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 convert vector-valued rewards into scalar advantages. Recent methods such as GDPO address scale discrepancies via per-dimension normalization, but they treat reward coordinates as uncorrelated. Under a sum, redundant objectives are then double-counted. For anti-correlated rewards, per-dimension normalization leaves most of the variance in the imbalance between rewards, which aggregators such as Nash welfare and SoftMin penalize. Any linear sum also lets the policy trade a hard objective for an easy one. 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 relate the aggregators to power means, characterize how each preconditioner changes the covariance of the rewards, and show that under a sum every preconditioner reduces to a choice of reward weights. On four mathematical reasoning benchmarks, covariance sphering with batch statistics and Sum aggregation raises mean accuracy over GDPO from 46.8% to 48.4% while reducing mean response length from 3,897 to 2,518 tokens. On tool calling, most GEOMA configurations learn a format constraint that GRPO does not learn, and with Nash aggregation only covariance sphering learns it.

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). Policy optimization with these signals requires aggregating the reward vector into a scalar advantage.

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. Naïve summation therefore suffers from *scale dominance*, where high-variance rewards overpower low-variance ones, and *redundancy*, where correlated metrics (e.g., correctness and reasoning) are double-counted. When an auxiliary objective is anti-correlated with the primary task (Hong et al.,

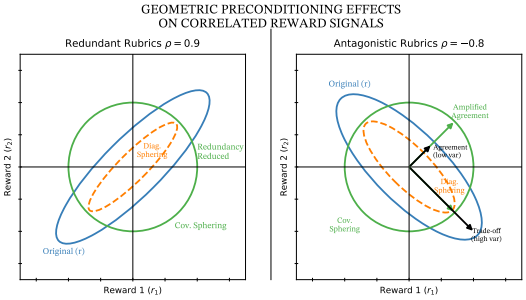


Figure 1: **Covariance sphering decorrelates rewards that diagonal scaling leaves correlated. Left (Redundancy, $\rho = 0.9$):** after diagonal sphering (GDPO) the rewards stay correlated; covariance sphering makes the distribution isotropic. **Right (Antagonism, $\rho = -0.8$):** the agreement direction (both rewards improve) has low variance and the trade-off direction has high variance; covariance sphering gives both unit variance. This changes the inputs to aggregators that penalize imbalance between rewards, such as Nash and SoftMin; under a sum, the preconditioner changes only the reward weights (Remark 4.1).

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.

2024; Yu et al., 2025), most of the reward variance lies along the trade-off direction, in which one reward improves at the expense of the other, and the agreement direction, in which both improve, has low variance.

Recent methods like GDPO (Liu et al., 2026) address scale discrepancies via per-dimension normalization (diagonal rescaling). Diagonal rescaling leaves the correlation between rewards unchanged, so a sum of normalized rewards still double-counts redundant signals. We treat the reward vector as a sample from a multivariate distribution and *precondition* it with its inverse square-root covariance, which accounts for both the scale of each reward and the covariance between rewards (Mahalanobis, 2018). Under a sum, the preconditioner changes only the weight of each reward; for aggregators that penalize imbalance between rewards, it also changes the scale of joint improvement relative to imbalance. Figure 1 shows the effect on correlated (redundant) and anti-correlated (antagonistic) rewards.

Parallel to these geometric concerns is the econometric challenge of substitution. Even in a standardized space, a linear sum can increase while one reward falls, provided an easier reward (e.g., length) rises by more. We therefore also consider aggregators from welfare economics (Nash et al., 1950), Max Nash Welfare and egalitarian welfare (SoftMin), which penalize low values of any single reward.

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 decorrelate the rewards, generalizing diagonal normalization. Second, Econometric Aggregation uses welfare functions such as *Nash* to penalize the collapse of any single reward. GRPO and GDPO are cells of this design space (Table 1). Our contributions are:

A Unified Framework (GEOMA): We decompose multi-reward aggregation into *geometric preconditioning* and *econometric aggregation*. GRPO and GDPO are the Identity-Sum and Diagonal-Sum cells of the resulting design space.

Analysis of Multi-Reward Failure Modes: We analyze *geometric redundancy* (double-counting correlated signals) and *econometric substitution* (raising the sum at the expense of one reward), and relate each to the choice of preconditioner and aggregator (Section 6). Under a sum, every preconditioner reduces to a choice of reward weights (Remark 4.1), so beyond these weights, the geometry of the preconditioned rewards matters only for nonlinear aggregators.

