreement nor reward contrast alone determines the final weight: amplification from RD remains conditioned on the geometric term, while geometric attenuation can be partially offset when the same group exhibits informative reward differentiation.

The shared coefficient $\alpha_G = \alpha_{\mathrm{base}} / \log G$ controls the overall departure from the neutral update while reducing sensitivity to rollout-group size. The functional forms above are chosen to instantiate the desired behavior of conservative treatment of small disagreement, bounded attenuation of unreliable variation, and bounded amplification of reward-differentiated groups. We do not claim that the quadratic mapping, clipping boundaries, or multiplicative composition are uniquely optimal; Figure 4 instead examines the empirical operating regime of this modulation and shows that intermediate modulation strength provides the best balance.

A remaining limitation is that reward dispersion measures reward contrast rather than reward quality. Large $\mathrm{RD}(q)$ can arise from genuinely informative differences among responses, but also from noisy or mis-specified rewards. The geometric term provides a complementary check on response structure, but does not guarantee that the reward signal itself is correct. GCPO therefore assumes that within-group reward differences are sufficiently meaningful to serve as a calibration signal.

F.2 Relation to KL-constrained Optimization

Recent work such as GVPO Zhang et al. (2025a) derives policy update weights from KL-constrained reward optimization and studies stability from an objective-level perspective. GCPO addresses a different level of the training process: given a GRPO-style update, it modulates the contribution of rollout groups according to semantic disagreement and reward contrast.

KL regularization constrains policy movement at the distribution level, whereas GCPO introduces query-level reweighting based on properties of the sampled rollout group. These mechanisms are therefore complementary rather than equivalent. Our goal is not to derive an optimal KL-constrained RL objective, but to study how rollout-level uncertainty can be quantified, calibrated, and utilized within group-based policy optimization.

F.3 Effect of Query-level Reweighting

GCPO does not preserve the expectation of the unweighted GRPO estimator. Let

$$\hat{g}_{\mathrm{GRPO}}(q) = \frac{1}{G} \sum_{i=1}^{G} \hat{A}_i \nabla_\theta \log \pi_\theta(y_i \mid q), \tag{38}$$

and let

$$w(q) = \omega_{\mathrm{geo}}(q) \omega_{\mathrm{RD}}(q). \tag{39}$$

GCPO uses

$$\hat{g}_{\mathrm{GCPO}}(q) = w(q) \hat{g}_{\mathrm{GRPO}}(q), \tag{40}$$

so in general

$$\mathbb{E}\left[\hat{g}_{\mathrm{GCPO}}(q)\right] \neq \mathbb{E}\left[\hat{g}_{\mathrm{GRPO}}(q)\right]. \tag{41}$$

This difference is the consequence of changing the relative contribution of rollout groups according to the estimated reliability and informativeness of their signals. Moreover, $w(q)$ is computed from the same sampled responses and rewards that determine the GRPO update, so the weight and the update are statistically dependent. We therefore do not claim a generic closed-form variance reduction or monotonic decrease in gradient variance. Rather, GCPO uses uncertainty and reward contrast to regulate which query-level updates receive greater or smaller weight. A formal convergence analysis under this adaptive, sample-dependent reweighting is left for future work.

F.4 On the Semantic-Gradient Alignment Assumption

The theoretical connection in Appendix C relies on a semantic–gradient alignment assumption: semantic distance should carry information about differences in the score-gradient directions induced by different responses. This is a modeling hypothesis rather than a mathematical identity. We do not assume that embedding distance exactly approximates gradient distance, or that the relationship holds for every response pair. Instead, the assumption only requires a statistical tendency that semantically similar responses are more likely to induce similar optimization effects than semantically distant ones.

This assumption also has an intuitive basis. Responses that differ only by paraphrasing or minor reasoning variations often preserve much of the same semantic and generation structure, whereas substantially different answers or reasoning paths can alter a larger portion of the trajectory and therefore induce more different gradients. This relationship is necessarily imperfect, since tokenization, sequence length, and model parameterization can all create mismatches between semantic and gradient spaces.

Importantly, we explicitly examine this assumption empirically rather than treating it as unverified. In Appendix D, we directly compute score-gradient variability from sampled rollouts and measure how closely different uncertainty signals track it. Geometry-aware measures show substantially stronger alignment on NarrativeQA and a weaker but still observable relationship on Qasper. These results provide empirical support for the assumption while also showing that its strength is task dependent.

Accordingly, semantic-gradient alignment should be viewed as a domain-dependent inductive bias rather than a universal guarantee. GCPO does not require semantic and gradient spaces to coincide; it only benefits when semantic geometry provides a useful proxy for optimization-relevant variation. Our analysis therefore provides structural justification under this assumption, while the empirical experiments examine whether the required relationship holds in practice.

F.5 Computational Overhead Analysis

The dominant cost in GRPO-style training is rollout generation by the policy model. GCPO adds semantic embedding, pairwise geometry, optional NLI clustering, and reward-dispersion computation. For group size G and embedding dimension d, CD requires $O(G^2 d)$ pairwise similarity computation after embedding extraction. BoT additionally performs greedy NLI clustering, requiring at most $O(GK)$ entailment comparisons for $K \leq G$ clusters, followed by cluster-centroid aggregation. RD itself requires only $O(G)$ arithmetic operations.

With the default $G = 16$, rollout generation remains the dominant cost in our experiments, and the additional uncertainty computations introduce modest overhead relative to policy generation. BoT is more expensive than CD because of the NLI clustering stage; this computational difference should be considered when choosing between the two variants.

F.6 On the OT Interpretation of BoT

Barycentric Transport is not an exact optimal-transport computation. It is a lightweight geometry-aware surrogate inspired by the barycentric view of OT: semantic modes are represented by normalized embedding centroids and compared with a shared weighted consensus direction using cosine-based deviation. No Wasserstein optimization problem is solved during training.

The purpose of this construction is therefore not to approximate an OT distance faithfully, but to retain mode-level semantic structure at low computational cost. Under the semantic–gradient alignment assumption, these mode-level deviations provide a structural proxy for inter-mode score-gradient variation. Appendix D shows that BoT provides strong alignment on NarrativeQA, while the advantage is smaller and less consistent on Qasper, emphasizing that the usefulness of the surrogate remains task dependent.

BoT additionally inherits sensitivity from both the embedding model and the semantic clustering procedure. In our implementation, clustering uses a fixed NLI model, a threshold of 0.35, and greedy order-dependent assignment, so changes to these components can alter the induced semantic modes. The results in this paper show that this lightweight surrogate is effective in the evaluated settings, but do not imply that exact OT computation is unnecessary in general.