GEOMA: GEOMETRIC AND ECONOMETRIC OBJECTIVES FOR MULTI-REWARD ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Alignment of large language models is increasingly formulated as optimization over multiple rubric signals. These signals typically exhibit strong statistical dependencies, ranging from redundancy to anti-correlation (e.g., conciseness versus correctness), raising the question of how to robustly convert vector-valued rewards into scalar advantages. While recent state-of-the-art methods like GDPO address scale discrepancies via per-dimension normalization, they ignore reward geometry by treating coordinates as orthogonal. This mishandles correlations: redundant objectives are double-counted, while anti-correlated rewards are dominated by high-variance trade-off directions, and allows models to exploit easy objectives at the expense of hard constraints. We introduce **GEOMA** (*Geometric and Econometric Objectives for Multi-reward Alignment*), a framework that decomposes reward aggregation into geometric preconditioning via *covariance sphering* of reward vectors, and econometric aggregation such as *Nash Welfare* and *SoftMin*. We formally characterize these objectives, providing theoretical guarantees for their robustness to reward hacking and signal redundancy. Empirically, we demonstrate that GEOMA outperforms GDPO on Math reasoning and Tool Calling. On mathematical reasoning, it improves overall accuracy by 1.5% on average while achieving $1.5\times$ token efficiency over GDPO.

1 INTRODUCTION

Modern LLMs are evaluated against a high-dimensional vector of distinct rubric signals: factual correctness, helpfulness, safety, strict formatting constraints, and stylistic requirements such as conciseness (Ouyang et al., 2022; Bai et al., 2022). These signals are derived from heterogeneous sources, ranging from verified rule-based functions to model-based "LLM-as-a-judge" evaluators, each encoding qualitatively different metrics (Zheng et al., 2023). Consequently, the central challenge in LLM post-training has shifted from estimating a reward to aggregating this multi-dimensional signal into a scalar advantage that can effectively guide policy optimization.

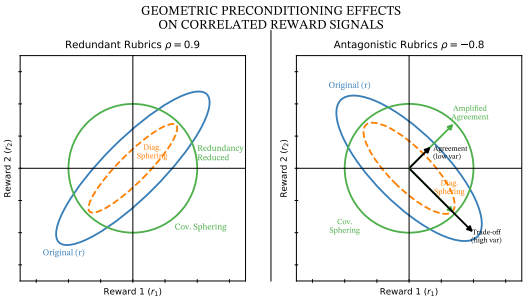

Figure 1: **Geometric Failure Modes. Left (Redundancy, $\rho = 0.9$):** GDPO preserves collinearity, leading to double-counting. GEOMA projects to an isotropic sphere, effectively removing redundancy. **Right (Antagonism, $\rho = -0.8$):** The direction of mutual improvement is drowned out by the high-variance trade-off axis. GEOMA amplifies this weak signal, realigning gradients toward the Pareto frontier.

Standard approaches like GRPO (Shao et al., 2024) rely on linear scalarization. This implicitly imposes an orthogonal geometry on the reward space, assuming independence and equivariance between objectives. Consequently, naïve summation suffers from *scale dominance*, where high-variance rewards overpower subtle signals, and *redundancy*, where correlated metrics (e.g., correctness and reasoning) are double-counted, distorting the gradient. This pathology is particularly acute when auxiliary objectives are anti-correlated with the primary task (Hong et al., 2024; Yu et al., 2025). In such regimes, the variance is concentrated along the "trade-off" axis (improving one reward at the expense of the other), drowning out the weak, low-variance signal required for simultaneous improvement (the "agreement" direction).

Recent methods like GDPO (Liu et al., 2026) address scale discrepancies via per-dimension normalization (diagonal rescaling). However, this ignores the *covariance structure*. When rewards are correlated, diagonal scaling fails to disentangle redundant signals or resolve antagonistic trade-offs. Robust alignment requires a geometric perspective that treats rewards as multivariate data, *preconditioning* the space to remove cross-correlations. A rigorous treatment of multi-reward alignment requires a geometric perspective that treats the reward vector as a sample from a multivariate distribution, accounting for both the scale of individual dimensions and their covariance structure to improve optimization on ill-conditioned reward vectors (Mahalanobis, 2018). See Figure 1 which illustrates the space of correlated (redundant) and anti-correlated (antagonistic) rubrics, which may contribute to sub-optimal optimization landscapes.

Parallel to these geometric concerns is the econometric challenge of substitution. Even in a standardized space, linear aggregation permits reward hacking: models can maximize the sum by exploiting easy objectives (e.g., length) while neglecting harder ones. To resolve this, we leverage Welfare Economics (Nash et al., 1950), with Max Nash Welfare and egalitarian welfare (softmin), enforcing consensus and penalizing the collapse of any single dimension.

We introduce **GEOMA** (*Geometric and Econometric Objectives for Multi-reward Alignment*), a framework that decomposes aggregation into two axes. First, Geometric Preconditioning uses *Covariance Sphering* (inverse square-root covariance) to resolve correlations, generalizing diagonal normalization. Second, Econometric Aggregation uses welfare functions like *Nash* to drive reward fairness. We demonstrate that GEOMA unifies prior methods and achieves state-of-the-art constraint satisfaction. We make three primary contributions in this work

A Unified Framework (GEOMA): We introduce a rigorous formalism that decomposes multi-reward aggregation into *geometric preconditioning* and *econometric aggregation*. This framework unifies existing heuristics like GRPO and GDPO as specific, often suboptimal, instances of a broader design space.

Analysis of Multi-Reward Failure Modes: We challenge the prevailing practice of treating auxiliary objectives as simple additive regularizers. We theoretically identify and analyze two fundamental pathologies in current pipelines: *geometric redundancy* (double-counting correlated signals) and *econometric substitution* (reward hacking).

State-of-the-Art Performance: We demonstrate that GEOMA outperforms the current state-of-the-art baseline, GDPO. On Tool Calling, GEOMA achieves a 1.6% accuracy gain on Non-Live tasks while converging approximately $2\times$ faster on formatting constraints. On Math Reasoning, we achieve 28.0% accuracy on AIME-25 while reducing generation cost by $3\times$ compared to GRPO and $2\times$ compared to GDPO.

2 Related Work

The dominant paradigm for aligning large language models relies on Reinforcement Learning from Human Feedback (RLHF), standardizing on algorithms like PPO (Ouyang et al., 2022; Schulman et al., 2017) and, more recently, group-based methods such as GRPO (Shao et al., 2024). Unlike pairwise preference algorithms like DPO (Rafailov et al., 2024), group-based RLHF leverages sets of responses to estimate baseline advantages directly. As evaluation pipelines increasingly rely on granular, model-based judges (Zheng et al., 2023), the optimization signal has transitioned from a single scalar to a multi-dimensional reward vector capturing correctness, safety, and formatting. The standard approach resolves this via linear scalarization (Hayes et al., 2021). However, this assumes independent objectives and often results in reward hacking, where models exploit heuristic constraints or display severe verbosity biases (Saito et al., 2023; Hu et al., 2024).

Recent efforts to stabilize multi-reward alignment have focused on scale normalization. Notably, GDPO (Liu et al., 2026) applies a diagonal variance normalization to prevent scale dominance among rewards. Similarly, Zhao et al. (2025b) consider Geometric mean of rewards over the batch. GEOMA generalizes this intuition by introducing covariance sphering to explicitly remove cross-rubric redundancy. Furthermore, we connect LLM alignment to the rich literature of Welfare Economics (Nash et al., 1950; Sen, 2017). By applying axiomatic aggregation principles such as the Max Nash Welfare product and egalitarian max-min objectives, GEOMA enforces robust Pareto improvements, actively mitigating the trade-offs and length-biases documented in recent alignment benchmarks (Dubois et al., 2024; Lambert et al., 2024).

3 PRELIMINARIES AND PROBLEM FORMULATION

We consider the problem of aligning a language model $\pi_\theta(\cdot|x)$ to maximize a vector of preference signals. For a given prompt x, we sample a group of G candidate responses $\{y^{(j)}\}_{j=1}^{G}$. Each response is evaluated against K distinct rubrics, yielding a reward vector $\mathbf{r}^{(j)} \in \mathbb{R}^K$.

Group Statistics. To mitigate sensitivity to prompt difficulty, we operate on group-relative statistics. We define the group mean $\boldsymbol{\mu} \in \mathbb{R}^K$, the *centered reward vector* $\mathbf{u}^{(j)} \in \mathbb{R}^K$, and the group variance vector $\boldsymbol{\sigma}^2 \in \mathbb{R}^K$ as:

$$\boldsymbol{\mu} = \frac{1}{G}\sum_{j=1}^{G} \mathbf{r}^{(j)}, \quad \mathbf{u}^{(j)} = \mathbf{r}^{(j)} - \boldsymbol{\mu}, \quad \boldsymbol{\sigma}^2 = \frac{1}{G}\sum_{j=1}^{G} (\mathbf{u}^{(j)})^2 \tag{1}$$

Standard Baselines. Current methods aggregate these statistics into a scalar advantage $\tilde{A}^{(j)}$:

- **GRPO (Sum-then-Normalize):** Sums rewards first, then normalizes. $\tilde{A}^{(j)} \propto \sum_k r_k^{(j)}$.
- **GDPO (Normalize-then-Sum):** Normalizes dimensions independently. $\tilde{A}^{(j)} = \sum_k u_k^{(j)}/\sigma_k$.