Shorter responses than GDPO at higher accuracy: On four math benchmarks, covariance sphering with batch statistics and Sum aggregation raises mean accuracy from 46.8% (GDPO) to 48.4% while reducing mean response length from 3,897 to 2,518 tokens; GRPO keeps the highest mean accuracy (50.2%) with 6,848 tokens. On tool calling, most GEOMA configurations learn the format constraint that GRPO does not learn (Figure 2), and with Nash aggregation only covariance sphering learns it. All six covariance configurations score above GRPO and GDPO in BFCL average accuracy, with Covariance-Nash highest at 45.28%.

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 uses 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 the policy exploits heuristic constraints or develops verbosity biases (Saito et al., 2023; Hu et al., 2024).

Recent efforts to stabilize multi-reward alignment have focused on scale normalization. GDPO (Liu et al., 2026) applies a diagonal variance normalization to prevent scale dominance among rewards, and Zhao et al. (2025b) use the geometric mean of rewards over the batch. GEOMA replaces the diagonal normalization with covariance sphering, which also accounts for the correlation between rubrics. We also connect LLM alignment to welfare economics (Nash et al., 1950; Sen, 2017): the

Max Nash Welfare product and egalitarian max-min objectives penalize outcomes in which a single reward collapses, which is relevant to the length biases documented in recent alignment benchmarks (Dubois et al., 2024; Lambert et al., 2024).

Multi-objective optimization offers other ways to combine objectives. Nash-MTL (Navon et al., 2022) applies Nash bargaining to task gradients, and MGDA (Sener & Koltun, 2018) and PCGrad (Yu et al., 2020) combine or project per-task gradients. Smooth Tchebycheff scalarization (Lin et al., 2024b) smooths the worst objective with a log-sum-exp, which is close to our SoftMin aggregator, and MaxMin-RLHF (Chakraborty et al., 2024) optimizes the worst-off preference group. These methods act on gradients or on reward weights. GEOMA acts on the group-relative advantage vector before the policy gradient is formed. Liu et al. (2025b) show that GRPO's per-group standard-deviation normalization introduces a difficulty bias, which bears on the choice between group and batch statistics in Section 5.

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. $\tilde{A}^{(j)} = \mathrm{Norm}_g\big(\sum_k r_k^{(j)}\big)$.
- **GDPO (Normalize-then-Sum):** Normalizes dimensions independently, then sums. $\tilde{A}^{(j)} = \mathrm{Norm}_g\big(\sum_k u_k^{(j)}/\sigma_k\big)$.

Here $\mathrm{Norm}_g(x^{(j)}) = (x^{(j)} - \bar{x})/(\mathrm{std}_g(x) + \varepsilon)$ standardizes a scalar over the G responses of a group. We apply the same step to the output of every GEOMA aggregator (Section 5).

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:

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

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:

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

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

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 $\tilde{A}^{(j)} = \mathcal{G}(a^{(j)})$, which is normalized within the group (Norm_g, Section 3) before it weights policy updates. GRPO is the (None, Sum) cell and GDPO the (Diagonal, Sum) cell. In the Sum column, the preconditioner acts only through the reward weights $P^\top \mathbf{1}$ (Remark 4.1). 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})	$a^{(j)} = (\Sigma + \varepsilon I)^{-1/2} u^{(j)}$	$\sum_{k=1}^{K} a_k^{(j)}$	$\sum_{k=1}^{K} \log \psi\left(a_k^{(j)}\right)$	$-\log\left(\frac{1}{K}\sum_{k=1}^{K} e^{-a_k^{(j)}}\right)$

P_{Id} corresponds to GRPO. P_{Diag} recovers GDPO-style diagonal normalization. P_{Cov} is our proposed **Covariance Sphering** (ZCA whitening), which decorrelates the reward coordinates and gives each unit variance (Table 1). Among whitening transforms, ZCA stays closest to the original coordinates (Kessy et al., 2018), so Nash and SoftMin still act on coordinates that correspond to the original rubrics.

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

- **Utilitarian (Sum):** $\mathcal{G}_{\mathrm{sum}}(\mathbf{a}) = \sum_k a_k$. Treats coordinates as substitutes at a constant one-to-one rate.
- **Max Nash Welfare (MNW):** $\mathcal{G}_{\mathrm{MNW}}(\mathbf{a}) = \sum_k \log \sigma(a_k)$. Maximizes the product of logistic utilities. Since $\log \sigma(a) = -\mathrm{softplus}(-a)$, each term is close to linear for $a_k \ll 0$ and close to zero for $a_k \gg 0$, so MNW penalizes coordinates below the group mean and saturates above it.
- **SoftMin:** $\mathcal{G}_{\mathrm{softmin}}(\mathbf{a}) = -\log\left(\frac{1}{K}\sum_k e^{-a_k}\right)$. A smooth approximation of $\min_k a_k$ that gives the most weight to the lowest coordinate.

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 leaves the correlation between rewards unchanged.

Redundancy and Signal Double-Counting. Consider two rubrics, r_1 and r_2, measuring semantically similar attributes such as truthfulness and honesty ($\rho \approx 1$). Diagonal preconditioning rescales them independently. Under linear aggregation ($\tilde{A} = \sum_k a_k$), their contributions add, which doubles the weight of the shared latent factor relative to an independent third rubric r_3. The inverse covariance root $P = \Sigma^{-1/2}$ reduces this weight from 2 to $\sqrt{2}$ when the two rewards are identical (Section 6.2).

Ill-Conditioned Reward Vectors. Conversely, consider anti-correlated rubrics such as conciseness and correctness, where increasing one degrades the other. The covariance matrix Σ has negative off-diagonal entries; for unit-variance rewards, the principal axes are the agreement direction $[1, 1]^\top$ (low variance) and the trade-off direction $[1, -1]^\top$ (high variance). Diagonal normalization leaves this structure unchanged. **Covariance Sphering** ($P = \Sigma^{-1/2}$) whitens the distribution, scaling each eigenvector by its inverse square-root eigenvalue, which amplifies the low-variance agreement direction relative to the trade-off direction. Nash and SoftMin penalize the component of $\mathbf{a}$ along the trade-off direction, so covariance sphering raises the scale of joint improvement relative to the imbalance they penalize (Appendix B.3).

Remark 4.1 (Sum uses the preconditioner only through reward weights). For any preconditioner, $\mathcal{G}_{\mathrm{sum}}(P\mathbf{u}) = \mathbf{1}^\top P\mathbf{u} = \mathbf{w}^\top \mathbf{u}$ with $\mathbf{w} = P^\top \mathbf{1}$. After group normalization, each (P, Sum) cell of Table 1 is therefore GRPO applied to the scalar reward $\mathbf{w}^\top \mathbf{r}$, and GRPO, GDPO and Covariance-Sum differ only in $\mathbf{w}$. Because the sum depends on $\mathbf{a}$ only through its component along $\mathbf{1}$, the isotropy of

$\mathbf{a}$ (Theorem 6.1) and the rescaling of the agreement and trade-off directions described above affect the Sum advantage only through $\mathbf{w}$. For $K = 2$ rewards with equal variance, $\mathbf{w} \propto \mathbf{1}$ under all three preconditioners, and the three Sum cells coincide. MNW and SoftMin penalize imbalance between coordinates, which is measured by the components of $\mathbf{a}$ orthogonal to $\mathbf{1}$, so their advantage depends on the whole matrix P. Appendix B.3 gives the proofs, a two-reward example, and the case of duplicated rewards.

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.

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 $D = \mathrm{diag}(\Sigma)$ 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:

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

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).

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 $P \in \{I, P_{\mathrm{diag}}, P_{\mathrm{cov}}\}$
Require: Aggregator $\mathcal{W} \in \{\mathrm{Sum}, \mathrm{Nash}, \mathrm{SoftMin}\}$
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 $P = P_{\mathrm{diag}}$ **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
  $\Sigma \leftarrow \frac{1}{G} \sum_j \mathbf{u}^{(j)} \mathbf{u}^{(j)\top}$ Batch-Cov: pool over batch
  $\mathbf{a}^{(j)} \leftarrow (\Sigma + \epsilon I)^{-1/2} \mathbf{u}^{(j)}$
end if
// **Step 4: Econometric Aggregation**
if $\mathcal{W} = \mathrm{Sum}$ **then**
  $\tilde{A}^{(j)} \leftarrow \sum_{k=1}^{K} a_k^{(j)}$
else if $\mathcal{W} = \mathrm{Nash}$ **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
// **Step 5: Normalize within Group**
$\tilde{A}^{(j)} \leftarrow (\tilde{A}^{(j)} - \bar{A}) / (\mathrm{std}_g(\tilde{A}) + \epsilon)$
Return $\{\tilde{A}^{(j)}\}_{j=1}^{G}$

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 $\boldsymbol{\nu} = [\nu_1, \ldots, \nu_K] \in \mathbb{R}^K_{\geq 0}$ 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, we normalize the scalar advantages within each group, $\tilde{A}^{(j)} \leftarrow \text{Norm}_g(\tilde{A}^{(j)})$ (Section 3), in all experiments. This makes the update magnitude invariant to the scale of the aggregator, and it makes the (None, Sum) and (Diagonal, Sum) cells coincide exactly with GRPO and GDPO.