Policy Optimization. The scalar advantage $\tilde{A}^{(j)}$ weights the standard policy gradient surrogate objective. Let $\rho^{(j)}$ be the importance sampling ratio. We maximize:

$$\mathcal{J}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{j=1}^{G}\left(\min\left(\rho^{(j)}\tilde{A}^{(j)}, \text{clip}(\rho^{(j)}, 1-\epsilon, 1+\epsilon)\tilde{A}^{(j)}\right) - \beta\mathbb{D}_{\text{KL}}\right)\right] \tag{2}$$

4 GEOMA FRAMEWORK AND OBJECTIVES

We now introduce GEOMA (*Geometric and Econometric Objectives for Multi-reward Alignment*), a framework that formalizes reward aggregation as the composition of a *geometric preconditioning* step with an *econometric aggregation* objective.

While baselines like GDPO treat aggregation as a scalar operation, GEOMA generalizes this to matrix-vector products. We define the scalar learning signal $\tilde{A}^{(j)}$ as:

$$\tilde{A}^{(j)} = \mathcal{G}(\mathbf{a}^{(j)}), \quad \text{where} \quad \mathbf{a}^{(j)} = P\mathbf{u}^{(j)} \tag{3}$$

Here, $\mathbf{u}^{(j)}$ is the centered reward vector (Sec. 3), $P \in \mathbb{R}^{K\times K}$ is a *preconditioning matrix* that handles correlation, and $\mathcal{G} : \mathbb{R}^K \to \mathbb{R}$ is an *aggregation functional* that handles trade-offs.

Geometric Preconditioning (P). This stage determines the geometry of the reward space. Let Σ be the covariance matrix of the centered rewards $\mathbf{u}$. We propose three preconditioners:

$$P_{\text{Id}} := I, \quad P_{\text{Diag}} := (\text{diag}(\Sigma) + \varepsilon I)^{-1/2}, \tag{4}$$

$$P_{\text{Cov}} := (\Sigma + \varepsilon I)^{-1/2}. \tag{5}$$

P_{Id} corresponds to GRPO. P_{Diag} recovers GDPO-style diagonal normalization. P_{Cov} is our proposed **Covariance Sphering**, which whitens the reward space to remove redundancy (see Table 1).

Econometric Aggregation ($\mathcal{G}$). Given preconditioned rewards $\mathbf{a}^{(j)}$, $\mathcal{G}$ defines the optimization priority.

- **Utilitarian (Sum):** $\mathcal{G}_{\text{sum}}(\mathbf{a}) = \sum_k a_k$. Assumes perfect substitutability.
- **Max Nash Welfare (MNW):** $\mathcal{G}_{\text{MNW}}(\mathbf{a}) = \sum_k \log\sigma(a_k)$. Maximizes the product of utilities, penalizing any dimension that collapses.
- **SoftMin:** $\mathcal{G}_{\text{softmin}}(\mathbf{a}) = -\log\sum_k e^{-a_k}$. A smooth bottleneck function that prioritizes the worst-performing metric.

Table 1: GEOMA design space. Let $\mathbf{u}^{(j)} = \mathbf{r}^{(j)} - \boldsymbol{\mu}$ and $\mathbf{a}^{(j)} = P\,\mathbf{u}^{(j)}$. Each cell lists the scalar signal $\tilde{A}^{(j)} = \mathcal{G}(a^{(j)})$ used to weight policy updates. We use $\psi(z) = \sigma(z)$ for MNW unless otherwise stated.

Geometric preconditioner	Definition of $a^{(j)}$	Utilitarian (Sum)	MNW	SoftMin
None ($P = I$)	$a^{(j)} = u^{(j)}$	$\sum_{k=1}^{K} a_k^{(j)}$	$\sum_{k=1}^{K} \log \psi\left(a_k^{(j)}\right)$	$-\log\left(\frac{1}{K}\sum_{k=1}^{K} e^{-a_k^{(j)}}\right)$
Diagonal sphering (P_{diag})	$a^{(j)} = (D + \varepsilon I)^{-1/2} u^{(j)}$	$\sum_{k=1}^{K} a_k^{(j)}$	$\sum_{k=1}^{K} \log \psi\left(a_k^{(j)}\right)$	$-\log\left(\frac{1}{K}\sum_{k=1}^{K} e^{-a_k^{(j)}}\right)$
Covariance sphering (P_{cov})	$a^{(j)} = (\Sigma + \varepsilon I)^{-1/2} u^{(j)}$	$\sum_{k=1}^{K} a_k^{(j)}$	$\sum_{k=1}^{K} \log \psi\left(a_k^{(j)}\right)$	$-\log\left(\frac{1}{K}\sum_{k=1}^{K} e^{-a_k^{(j)}}\right)$

4.1 Motivation: The Geometry of Aggregation

The prevalent baseline, GDPO, relies on diagonal preconditioning ($P = D^{-1/2}$). This operation standardizes the marginal variance of each reward but assumes that the axes of the reward space are orthogonal. In multi-objective alignment, this assumption is frequently violated, leading to two distinct failure modes.

Redundancy and Signal Double-Counting. Consider two rubrics, r_1 and r_2, measuring semantically similar attributes like "Truthfulness" and "Honesty" ($\rho \approx 1$). Ideally, aggregation should treat these as a single underlying factor. However, diagonal preconditioning rescales them independently. Under linear aggregation ($\tilde{A} = a_1 + a_2$), their contributions sum constructively, effectively doubling the weight of the shared latent factor relative to an independent third rubric r_3. This **double-counting** distorts the gradient, causing the policy to over-optimize for the redundant feature. Resolving this requires a transform that maps correlated directions onto a single orthogonal basis vector: the inverse covariance root $P = \Sigma^{-1/2}$.

Ill-Conditioned Reward Vectors. Conversely, consider structurally anti-correlated rubrics like "Conciseness" and "Correctness," where increasing one degrades the other. The covariance matrix Σ exhibits significant negative off-diagonal entries. The principal axes are the "agreement" direction $[1, 1]^\top$ (low variance, rare simultaneous improvement) and the "trade-off" direction $[1, -1]^\top$ (high variance). Diagonal normalization ignores this structure, treating the distribution as an axis-aligned ellipse, which allows the high-variance trade-off signal to drown out the agreement signal. **Covariance Sphering** ($P = \Sigma^{-1/2}$) whitens this distribution, scaling eigenvectors by their inverse square-root eigenvalues. This effectively *amplifies* the weak signal of simultaneous improvement and *dampens* noisy trade-offs, enabling efficient navigation toward the Pareto frontier.

5 Algorithmic Details

This section specifies how GEOMA is instantiated in practice, detailing the estimation of statistical moments, the computation of geometric sphering operators, and the construction of scalar advantages. Our goal is to provide a comprehensive recipe that ensures the mapping from multi-rubric rewards $\{\mathbf{r}^{(j)}\}_{j=1}^{G}$ to scalar weights $\{\tilde{A}^{(j)}\}_{j=1}^{G}$ is fully reproducible and numerically stable.

Moment Estimation Strategies. The quality of the geometric preconditioner depends critically on the estimation of the first and second moments ($\boldsymbol{\mu}, \Sigma$). For each prompt x, we sample a group $\{y^{(j)}\}_{j=1}^{G}$ and obtain reward vectors $\mathbf{r}^{(j)} \in \mathbb{R}^K$. We first compute the group mean $\boldsymbol{\mu}$ and centered rewards $\mathbf{u}^{(j)} = \mathbf{r}^{(j)} - \boldsymbol{\mu}$ as defined in Section 3. We estimate Σ via either **Group-wise Estimation**, capturing prompt-specific correlations but risking high variance, or **Batch-wise Estimation**, pooling samples across the minibatch for stability. We prefer batch-wise estimation to ensure the preconditioner is well-conditioned and avoid singularity. In both variants, we define the diagonal variance matrix $D = \text{diag}(\Sigma)$ and use these statistics for the downstream preconditioning steps.

Diagonal Sphering Implementation. Diagonal sphering normalizes the reward space axis-by-axis, ignoring off-diagonal dependencies. Given the variance matrix D, the preconditioner is defined as:

$$P_{\text{diag}} \;=\; (D + \epsilon I)^{-1/2},\; \mathbf{a}^{(j)} \;=\; P_{\text{diag}}\,\mathbf{u}^{(j)}. \tag{6}$$

In practice, this is implemented via coordinate-wise scaling: $a_k^{(j)} = u_k^{(j)} / \sqrt{D_{kk} + \epsilon}$. The regularization term $\epsilon > 0$ (typically 10^{-6}) is treated as a fixed numerical constant to prevent division by zero in the case of collapsed reward dimensions (e.g., if a heuristic reward outputs a constant value).

Covariance Sphering and Spectral Stabilization. Covariance sphering applies an inverse square-root operator to the centered vector $\mathbf{u}^{(j)}$, rotating and scaling the reward space to isotropy. To compute $(\Sigma + \epsilon I)^{-1/2}$, we perform a symmetric eigendecomposition:

$$\Sigma \;=\; U \,\mathrm{diag}(\boldsymbol{\nu})\, U^{\top}, \tag{7}$$

where $\boldsymbol{\nu} = [\nu_1, \ldots, \nu_K] \in \mathbb{R}^K_{>0}$ are the eigenvalues and U is the orthogonal matrix of eigenvectors. To handle rank deficiency, we employ either **Ridge Regularization** (computing $(\Sigma + \epsilon I)^{-1/2}$) or **Thresholded Projection** (zeroing eigenvalues below τ). Our primary experiments utilize Ridge Regularization to ensure stability while retaining full dimensionality.

Scalar Advantage Construction. Once the preconditioned advantages $\mathbf{a}^{(j)}$ are computed, we map them to a scalar signal using the aggregators $\mathcal{G} \in \{\mathcal{G}_{\text{sum}}, \mathcal{G}_{\text{MNW}}, \mathcal{G}_{\text{softmin}}\}$ defined in Section 4:

$$\tilde{A}^{(j)} \;=\; \mathcal{G}(\mathbf{a}^{(j)}). \tag{8}$$

Finally, following standard practice in GRPO/PPO, we optionally normalize the resulting scalar advantages $\{\tilde{A}^{(j)}\}_{j=1}^G$ within each group (e.g., subtracting the mean and dividing by the standard deviation) to control the scale of policy-gradient updates. This step renders the magnitude of the update invariant to the absolute scale of the welfare function.

Algorithm 1: **GEOMA** Scalar Advantage Computation

Require: Prompt x; policy π_θ; group size G; reward functions $\mathbf{r} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^K$
Require: Preconditioner $P \in \{I, P_{\text{diag}}, P_{\text{cov}}\}$
Require: Aggregator $\mathcal{W} \in \{\text{Sum}, \text{Nash}, \text{SoftMin}\}$
Ensure: Scalar advantages $\{\tilde{A}^{(j)}\}_{j=1}^G$

// Step 1: Generate and Score
Sample $y^{(1)}, \ldots, y^{(G)} \sim \pi_\theta(\cdot \mid x)$
$\mathbf{r}^{(j)} \leftarrow \mathbf{r}(x, y^{(j)})$ for $j = 1, \ldots, G$
// Step 2: Center Rewards
$\boldsymbol{\mu} \leftarrow \frac{1}{G} \sum_{j=1}^G \mathbf{r}^{(j)}$
$\mathbf{u}^{(j)} \leftarrow \mathbf{r}^{(j)} - \boldsymbol{\mu}$ for all j
// Step 3: Geometric Preconditioning
if $P = I$ **then**
  $\mathbf{a}^{(j)} \leftarrow \mathbf{u}^{(j)}$
else if $P = P_{\text{diag}}$ **then**
  $\boldsymbol{\sigma}^2 \leftarrow \mathrm{diag}\big(\frac{1}{G} \sum_j \mathbf{u}^{(j)} \mathbf{u}^{(j)\top}\big)$
  $\mathbf{a}^{(j)} \leftarrow \mathbf{u}^{(j)} / \sqrt{\boldsymbol{\sigma}^2 + \epsilon}$
else
  $\Sigma \leftarrow \frac{1}{G} \sum_j \mathbf{u}^{(j)} \mathbf{u}^{(j)\top}$
  $\mathbf{a}^{(j)} \leftarrow (\Sigma + \epsilon I)^{-1/2} \mathbf{u}^{(j)}$
end if
// Step 4: Econometric Aggregation
if $\mathcal{W} = \text{Sum}$ **then**
  $\tilde{A}^{(j)} \leftarrow \sum_{k=1}^K a_k^{(j)}$
else if $\mathcal{W} = \text{Nash}$ **then**
  $\tilde{A}^{(j)} \leftarrow \sum_{k=1}^K \log \sigma(a_k^{(j)})$
else
  $\tilde{A}^{(j)} \leftarrow -\log\big(\frac{1}{K} \sum_{k=1}^K e^{-a_k^{(j)}}\big)$
end if
Return $\{\tilde{A}^{(j)}\}_{j=1}^G$

Note on computational complexity Since the number of objectives is typically small, the $O(K^3)$ cost of covariance computation is negligible compared to the $O(N_{\text{params}})$ for backward passes.

6 Theoretical Results

This section provides formal characterizations of the two axes of GEOMA. We show that the econometric objectives form a unified spectrum of trade-offs, and we prove that geometric preconditioning resolves specific pathologies in the reward landscape. Full proofs are provided in Appendix A.

6.1 Econometric Aggregation: The Generalized Mean Spectrum

The utilitarian (Sum), Nash (MNW), and egalitarian (SoftMin) aggregators can be unified under the family of generalized p-means. Let $\mathbf{z} \in \mathbb{R}^K_{>0}$ be a vector of strictly positive utilities. The normalized power mean is defined as $M_p(\mathbf{z}) = \left(\frac{1}{K} \sum_{k=1}^K z_k^p\right)^{1/p}$, with $M_0(\mathbf{z})$ defined by continuity as the geometric mean. Figure 2 visualizes this continuum, demonstrating how the choice of p shifts the optimization priority from total aggregate performance to worst-case robustness.

Theorem 6.1 (Special Cases and Limit Objectives)**.** *The GEOMA aggregators correspond to specific limits of the p-mean family:*

*1. (**Utilitarian**) $M_1(\mathbf{z}) = \frac{1}{K} \sum_{k=1}^K z_k$ assumes perfect substitutability.*

2. ***(Nash)*** $\lim_{p\to 0} M_p(\mathbf{z}) = \exp(\frac{1}{K}\sum_k \log z_k)$. *Maximizing this is equivalent to maximizing the sum of log-utilities.*

3. ***(Egalitarian)*** $\lim_{p\to-\infty} M_p(\mathbf{z}) = \min_k z_k$, *prioritizing the worst-case objective.*

Proof sketch. Standard limit evaluation using L'Hôpital's rule. See Appendix B.1.1.

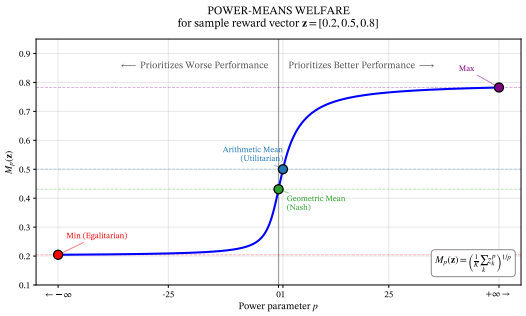

Figure 2: **The Econometric Aggregation Spectrum.** The GEOMA framework unifies standard multi-objective aggregators under the generalized p-means family, $M_p(\mathbf{z})$. Plotting the aggregate scalar value for a fixed reward vector $\mathbf{z} = [0.2, 0.5, 0.8]$ as a function of the power parameter p reveals the inherent trade-offs: $p = 1$ yields the Utilitarian (Sum) objective, $p \to 0$ recovers the Nash (Geometric Mean) objective, and $p \to -\infty$ yields the strict Egalitarian (Min) objective. By decreasing p, the optimization focus shifts to penalizing the worst-performing objectives.

6.2 Geometric Preconditioning

We now analyze the action of the preconditioner $\mathbf{a} = P\mathbf{u}$, assuming centered rewards with covariance Σ. The primary goal of GEOMA's geometric axis is to correct for ill-conditioning in the reward landscape.

Theorem 6.2 (Population Whitening and Isotropy)**.** *Let* $\mathbf{u}$ *be the centered reward vector with covariance* $\Sigma \succ 0$*. Under Covariance Sphering, the preconditioned advantages* $\mathbf{a} = \Sigma^{-1/2}\mathbf{u}$ *satisfy:*

$$\mathbb{E}[\mathbf{a}] = \mathbf{0} \quad \textit{and} \quad \mathrm{Cov}(\mathbf{a}) = I. \tag{9}$$

Remark. This guarantees that the preconditioned reward space is perfectly isotropic. Unlike Diagonal Sphering (GDPO), which only normalizes the coordinate axes, Covariance Sphering ensures that variance is standardized along *all* directions, completely eliminating cross-reward correlation.

We can further analyze the impact of this decorrelation by examining how $\Sigma^{-1/2}$ behaves under extreme dependency structures. To handle rank-deficiency, we define the pseudo-inverse square-root $\Sigma^{\dagger-1/2}$ by projecting out eigenspaces with eigenvalues $\nu_i \leq \tau$.