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 relate the aggregators to the power-mean family, and we characterize how covariance sphering changes the reward covariance and how it treats redundant rewards. Full proofs are in Appendix B.

6.1 Econometric Aggregation: The Generalized Mean Spectrum

The utilitarian (Sum), Nash (MNW), and egalitarian (SoftMin) aggregators are power means of utilities. With the normalized power mean $M_p(\mathbf{z}) = \left(\frac{1}{K}\sum_{k=1}^{K} z_k^p\right)^{1/p}$ and M_0 the geometric mean, the utilities $z_k = e^{a_k}$ give $\mathcal{G}_{\text{sum}}(\mathbf{a}) = K \log M_0(\mathbf{z})$ and $\mathcal{G}_{\text{softmin}}(\mathbf{a}) = \log M_{-1}(\mathbf{z})$, and the utilities $z_k = \sigma(a_k)$ give $\mathcal{G}_{\text{MNW}}(\mathbf{a}) = K \log M_0(\mathbf{z})$. Lowering p moves M_p from the arithmetic mean toward the minimum utility, so the aggregators differ in how much weight they give to the lowest coordinate. Appendix B.1.1 states the limits (Theorem B.2) and plots the spectrum (Figure 6).

6.2 Geometric Preconditioning

We now analyze the action of the preconditioner $\mathbf{a} = P\mathbf{u}$ on centered rewards with covariance Σ.

Theorem 6.1 (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:*

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

Remark. In the population, the sphered rewards are isotropic, with unit variance along every direction, whereas diagonal sphering (GDPO) standardizes only the coordinate axes. With the ridge term ε and a finite-sample estimate of Σ, the sphered covariance is close to I but not equal to it (Theorem B.12).

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

For two identical rewards with standard deviation σ and an uncorrelated third reward with standard deviation σ_3, the null direction is the difference of the identical pair, and the sphered sum is $\mathbf{1}^\top \mathbf{a} = \sqrt{2}\, u_1/\sigma + u_3/\sigma_3$, compared with $2u_1/\sigma + u_3/\sigma_3$ under diagonal sphering. Covariance sphering therefore reduces double counting under Sum without removing it; weighting by the precision, $\mathbf{w} \propto \Sigma^\dagger \mathbf{1}$, gives the same advantage whether or not a reward is duplicated (Appendix B.3).

7 Experiments

7.1 Tool Calling

Setup. We follow the tool-use setting of ToolRL (Qian et al., 2025), whose training data combine ToolACE (Liu et al., 2025a), Hammer (Lin et al., 2024a) and xLAM (Zhang et al., 2024). The reward

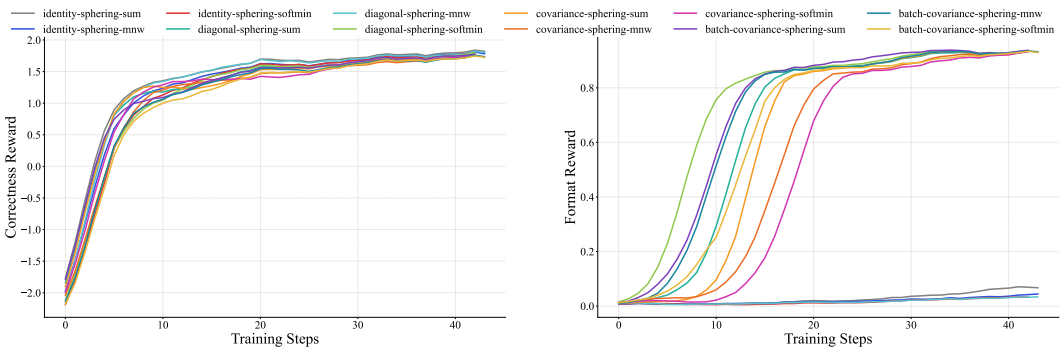


Figure 2: **Every covariance configuration learns the format constraint that GRPO misses.** Correctness (left) and format (right) training rewards on the tool-calling task (Qwen2.5-1.5B-Instruct), steps 0 to 42, by which point format rewards have stabilized. Correctness rises at a similar rate for all configurations. The format reward exceeds 0.85 by about step 25 for all six covariance configurations and for Diagonal-Sum (GDPO) and Diagonal-SoftMin; it stays below 0.1 for Identity-Sum (GRPO, gray), Identity-MNW, Identity-SoftMin and Diagonal-MNW.