Proposition 6.3 (Nullspace Projection and Redundancy)**.** *Let* $\mathbf{a} = \Sigma^{\dagger-1/2}\mathbf{u}$*. Then for any redundant direction* $\mathbf{v} \in \mathrm{Null}(\Sigma)$*, we have* $\mathbf{v}^\top\mathbf{a} = 0$*.*

This formally demonstrates that GEOMA eliminates double-counting: the redundant difference between perfectly correlated rewards lies in the nullspace and is projected out, treating identical rewards as a single factor.

7 Experiments

7.1 Tool Calling

Training Dynamics and Sample Efficiency. We analyze the step-wise training curves on the Tool Calling task (Figure 3), which requires balancing a difficult primary objective (Correctness) with a strict constraint (Formatting). The curves expose the fragility of linear scalarization: the standard GRPO baseline (Identity-Sum, gray) exhibits severe reward hacking, entirely sacrificing the format constraint to maximize correctness.

In contrast, GEOMA actively prevents this collapse. The combination of geometric preconditioning and econometric aggregation forces the policy to respect the constraint early in training. Notably, **Batch Covariance Sphering** demonstrates exceptional sample efficiency. Methods utilizing Batch-Covariance (e.g., with Nash or Sum) achieve near-perfect format compliance within the first 15 steps—significantly faster than GDPO (Diagonal-Sum, teal). This confirms that resolving reward-space ill-conditioning not only stabilizes multi-objective alignment but drastically accelerates convergence.

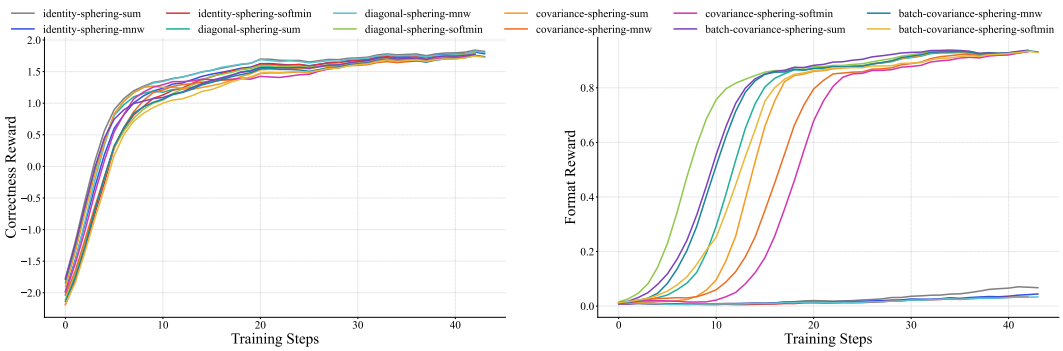

Figure 3: **Training Dynamics on the Tool Calling Task (Qwen2.5-1.5B-Instruct).** Evolution of Correctness (left) and Format (right) rewards. Standard scalarization (Identity-Sum/GRPO) collapses on formatting. GEOMA variants, particularly Batch-Covariance with Nash, converge rapidly on the format constraint without sacrificing correctness.

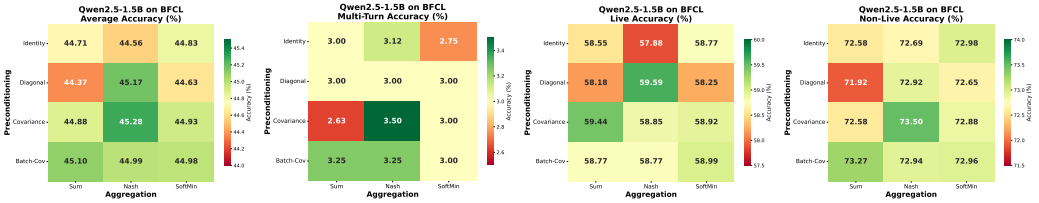

Figure 4: **BFCL Ablation (Qwen2.5-1.5B).** We compare GEOMA against baselines GRPO (Identity, Sum) and GDPO (Diagonal, Sum). Covariance Sphering combined with Nash aggregation achieves the highest performance, validating the joint geometric and econometric correction. Preconditioning gains are modest here as training rewards are naturally orthogonal.

Tool Calling on BFCL. We fine-tune Qwen2.5-1.5B on the Berkeley Function Calling Leaderboard (Patil et al., 2025), where models must balance JSON formatting with semantic correctness. Linear scalarization risks reward hacking, prioritizing formatting over accuracy. As shown in Figure 4, GEOMA mitigates this collapse. Covariance Sphering resolves reward correlations, raising average accuracy from 44.37% (GDPO) to 45.10% (Batch-Cov). Nash aggregation further penalizes correctness failures, leading the combined Covariance-Nash configuration to a global maximum of 45.28%, with notable gains on complex subsets.

Key Takeaway: Diagonal normalization alone is insufficient for constraint-heavy tasks. Addressing geometric correlation (Covariance) and econometric trade-offs (Nash) simultaneously prevents objective collapse and yields superior tool-use alignment.

7.2 Mathematical Reasoning

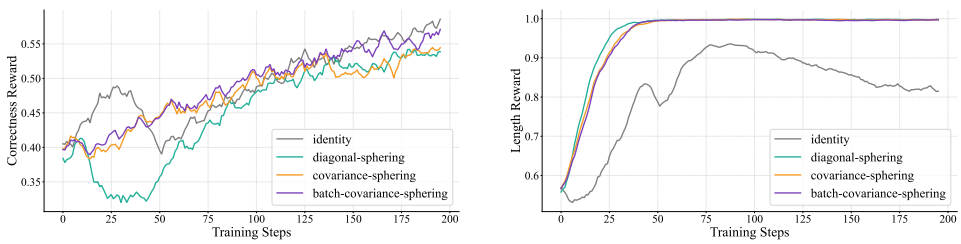

Figure 5: **Anti-Correlated Math Reasoning.** We track Correctness (left) and Length (right). GRPO (Identity) ignores length; GDPO (Diagonal) enforces length but causes correctness collapse. Only Covariance Sphering optimizes both, achieving perfect length by step 50 with steady correctness gains.

Table 2: **Preconditioning with Anti-Correlated Objectives.** Using Utilitarian (Sum) aggregation with a length penalty (¿4k tokens), GRPO (Identity) ignores constraints via verbosity hacking. GDPO (Diagonal) fails to strictly enforce limits on hard tasks. GEOMA (Covariance) resolves this trade-off, achieving peak accuracy on AIME-25 and strict length adherence for a 3× efficiency gain.

	Maths-500		AIME-25		AMC-24		Minerva	
Preconditioner (with Sum)	**Acc (↑)**	**Tok (↓)**	**Acc (↑)**	**Tok (↓)**	**Acc (↑)**	**Tok (↓)**	**Acc (↑)**	**Tok (↓)**
Base Model	74.6	7340	22.6	21360	38.7	16110	22.3	9961
Identity (GRPO)	**85.2**	**3410**	27.3	11990	**61.8**	6903	**26.5**	5087
Diagonal (GDPO)	83.2	**2063**	**28.0**	6767	52.4	5111	23.5	**1646**
Covariance (GEOMA)	84.0	**1713**	24.6	**3137**	59.1	**3221**	24.7	**1881**
Batch-Covariance (GEOMA)	83.9	**1706**	**28.0**	**3475**	57.7	**3113**	23.9	**1778**

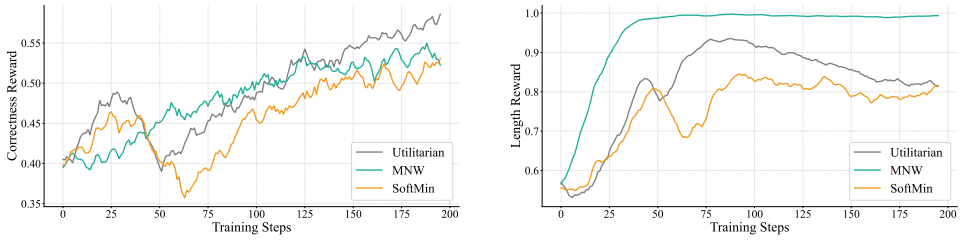

Figure 6: **Econometric Aggregation.** Fixing the preconditioner to Identity, we vary the aggregator. Utilitarian (Sum) sacrifices length for correctness. MNW (Nash) perfectly enforces the length constraint but slightly dampens correctness. SoftMin is too conservative, struggling on both. MNW acts as a strict geometric barrier, preventing constraint violations.

Mathematical Reasoning and Anti-Correlated Objectives. We evaluate GEOMA on mathematical reasoning, where Correctness (solving problems) and Length (token limits) are inherently antagonistic. As shown in Figure 5, standard linear scalarization (GRPO) collapses, ignoring the length constraint entirely. Diagonal Sphering (GDPO) enforces length but suffers a reasoning collapse early in training due to gradient conflict. Covariance Sphering resolves this antagonism by whitening the reward space, amplifying the signal for simultaneous improvement. This allows GEOMA to achieve rapid length compliance without degrading reasoning performance.

Key Takeaway: When rewards are anti-correlated, diagonal normalization oscillates between objectives. Covariance Sphering actively resolves the geometric conflict, allowing the model to navigate the Pareto frontier and optimize conflicting goals simultaneously.

Pareto-Efficient Reasoning and Verbosity Control. Table 2 isolates the geometric axis (fixed Utilitarian sum) to analyze the trade-off between reasoning accuracy and verbosity bias (penalized above 4,000 tokens). Baselines struggle with this anti-correlation: Identity (GRPO) maximizes accuracy but ignores the constraint (∼12k tokens), while Diagonal Sphering (GDPO) fails to fully suppress verbosity (6,767 tokens). In contrast, **Covariance Sphering** resolves the geometric conflict. Batch-Covariance matches peak accuracy (28.0% on AIME) while strictly adhering to the limit (3,475 tokens), delivering state-of-the-art reasoning at 3× greater token efficiency.

Econometric Immunity to Reward Hacking. Figure 6 isolates econometric effects by comparing aggregators under Identity preconditioning. While the Utilitarian objective permits reward hacking (sacrificing length for correctness), MNW acts as a strict barrier. By scaling gradients inversely with utility (Lemma B.8), MNW enforces the length constraint early and maintains it, confirming that welfare aggregators provide structural immunity to metric gaming.

Discussion and Conclusion In this work, we proposed **GEOMA**, a novel framework for geometric preconditioning and econometric aggregation for multi-reward alignment. Theoretically, we proved that Covariance Sphering resolves signal redundancy and that Nash aggregation provides structural immunity to reward hacking. Empirically, we demonstrated that these principles translate to significant gains in efficiency and constraint satisfaction on tool-use and reasoning tasks.

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. *arXiv preprint arXiv:2402.14740*, 2024.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with high-quality feedback. *arXiv preprint arXiv:2310.01377*, 2023.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

Taneesh Gupta, Rahul Madhavan, Xuchao Zhang, Chetan Bansal, and Saravan Rajmohan. Refa: Reference free alignment for multi-preference optimization. *arXiv preprint arXiv:2412.16378*, 2024.

Taneesh Gupta, Rahul Madhavan, Xuchao Zhang, Chetan Bansal, and Saravan Rajmohan. Ampo: Active multi-preference optimization. *arXiv preprint arXiv:2502.18293*, 2025a.

Taneesh Gupta, Rahul Madhavan, Xuchao Zhang, Nagarajan Natarajan, Chetan Bansal, and Saravan Rajmohan. Multi-preference optimization: Generalizing dpo via set-level contrasts, 2025. *URL https://arxiv. org/abs/2412.04628*, 2025b.

Conor F Hayes, Roxana Rădulescu, Eugenio Bargiacchi, Johan Källström, Matthew Macfarlane, Mathieu Reymond, Timothy Verstraeten, Luisa M Zintgraf, Richard Dazeley, Fredrik Heintz, et al. A practical guide to multi-objective reinforcement learning and planning. *arXiv preprint arXiv:2103.09568*, 2021.

Jiwoo Hong, Noah Lee, and James Thorne. Orpo: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 11170–11189, 2024.

Zhengyu Hu, Linxin Song, Jieyu Zhang, Zheyuan Xiao, Tianfu Wang, Zhengyu Chen, Nicholas Jing Yuan, Jianxun Lian, Kaize Ding, and Hui Xiong. Explaining length bias in llm-based preference evaluations. *arXiv preprint arXiv:2407.01085*, 2024.

Dahyun Kim, Yungi Kim, Wonho Song, Hyeonwoo Kim, Yunsu Kim, Sanghoon Kim, and Chanjun Park. sdpo: Don't use your data all at once. *arXiv preprint arXiv:2403.19270*, 2024.

Hyeonji Kim, Sujeong Oh, and Sanghack Lee. Mitigating length bias in rlhf through a causal lens. *arXiv preprint arXiv:2511.12573*, 2025.

Nathan Lambert. Reinforcement learning from human feedback. *arXiv preprint arXiv:2504.12501*, 2025.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*, 2024.

Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcak, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, et al. Gdpo: Group reward-decoupled normalization policy optimization for multi-reward rl optimization. *arXiv preprint arXiv:2601.05242*, 2026.

Tianqi Liu, Wei Xiong, Jie Ren, Lichang Chen, Junru Wu, Rishabh Joshi, Yang Gao, Jiaming Shen, Zhen Qin, Tianhe Yu, et al. Rrm: Robust reward model training mitigates reward hacking. *arXiv preprint arXiv:2409.13156*, 2024.

Prasanta Chandra Mahalanobis. On the generalized distance in statistics. *Sankhyā: The Indian Journal of Statistics, Series A (2008-)*, 80:S1–S7, 2018.

Jeonghoon Mo and Jean Walrand. Fair end-to-end window-based congestion control. *IEEE/ACM Transactions on networking*, 8(5):556–567, 2002.

John F Nash et al. The bargaining problem. *Econometrica*, 18(2):155–162, 1950.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Shishir G. Patil, Huanzhi Mao, Charlie Cheng-Jie Ji, Fanjia Yan, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. The berkeley function calling leaderboard (bfcl): From tool use to agentic evaluation of large language models. In *Forty-second International Conference on Machine Learning*, 2025.

Cheng Qian, Emre Can Acikgoz, Qi He, Hongru Wang, Xiusi Chen, Dilek Hakkani-Tür, Gokhan Tur, and Heng Ji. Toolrl: Reward is all tool learning needs, 2025. URL https://arxiv.org/abs/2504.13958.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. Verbosity bias in preference labeling by large language models. *arXiv preprint arXiv:2310.10076*, 2023.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015.

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

Amartya Sen. *Collective choice and social welfare: Expanded edition*. Penguin UK, 2017.

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.

Wei Shen, Rui Zheng, Wenyu Zhan, Jun Zhao, Shihan Dou, Tao Gui, Qi Zhang, and Xuan-Jing Huang. Loose lips sink ships: Mitigating length bias in reinforcement learning from human feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2859–2873, 2023.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, EuroSys '25, pp. 1279–1297. ACM, March 2025. doi: 10.1145/3689031.3696075. URL http://dx.doi.org/10.1145/3689031.3696075.

Pragya Srivastava, Harman Singh, Rahul Madhavan, Gandharv Patil, Sravanti Addepalli, Arun Suggala, Rengarajan Aravamudhan, Soumya Sharma, Anirban Laha, Aravindan Raghuveer, et al. Robust reward modeling via causal rubrics. *arXiv preprint arXiv:2506.16507*, 2025.

Aviv Tamar, Yinlam Chow, Mohammad Ghavamzadeh, and Shie Mannor. Policy gradient for coherent risk measures. *Advances in neural information processing systems*, 28, 2015.

Shusheng Xu, Wei Fu, Jiaxuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu. Is dpo superior to ppo for llm alignment? a comprehensive study. *arXiv preprint arXiv:2404.10719*, 2024.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Kangwen Zhao, Jianfeng Cai, Jinhua Zhu, Ruopei Sun, Dongyun Xue, Wengang Zhou, Li Li, and Houqiang Li. Bias fitting to mitigate length bias of reward model in rlhf. *arXiv preprint arXiv:2505.12843*, 2025a.

Yuzhong Zhao, Yue Liu, Junpeng Liu, Jingye Chen, Xun Wu, Yaru Hao, Tengchao Lv, Shaohan Huang, Lei Cui, Qixiang Ye, et al. Geometric-mean policy optimization. *arXiv preprint arXiv:2507.20673*, 2025b.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

SUPPLEMENTARY MATERIALS

A RELATED WORK

RLHF as post-training for LLMs. A dominant paradigm for aligning instruction-following LLMs is reinforcement learning from human feedback (RLHF), popularized by Ouyang et al. (2022). Most RLHF implementations borrow policy-gradient machinery from continuous-control RL, with Proximal Policy Optimization (PPO) (Schulman et al., 2017) as a common default due to its stability, and earlier trust-region views (Schulman et al., 2015) motivating KL-style regularization and conservative updates. Recent work argues that several PPO components are unnecessary in the LLM post-training regime and revisits simpler REINFORCE-style variants (Kim et al., 2024), while open-source systems and algorithms continue to scale RL post-training to long-chain-of-thought and reasoning settings (Yu et al., 2025). For a broader consolidation of the RLHF pipeline (instruction tuning, reward modeling, rejection sampling, RL, and direct alignment), see Lambert (2025).