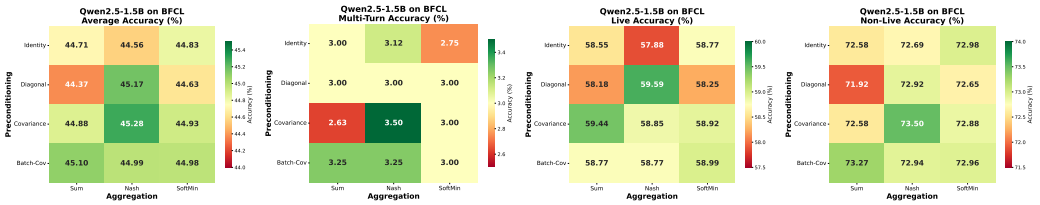


Figure 3: **Every covariance cell scores above both GRPO and GDPO in average accuracy.** BFCL-v3 accuracy (%) of Qwen2.5-1.5B-Instruct after tool-calling training, with preconditioners as rows and aggregators as columns. Panels, left to right: average, multi-turn, live and non-live accuracy. GRPO is the (Identity, Sum) cell and GDPO the (Diagonal, Sum) cell. Covariance-Nash has the highest average accuracy (45.28%).

has a binary format reward $R_{\text{format}} \in \{0, 1\}$ and a correctness reward $R_{\text{correct}} \in [-3, 3]$. We fine-tune Qwen2.5-1.5B-Instruct (Qwen et al., 2025) with verl (Sheng et al., 2025) for 100 steps ($G = 4$, batch size 512), with one training run per configuration, and evaluate on BFCL-v3 (Patil et al., 2025); Appendix C gives details.

Covariance sphering learns the format constraint under every aggregator. Figure 2 shows the first 42 steps of training on the tool-calling task, which pairs a correctness reward in $[-3, 3]$ with a binary format reward. Under GRPO (Identity-Sum), the format reward stays below 0.1 while correctness rises, which is consistent with scale dominance, since the correctness reward has six times the range of the format reward. All six covariance configurations (group and batch statistics, each with Sum, MNW and SoftMin) exceed 0.85 by about step 25 while their training correctness stays within 0.1 of GRPO's on a reward range of 6. Batch-Covariance-Sum and Batch-Covariance-MNW pass 0.85 by step 16. With Identity or Diagonal preconditioning, only Diagonal-Sum (GDPO) and Diagonal-SoftMin learn the constraint; Identity-MNW, Identity-SoftMin and Diagonal-MNW keep the format reward below 0.1. With MNW aggregation, the format constraint is therefore learned only after covariance sphering.

Tool Calling on BFCL. We evaluate the trained policies on the Berkeley Function Calling Leaderboard (Patil et al., 2025). Figure 3 reports average, multi-turn, live and non-live accuracy for every configuration. All six covariance-based configurations score above both GRPO (44.71%) and GDPO (44.37%) in average accuracy, and Covariance-Nash scores highest (45.28%). Averaged over the three aggregators, the covariance rows score 45.03% and 45.02%, against 44.70% for Identity and 44.72% for Diagonal preconditioning. The spread across configurations is under one point. One explanation is that the two training rewards measure different properties, format and correctness,

Table 2: **Covariance sphering gives higher mean accuracy than GDPO with shorter responses.** Pass@1 accuracy (%) and mean response length (tokens) on four math benchmarks after 200 training steps with Sum aggregation and a length penalty above 4,000 tokens, averaged over 5 generations per question. Bold marks the best trained model in each column. Batch-Covariance matches or exceeds GDPO's accuracy on all four benchmarks and reduces mean length from 3,897 to 2,518 tokens; on AIME-25 it matches GDPO's 28.0% with 3,475 tokens, against 6,767 for GDPO and 11,990 for GRPO.