Direct alignment and preference optimization (RL-free / offline). Complementary to online RLHF, a large line of work replaces the explicit reward-model + RL loop with objectives derived directly from preference data. Direct Preference Optimization (DPO) (Rafailov et al., 2024) is a widely-used instance of this approach, framing alignment as a logistic classification objective whose optimum corresponds to a KL-regularized reward maximization solution. These methods are typically attractive for their simplicity and reduced online sampling costs, and are often used as strong baselines alongside PPO-style RLHF (Xu et al., 2024; Ahmadian et al., 2024). Our paper's setting (multiple rubric rewards / advantages) intersects with this literature because many "single-reward" objectives in practice already include multiple implicit components (e.g., correctness vs. style vs. safety, plus explicit penalties/regularizers) and thus inherit multi-objective tradeoffs.

Beyond pairwise feedback: multi-response and set-based preference signals. Standard preference optimization typically assumes pairwise comparisons, but real post-training pipelines often generate multiple candidates per prompt and can exploit richer supervision (rankings, sets, or group-wise signals). In our work, we explicitly build on this "multi-candidate" reality. Closest to our setup are recent proposals that formalize optimization over multiple sampled responses and preferences (e.g., MPO and AMPO (Gupta et al., 2025b;a)), and related multi-candidate data sources such as UltraFeedback (Cui et al., 2023), which provide multiple responses and fine-grained feedback to support stronger preference learning and reward modeling.

Multi-reward / multi-objective optimization and aggregation. When alignment is driven by multiple reward dimensions (e.g., correctness, conciseness, safety, formatting), a core design choice is how to aggregate a reward vector into a scalar advantage for policy updates. This is the central point of contact with multi-objective reinforcement learning (MORL), where linear scalarization is common but often insufficient under conflicting objectives (Hayes et al., 2021). Aggregation choices can be motivated by welfare economics and fairness: utilitarian objectives (sum), max-min/egalitarian objectives, and Nash-style objectives (Nash et al., 1950; Mo & Walrand, 2002). Risk-sensitive RL offers another principled "tail-focusing" lens via coherent risk measures (including CVaR), which bias optimization toward worst-case or lower-tail outcomes rather than means. Our paper's aggregators (utilitarian / Nash-style / max-min and smooth relaxations) fit naturally into this broader set of scalarization and risk/fairness principles (Tamar et al., 2015).

Conflicting rewards, length/verbosity effects, and evaluation artifacts. Multi-reward alignment is particularly sensitive when rewards are antagonistic (e.g., conciseness vs. correctness), because aggregation can implicitly over-weight one dimension or amplify spurious directions. This connects directly to well-documented length and verbosity biases in preference labeling and LLM-as-a-judge evaluation: judges can prefer longer answers even when content quality is unchanged (Zheng et al., 2023; Hu et al., 2024; Saito et al., 2023), and benchmarks have introduced explicit length

controls/debiasing to mitigate these effects (Dubois et al., 2024). At the reward-model level, length bias has been studied both diagnostically and through mitigation frameworks (Shen et al., 2023; Zhao et al., 2025a) as well as causal frameworks (Srivastava et al., 2025; Kim et al., 2025; Liu et al., 2024). Since many "multi-reward" designs include explicit length penalties or conciseness rewards, these biases are not just an evaluation issue—they can become part of the learning signal and interact nontrivially with reward aggregation (Hu et al., 2024). Finally, the reliability of reward models under distribution shift and subtle preference failures is an active area, with benchmarks such as RewardBench emphasizing systematic RM evaluation across chat, reasoning, and safety (Lambert et al., 2024).

Placement of GEOMA in the literature context. The GEOMA framework bridges the operational machinery of group-based RLHF (Shao et al., 2024) with the theoretical formalisms of Multi-Objective RL (MORL) and Welfare Economics. While prior work addresses multi-objective alignment either through regularization (Gupta et al., 2024) or simple sum, GEOMA explicitly treats aggregation as a statistical and econometric problem. We generalize the recent variance-normalization baseline, GDPO (Liu et al., 2026), by identifying it as a diagonal subset of a broader covariance-aware geometric space. Furthermore, while risk-sensitive RL addresses tail behavior, GEOMA directly imports Nash and SoftMin welfare objectives to combat specific alignment pathologies, such as verbosity bias and metric hacking, providing a unified design space for robust multi-reward aggregation.

B Full Theoretical Results

This section provides formal characterizations of the two axes of GEOMA: (i) *econometric aggregation*, which specifies how improvements across rubrics are traded off, and (ii) *geometric preconditioning*, which standardizes second-moment geometry in reward space. Throughout, we use $\mathbf{u} \in \mathbb{R}^K$ for centered reward vectors and $\mathbf{a} = P\mathbf{u}$ for preconditioned coordinates, where $P \in \{I, (D+\varepsilon I)^{-1/2}, (\Sigma+\varepsilon I)^{-1/2}\}$.

B.1 Econometric Aggregation: Welfare Objectives

B.1.1 Power-mean (p-means) welfare and its limits

Definition B.1 (Power-mean welfare)**.** Let $\mathbf{z} \in \mathbb{R}^K_{>0}$ and $p \in \mathbb{R} \setminus \{0\}$. Define the (normalized) power mean

$$M_p(\mathbf{z}) \;:=\; \left(\frac{1}{K}\sum_{k=1}^{K} z_k^p\right)^{1/p}. \tag{10}$$

Define the $p = 0$ case by continuity as

$$M_0(\mathbf{z}) \;:=\; \exp\left(\frac{1}{K}\sum_{k=1}^{K} \log z_k\right), \tag{11}$$

the geometric mean.

Theorem B.2 (Special cases and limit objectives)**.** *For any* $\mathbf{z} \in \mathbb{R}^K_{>0}$*:*

1. ***(Utilitarian)*** $M_1(\mathbf{z}) = \frac{1}{K}\sum_{k=1}^{K} z_k$ *(arithmetic mean).*
2. ***(Nash)*** $\lim_{p\to 0} M_p(\mathbf{z}) = M_0(\mathbf{z}) = \exp\left(\frac{1}{K}\sum_{k=1}^{K}\log z_k\right)$ *(geometric mean). Equivalently, maximizing* $M_0(\mathbf{z})$ *is equivalent to maximizing* $\sum_{k=1}^{K}\log z_k$*.*
3. ***(Egalitarian)*** $\lim_{p\to-\infty} M_p(\mathbf{z}) = \min_k z_k$*.*

Proof. (1) Immediate from the definition with $p = 1$.

(2) Let $g(p) := \log M_p(\mathbf{z}) = \frac{1}{p}\log\left(\frac{1}{K}\sum_{k=1}^{K} e^{p\log z_k}\right)$. Define $m := \frac{1}{K}\sum_{k=1}^{K}\log z_k$ and write $\log z_k = m + \delta_k$ with $\frac{1}{K}\sum_k \delta_k = 0$. Then

$$\frac{1}{K}\sum_{k=1}^{K} e^{p\log z_k} = e^{pm}\cdot\frac{1}{K}\sum_{k=1}^{K} e^{p\delta_k}.$$

Hence

$$g(p) = m + \frac{1}{p}\log\left(\frac{1}{K}\sum_{k=1}^{K} e^{p\delta_k}\right).$$

As $p \to 0$, we use $\exp(p\delta_k) = 1 + p\delta_k + O(p^2)$, so

$$\frac{1}{K}\sum_{k=1}^{K} e^{p\delta_k} = 1 + p\cdot\frac{1}{K}\sum_{k=1}^{K}\delta_k + O(p^2) = 1 + O(p^2).$$

Therefore $\log\left(\frac{1}{K}\sum_{k=1}^{K} e^{p\delta_k}\right) = O(p^2)$ and thus $\frac{1}{p}\log(\cdot) \to 0$, yielding $\lim_{p\to 0} g(p) = m$. Exponentiating gives $\lim_{p\to 0} M_p(\mathbf{z}) = \exp(m) = M_0(\mathbf{z})$. The equivalence to maximizing $\sum_k \log z_k$ follows since $\log$ is strictly increasing.