	MATH-500		AIME-25		AMC-24		Minerva		Mean	
Preconditioner (with Sum)	**Acc (↑)**	**Tok (↓)**	**Acc (↑)**	**Tok (↓)**	**Acc (↑)**	**Tok (↓)**	**Acc (↑)**	**Tok (↓)**	**Acc (↑)**	**Tok (↓)**
Base Model	74.6	7340	22.6	21360	38.7	16110	22.3	9961	39.6	13693
Identity (GRPO)	**85.2**	3410	27.3	11990	**61.8**	6903	**26.5**	5087	**50.2**	6848
Diagonal (GDPO)	83.2	2063	**28.0**	6767	52.4	5111	23.5	**1646**	46.8	3897
Covariance	84.0	1713	24.6	**3137**	59.1	3221	24.7	1881	48.1	**2488**
Batch-Covariance	83.9	**1706**	**28.0**	3475	57.7	**3113**	23.9	1778	48.4	2518

which we expect to be nearly uncorrelated, so preconditioning has less correlation to correct than on math.

7.2 Mathematical Reasoning

Setup. We initialize from DeepSeek-R1-Distill-Qwen-1.5B (DeepSeek-AI, 2025) and train on DeepScaleR-Preview (Tan et al., 2025) for 200 steps with a binary correctness reward on the final answer and a piecewise-linear length reward that penalizes responses longer than 4,000 tokens, up to the maximum context length. The global batch size is 512 with $G = 8$ rollouts per prompt, and each configuration is trained once. We evaluate on MATH-500 (Hendrycks et al., 2021), AIME-25 (MAA, 2025), AMC-24 (MAA, 2024) and Minerva (Lewkowycz et al., 2022), and report pass@1 and mean response length, each averaged over 5 generations per question.

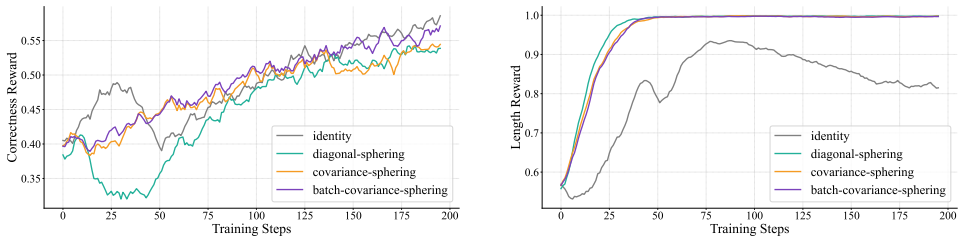


Figure 4: **Covariance sphering reaches the length limit without the correctness dip seen under GDPO.** Training correctness (left) and length reward (right) over 200 steps with Sum aggregation. GRPO (Identity) keeps the highest final correctness, but its length reward peaks near 0.93 and falls to about 0.82. GDPO (Diagonal) reaches the length limit while its correctness drops to about 0.32 between steps 10 and 45. Covariance and Batch-Covariance reach the length limit by about step 50 with no such drop.

Covariance sphering holds the length limit while correctness keeps rising. Figure 4 compares the preconditioners over 200 training steps under Sum aggregation, with a correctness reward and a length penalty that conflict. GDPO reaches the length limit, but its training correctness falls from about 0.41 to 0.32 between steps 10 and 45 before recovering. Covariance and Batch-Covariance reach the length limit by about step 50 without this dip and reach a training correctness of 0.45 about 30 steps before GDPO; from about step 55 to step 115, Batch-Covariance has the highest training correctness of all four preconditioners. GRPO ends with the highest training correctness, but its length reward peaks near 0.93 and declines to about 0.82. Under Sum, the covariance variants differ from GDPO in the reward weights $\mathbf{w} = P^\top \mathbf{1}$ (Remark 4.1), which covariance sphering sets from the full reward covariance, including the correlation between correctness and length.

Table 2 holds the aggregator fixed at Sum and varies the preconditioner, which under Sum sets the reward weights $\mathbf{w} = P^\top \mathbf{1}$ (Remark 4.1). Averaged over the four benchmarks, Batch-Covariance reaches 48.4% accuracy with 2,518 tokens, and Covariance 48.1% with 2,488, against 46.8% and 3,897 tokens for GDPO. Batch-Covariance matches or exceeds GDPO's accuracy on every benchmark

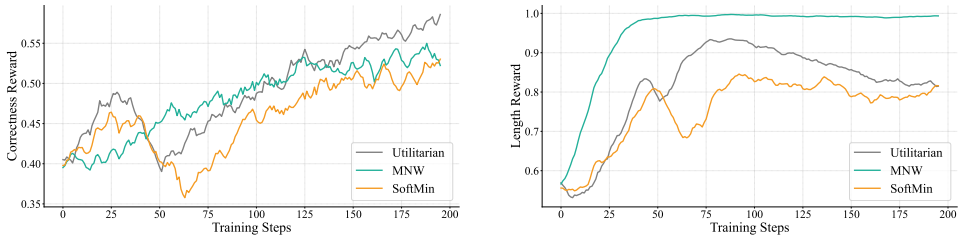


Figure 5: **Under Identity preconditioning, MNW enforces the length limit and Sum does not.** Training correctness (left) and length reward (right) over 200 steps for the three aggregators. Sum ends with the highest correctness (about 0.585), while its length reward peaks near 0.93 and declines to about 0.82. MNW reaches a length reward near 1.0 by about step 40 and holds it, ending at about 0.53 correctness. SoftMin ends with lower correctness than Sum and about the same length reward.

and uses fewer tokens on all but Minerva; on AIME-25 it matches GDPO's 28.0% with 3,475 tokens, against 6,767 for GDPO and 11,990 for GRPO. GRPO has the highest mean accuracy (50.2%) with 6,848 tokens, and its length reward falls to about 0.82 during training (Figure 4). By Proposition B.16(i), Batch-Covariance with Sum is GRPO on the reward $\mathbf{w}^\top \mathbf{r}$ with weights set from the batch covariance, so the batch statistics set the relative weight of correctness and length.

With Nash or SoftMin aggregation, the covariance preconditioners give the shortest responses. Averaged over the four benchmarks in Table 3, Batch-Covariance-SoftMin uses 2,330 tokens at 47.3% accuracy and Covariance-SoftMin 2,442 at 47.2%, against 2,606 tokens at 48.4% for Diagonal-SoftMin; Covariance-Nash uses 2,420 tokens at 45.5%, against 2,553 at 47.5% for Diagonal-Nash. Six of the nine configurations other than GRPO and GDPO have both higher mean accuracy and fewer mean tokens than GDPO, four of them with covariance sphering. Three of the six configurations on the Pareto frontier of mean accuracy and mean tokens use covariance sphering (Covariance-Sum, Batch-Covariance-Sum and Batch-Covariance-SoftMin), and GDPO is not on it.

With Identity preconditioning, MNW enforces the length limit at a small cost in correctness. Figure 5 fixes the preconditioner to Identity and varies the aggregator. Under Sum, the length reward declines after about step 90 while correctness keeps rising. MNW reaches the length limit by about step 40 and holds it, and its final training correctness is about 0.53, against about 0.585 under Sum. The hinge form of MNW (Section 4) is consistent with this, since a response below the group mean on either reward is penalized roughly linearly, while one above the mean gains little. On tool calling, Identity-MNW and Diagonal-MNW do not learn the format reward, while Covariance-MNW and Batch-Covariance-MNW do (Figure 2), so there MNW learns the format constraint only in combination with covariance sphering.

8 Discussion and Conclusion

We proposed **GEOMA**, which splits multi-reward advantage computation into geometric preconditioning and econometric aggregation and has GRPO and GDPO as cells. Covariance sphering makes the reward distribution isotropic in the population. Under Sum, every preconditioner reduces to a choice of reward weights, and covariance sphering lowers the weight of a duplicated reward from 2 to $\sqrt{2}$ relative to an independent reward; beyond these weights, the geometry of the sphered rewards matters for Nash and SoftMin, which penalize imbalance between rewards. On mathematical reasoning, Batch-Covariance with Sum raised mean accuracy over GDPO from 46.8% to 48.4% and reduced mean response length from 3,897 to 2,518 tokens, and with Identity preconditioning, MNW enforced a length limit that the plain sum did not. On tool calling, every covariance configuration learned the format constraint that GRPO did not learn, with MNW only the covariance preconditioners learned it, and all six covariance configurations scored above GRPO and GDPO in BFCL average accuracy, with Covariance-Nash highest at 45.28%.

Ethics Statement

This work proposes GEOMA, a mathematical framework for aggregating multiple reward signals during the reinforcement learning phase of LLM alignment. We discuss the ethical dimensions of this research below.