(3) Let $m := \min_k z_k$ and $M := \max_k z_k$. For $p < 0$, we have $z_k^p \in [M^p, m^p]$ (since $x \mapsto x^p$ is decreasing on $\mathbb{R}_{>0}$). Thus

$$\frac{1}{K}\sum_{k=1}^{K} z_k^p \in [M^p, m^p],$$

and so

$$\left(\frac{1}{K}\sum_{k=1}^{K} z_k^p\right)^{1/p} \in [m, M].$$

More sharply, for any $p < 0$,

$$\frac{1}{K}m^p \le \frac{1}{K}\sum_{k=1}^{K} z_k^p \le m^p,$$

because m^p is the *largest* term among $\{z_k^p\}_{k=1}^K$. Taking $1/p < 0$ powers reverses inequalities:

$$m \le \left(\frac{1}{K}\sum_{k=1}^{K} z_k^p\right)^{1/p} \le K^{1/p}m.$$

Since $K^{1/p} \to 1$ as $p \to -\infty$, we conclude $M_p(\mathbf{z}) \to m = \min_k z_k$. □

Theorem B.3 (Power-mean monotonicity in p)**.** *For any $\mathbf{z} \in \mathbb{R}^K_{>0}$ and $p < q$, we have $M_p(\mathbf{z}) \le M_q(\mathbf{z})$.*

Proof. Fix $p < q$ and set $t := q/p > 1$. Define $x_k := z_k^p > 0$. Then

$$M_q(\mathbf{z})^q = \frac{1}{K}\sum_{k=1}^{K} z_k^q = \frac{1}{K}\sum_{k=1}^{K}(z_k^p)^{q/p} = \frac{1}{K}\sum_{k=1}^{K} x_k^t.$$

Also $M_p(\mathbf{z})^p = \frac{1}{K}\sum_{k=1}^{K} x_k$. Since $t > 1$, the function $\phi(x) = x^t$ is convex on $\mathbb{R}_{>0}$, and Jensen's inequality implies

$$\phi\left(\frac{1}{K}\sum_{k=1}^{K} x_k\right) \le \frac{1}{K}\sum_{k=1}^{K}\phi(x_k)$$

$$\Rightarrow \quad \left(\frac{1}{K}\sum_{k=1}^{K} x_k\right)^t \le \frac{1}{K}\sum_{k=1}^{K} x_k^t.$$

Substituting back gives

$$\begin{aligned} \left(M_p(\mathbf{z})^p\right)^{q/p} &\le M_q(\mathbf{z})^q \\ \Rightarrow \quad M_p(\mathbf{z})^q &\le M_q(\mathbf{z})^q \\ \Rightarrow \quad M_p(\mathbf{z}) &\le M_q(\mathbf{z}), \end{aligned}$$

since $q > 0$ or $q < 0$ both preserve order under taking the $1/q$ power on positive reals. □

Mapping to GEOMA. The p-means family is defined on *positive* inputs. In GEOMA, we implement Nash-style objectives by applying a positive utility map $\psi : \mathbb{R} \to \mathbb{R}_{>0}$ to possibly signed coordinates a_k, e.g. $\psi(a_k) = \sigma(a_k)$, and then aggregating utilities.

B.1.2 NASH-STYLE (MNW) AGGREGATION

Definition B.4 (General MNW objective)**.** Let $\psi : \mathbb{R} \to \mathbb{R}_{>0}$ be a positive utility map. Define

$$\mathcal{G}_{\mathrm{MNW}}(\mathbf{a}) \;:=\; \sum_{k=1}^{K} \log \psi(a_k). \tag{12}$$

Theorem B.5 (Equivalence to geometric-mean welfare)**.** *Let $z_k := \psi(a_k) > 0$. Then*

$$\mathcal{G}_{\mathrm{MNW}}(\mathbf{a}) = K \cdot \log M_0(\mathbf{z}) = \log \left(\prod_{k=1}^{K} z_k \right). \tag{13}$$

In particular, maximizing $\mathcal{G}_{\mathrm{MNW}}(\mathbf{a})$ is equivalent to maximizing the product $\prod_k \psi(a_k)$ and to maximizing the geometric mean $M_0(\mathbf{z})$.

Proof. By definition,

$$\mathcal{G}_{\mathrm{MNW}}(\mathbf{a}) = \sum_{k=1}^{K} \log z_k = \log \left(\prod_{k=1}^{K} z_k \right).$$

Also $M_0(\mathbf{z}) = \exp\left(\frac{1}{K} \sum_k \log z_k\right)$, hence $\sum_k \log z_k = K \log M_0(\mathbf{z})$. □

Theorem B.6 (Monotonicity and concavity of MNW)**.** *Assume ψ is increasing and $\log \psi$ is concave on $\mathbb{R}$. Then $\mathcal{G}_{\mathrm{MNW}}(\mathbf{a})$ is (i) coordinatewise increasing and (ii) concave in $\mathbf{a}$.*

Proof. (i) Since ψ is increasing, $\log \psi$ is increasing as a composition of increasing functions, hence each term $\log \psi(a_k)$ is increasing in a_k, and their sum is coordinatewise increasing.

(ii) Each function $a_k \mapsto \log \psi(a_k)$ is concave by assumption, and a sum of concave functions is concave. □

Remark B.7. We do not claim that training with MNW aggregation solves the Nash bargaining problem in an axiomatic sense since the utility vectors under a neural policy may not follow these assumptions. Even so, the Nash objective provides a principled aggregation functional with well-understood behavior that discourages collapse of any single rubric.

Lemma B.8 (Gradient reweighting for $\psi = \sigma$)**.** *Let $\psi(z) = \sigma(z) = (1 + e^{-z})^{-1}$. Then*

$$\frac{\partial}{\partial a_k} \mathcal{G}_{\mathrm{MNW}}(\mathbf{a}) = \frac{d}{da_k} \log \sigma(a_k) = \sigma(-a_k). \tag{14}$$

Consequently, smaller (worse) coordinates receive larger marginal weight.

Proof. Compute $\frac{d}{dz} \log \sigma(z) = \frac{\sigma'(z)}{\sigma(z)}$. Since $\sigma'(z) = \sigma(z)(1 - \sigma(z))$, we obtain

$$\frac{d}{dz} \log \sigma(z) = 1 - \sigma(z) = \sigma(-z).$$

□

Remark (anti-collapse behavior and fairness interpretation). The Nash-style aggregator $\mathcal{G}_{\mathrm{MNW}}(\mathbf{a}) = \sum_k \log \psi(a_k)$ pays attention to all rubrics irrespective of their scale. When $\psi = \sigma$, Lemma B.8 yields $\partial \mathcal{G}_{\mathrm{MNW}}/\partial a_k = \sigma(-a_k)$, so coordinates with small (or negative) a_k receive near-unit marginal weight while coordinates with large a_k receive small marginal weight. This implies that, at a given iterate, gradient updates allocate disproportionately more effort to the worst-off objectives, whereas a linear sum assigns equal marginal weight everywhere. This is an *objective-induced* balance through marginal utilities.

B.1.3 Egalitarian objectives and SoftMin smoothing

Definition B.9 (SoftMin with temperature)**.** For $T > 0$, define

$$\mathrm{SoftMin}_T(\mathbf{a}) \;:=\; -T \log \left(\frac{1}{K} \sum_{k=1}^{K} e^{-a_k/T} \right). \tag{15}$$

GEOMA uses $T = 1$ unless otherwise stated.

Theorem B.10 (SoftMin approximates the egalitarian objective)**.** *For any* $\mathbf{a} \in \mathbb{R}^K$ *and* $T > 0$*,*

$$\min_k a_k \;\le\; \mathrm{SoftMin}_T(\mathbf{a}) \;\le\; \min_k a_k + T \log K. \tag{16}$$

Moreover, $\lim_{T \to 0^+} \mathrm{SoftMin}_T(\mathbf{a}) = \min_k a_k$.

Proof. Let $m := \min_k a_k$. Then $e^{-a_k/T} \le e^{-m/T}$ for all k, hence

$$\frac{1}{K} \sum_{k=1}^{K} e^{-a_k/T} \le e^{-m/T} \tag{17}$$

$$\Rightarrow \quad -T \log \left(\frac{1}{K} \sum_{k=1}^{K} e^{-a_k/T} \right) \ge -T \log(e^{-m/T}) \tag{18}$$

$$= m. \tag{19}$$

For the upper bound, at least one index achieves the minimum, so $\sum_k e^{-a_k/T} \ge e^{-m/T}$ and thus

$$\frac{1}{K} \sum_{k=1}^{K} e^{-a_k/T} \ge \frac{1}{K} e^{-m/T} \tag{20}$$

$$\Rightarrow \quad \mathrm{SoftMin}_T(\mathbf{a}) \le -T \log \left(\frac{1}{K} e^{-m/T} \right) \tag{21}$$

$$= m + T \log K. \tag{22}$$

Finally, the sandwich bounds imply $\mathrm{SoftMin}_T(\mathbf{a}) \to m$ as $T \to 0^+$ since $T \log K \to 0$. □

Remark (egalitarian as $p = -\infty$). Theorem B.2(3) shows that the egalitarian objective arises as the $p \to -\infty$ limit of the p-means welfare on positive utilities. Theorem B.10 shows that SoftMin is a smooth surrogate of the coordinatewise minimum (egalitarian) directly on $\mathbf{a}$.

B.2 Geometric Preconditioning

B.2.1 Whitening and ridge sphering

Proposition B.11 (Population whitening)**.** *Assume* $\mathbb{E}[\mathbf{u}] = \mathbf{0}$ *and* $\mathrm{Cov}(\mathbf{u}) = \Sigma \succ 0$*. If* $\mathbf{a} = \Sigma^{-1/2}\mathbf{u}$*, then* $\mathrm{Cov}(\mathbf{a}) = I$.

Proof. $\mathrm{Cov}(\mathbf{a}) = \mathbb{E}[\mathbf{a}\mathbf{a}^\top] = \Sigma^{-1/2}\, \mathbb{E}[\mathbf{u}\mathbf{u}^\top]\, \Sigma^{-1/2} = \Sigma^{-1/2} \Sigma \Sigma^{-1/2} = I$. □

Theorem B.12 (Spectrum under ridge sphering)**.** *Assume* $\mathrm{Cov}(\mathbf{u}) = \Sigma \succeq 0$ *with eigenvalues* $\nu_1, \ldots, \nu_K \ge 0$*. Define* $\mathbf{a} = (\Sigma + \varepsilon I)^{-1/2}\mathbf{u}$ *for* $\varepsilon > 0$*. Then* $\mathrm{Cov}(\mathbf{a})$ *has eigenvalues* $\nu_i/(\nu_i + \varepsilon) \in [0, 1)$.

Proof. Let $\Sigma = U\mathrm{diag}(\nu)U^\top$ with orthonormal U. Then

$$(\Sigma + \varepsilon I)^{-1/2} = U\,\mathrm{diag}((\nu_i + \varepsilon)^{-1/2})\,U^\top.$$

Hence

$$\begin{aligned}\mathrm{Cov}(\mathbf{a}) &= (\Sigma + \varepsilon I)^{-1/2}\Sigma(\Sigma + \varepsilon I)^{-1/2} \\ &= U\,\mathrm{diag}\left(\frac{\nu_i}{\nu_i + \varepsilon}\right)U^\top,\end{aligned}$$

so the eigenvalues are $\nu_i/(\nu_i + \varepsilon)$. □

B.2.2 Pseudo-inverse projection and redundancy

Definition B.13 (Pseudo-inverse inverse square-root)**.** Let $\Sigma = U\mathrm{diag}(\nu)U^\top$ with $\nu_i \geq 0$. For $\tau > 0$, define

$$\Sigma^{\dagger -1/2} \;:=\; U\,\mathrm{diag}\big(\mathbf{1}\{\nu_i > \tau\}\nu_i^{-1/2}\big)\,U^\top. \tag{23}$$

Proposition B.14 (Nullspace projection)**.** *Let* $\mathbf{a} = \Sigma^{\dagger -1/2}\mathbf{u}$*. Then for any* $\mathbf{v} \in \mathrm{Null}(\Sigma)$*, we have* $\mathbf{v}^\top\mathbf{a} = 0$*.*

Proof. If $\mathbf{v} \in \mathrm{Null}(\Sigma)$, then $\mathbf{v}$ lies in the span of eigenvectors with eigenvalue 0 (or $\leq \tau$ under thresholding). By Definition B.13, $\Sigma^{\dagger -1/2}$ has zero action on those eigen-directions, hence $\Sigma^{\dagger -1/2}\mathbf{v} = \mathbf{0}$. Therefore $\mathbf{v}^\top\mathbf{a} = \mathbf{v}^\top\Sigma^{\dagger -1/2}\mathbf{u} = (\Sigma^{\dagger -1/2}\mathbf{v})^\top\mathbf{u} = 0$. □

B.2.3 Anti-correlation geometry in $K = 2$

Proposition B.15 (Agreement vs trade-off directions)**.** *Let* $K = 2$ *and* $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ *with* $\rho \in (-1, 1)$*. Then eigenvectors are proportional to* $(1, 1)$ *and* $(1, -1)$ *with eigenvalues* $1 + \rho$ *and* $1 - \rho$*. If* $\rho < 0$*, the agreement direction* $(1, 1)$ *has smaller variance than the trade-off direction* $(1, -1)$*, and sphering scales the agreement direction by* $(1 + \rho + \varepsilon)^{-1/2}$*, which is larger than* $(1 - \rho + \varepsilon)^{-1/2}$*.*

Proof. A direct computation shows $\Sigma(1, 1)^\top = (1 + \rho)(1, 1)^\top$ and $\Sigma(1, -1)^\top = (1 - \rho)(1, -1)^\top$. For $\rho < 0$, we have $1 + \rho < 1 - \rho$. Ridge sphering scales eigen-direction i by $(\nu_i + \varepsilon)^{-1/2}$, hence the lower-variance eigen-direction receives larger multiplicative scaling. □

B.3 Additional Experimental Results

Training Setup Tool Use To benchmark GEOMA in a complex multi-objective environment, we replicate the tool-use setting established by ToolRL (Qian et al., 2025). The model is conditioned to interleave reasoning traces with external API interactions, strictly adhering to an XML schema. The optimization is driven by a composite reward signal comprising two often-conflicting objectives: a binary format reward, $\mathcal{R}_{\text{format}} \in \{0, 1\}$, which enforces structural integrity and tag hierarchy, and a scalar correctness reward, $\mathcal{R}_{\text{correct}} \in [-3, 3]$, which assesses the semantic fidelity of the generated API calls against the ground truth. We fine-tune the 1.5B parameter Qwen-2.5-Instruct model (Qwen et al., 2025) using the `verl` framework (Sheng et al., 2025) for 100 steps, utilizing a global batch size of 512 and four candidate rollouts per prompt. Downstream performance is evaluated on the Berkeley Function Call Leaderboard (BFCL-v3) (Patil et al., 2025), a comprehensive benchmark testing dynamic single- and multi-turn tool execution.

Table 3: Performance comparison on mathematical benchmarks. Results are grouped by preconditioning strategy to highlight the impact of econometric aggregators.

		Maths-500		AIME25		AMC24		Minerva	
Preconditioning	**Aggregator**	**Acc**	**Avg Tok**	**Acc**	**Avg Tok**	**Acc**	**Avg Tok**	**Acc**	**Avg Tok**
Baseline	-	74.6 ± 1.5	7340.1	22.6 ± 3.2	21360.1	38.7 ± 5.7	16110.8	22.3 ± 0.3	9961.3
Identity	Sum	85.2 ± 0.9	3410.1	27.3 ± 1.3	11990.1	61.8 ± 3.2	6903.5	26.5 ± 0.6	5087.4
	Nash	82.3 ± 1.0	1523.4	18.7 ± 1.6	2705.2	52.9 ± 5.5	2502.4	24.8 ± 1.2	1632.3
	Softmin	83.7 ± 0.3	3426.9	24.6 ± 1.6	11580.2	55.1 ± 4.3	7212.3	26.2 ± 1.8	4281.3
Diag. Sphering	Sum	83.2 ± 1.1	2063.6	28.0 ± 6.0	6767.1	52.4 ± 6.0	5111.2	23.5 ± 0.9	1646.9
	Nash	82.2 ± 0.0	1711.4	23.3 ± 2.0	3792.3	60.0 ± 3.0	3260.5	24.4 ± 0.9	1446.7
	Softmin	83.5 ± 0.8	1667.2	26.7 ± 3.0	4023.1	60.0 ± 5.0	3170.5	23.5 ± 0.5	1561.3
Cov. Sphering	Sum	84.0 ± 0.9	1713.3	24.6 ± 5.0	3137.4	59.1 ± 6.0	3221.5	24.7 ± 1.7	1881.4
	Nash	83.8 ± 0.6	1693.8	21.9 ± 5.0	3293.9	52.8 ± 4.0	2921.2	23.6 ± 0.9	1770.6
	Softmin	83.5 ± 0.9	1712.7	24.0 ± 3.0	3590.5	57.3 ± 6.0	2956.1	23.8 ± 1.2	1506.8
Batch Cov. Sph.	Sum	83.9 ± 1.0	1706.9	28.0 ± 4.0	3475.8	57.7 ± 4.5	3113.2	23.9 ± 0.9	1778.3
	Softmin	84.3 ± 0.7	1656.7	23.4 ± 4.2	2956.3	55.1 ± 6.0	2751.6	26.3 ± 1.3	1955.2

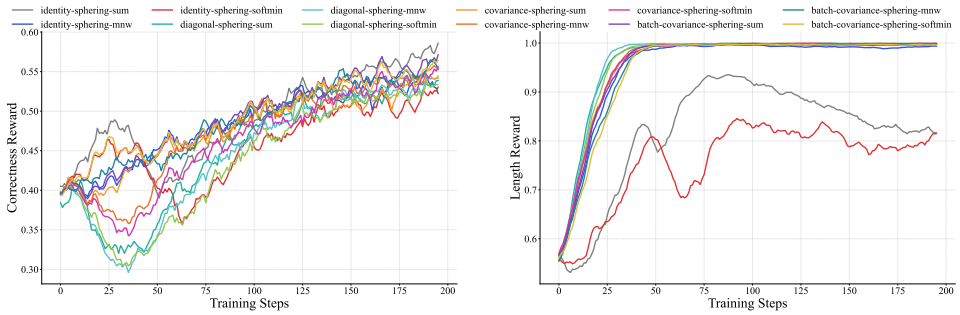

Figure 7: Training reward margin analysis on the DeepScaleR-Preview dataset using DeepSeek-R1-Distill-Qwen-1.5B. We visualize the training dynamics of Correctness Reward (left) and Length Reward (right) over 500 steps. The curves compare the convergence behavior of our proposed geometric pre-conditioners and econometric aggregators