Alignment Quality and Safety. GEOMA is designed for multi-objective alignment. Nash and SoftMin aggregation penalize a low value of any single reward, which can help keep a training run from trading a safety-relevant reward for an easier one.

No New Risks from the Framework Itself. GEOMA operates exclusively at the reward aggregation layer of the RLHF pipeline. It does not introduce new data collection practices, modify the underlying language model architecture, or alter the types of content that models can generate. The framework is agnostic to the specific reward signals being aggregated and thus does not introduce new dual-use risks beyond those already present in the base RLHF paradigm.

Potential for Misuse. As with any improvement to alignment methodology, more effective reward optimization could in principle be used to align models toward harmful objectives if the reward signals themselves encode undesirable behavior. However, this risk applies to RLHF in general and is not specific to our contribution. GEOMA's welfare objectives (Nash, SoftMin) penalize the neglect of any single reward.

Computational and Environmental Impact. GEOMA introduces negligible computational overhead relative to standard GRPO training. The covariance estimation and matrix operations are performed on $K \times K$ matrices (typically $K = 2$), which is computationally trivial compared to the forward and backward passes of the language model. Our experiments use 1.5B-parameter models trained for 100 and 200 steps.

Compliance with Ethical Standards. This research does not involve human subjects, crowdsourcing, or the collection of personally identifiable information. All datasets used (BFCL, MATH-500, AIME, AMC, Minerva) are publicly available benchmarks with established community usage norms. We confirm that this work conforms with the ICLR Code of Ethics.

Reproducibility Statement

Section 5 and Algorithm 1 specify the full GEOMA procedure, including covariance estimation, sphering, aggregation and group normalization, and Section 4 (Table 1) defines every preconditioner and aggregator we evaluate. Section 7 and Appendix C give the models, training data, reward functions, batch sizes, rollout counts and step counts for both tasks, together with the full math results (Table 3). Each configuration is trained once with `verl` and evaluated with vLLM (temperature 0.6, top-p 0.95, top-k 20); math scores are pass@1 averaged over 5 generations per question, and the $\pm$ values in Table 3 are standard deviations over these generations. Appendix B states the assumptions and gives complete proofs of the theoretical results, and Appendix E reports the compute resources. All benchmarks are publicly available, and all training data sources are cited in Section 7.

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.

Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Furong Huang, Dinesh Manocha, Amrit Singh Bedi, and Mengdi Wang. MaxMin-RLHF: Alignment with diverse human preferences. *arXiv preprint arXiv:2402.08925*, 2024.

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.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

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.

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, 2021. URL https://arxiv.org/abs/2103.03874.

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.

Agnan Kessy, Alex Lewin, and Korbinian Strimmer. Optimal whitening and decorrelation. *The American Statistician*, 72(4):309–314, 2018.

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.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models, 2022. URL https://arxiv.org/abs/2206.14858.

Qiqiang Lin, Muning Wen, Qiuying Peng, Guanyu Nie, Junwei Liao, Jun Wang, Xiaoyun Mo, Jiamu Zhou, Cheng Cheng, Yin Zhao, Jun Wang, and Weinan Zhang. Hammer: Robust function-calling for on-device language models via function masking, 2024a. URL https://arxiv.org/abs/2410.04587.

Xi Lin, Xiaoyuan Zhang, Zhiyuan Yang, Fei Liu, Zhenkun Wang, and Qingfu Zhang. Smooth Tchebycheff scalarization for multi-objective optimization. *arXiv preprint arXiv:2402.19078*, 2024b.

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.

Weiwen Liu, Xu Huang, Xingshan Zeng, Xinlong Hao, Shuai Yu, Dexun Li, Shuai Wang, Weinan Gan, Zhengying Liu, Yuanqing Yu, Zezhong Wang, Yuxian Wang, Wu Ning, Yutai Hou, Bin Wang, Chuhan Wu, Xinzhi Wang, Yong Liu, Yasheng Wang, Duyu Tang, Dandan Tu, Lifeng Shang, Xin Jiang, Ruiming Tang, Defu Lian, Qun Liu, and Enhong Chen. Toolace: Winning the points of llm function calling, 2025a. URL https://arxiv.org/abs/2409.00920.

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*, 2025b.

MAA. American invitational mathematics examination - amc. In *American Invitational Mathematics Examination - AMC*, 2024.

MAA. American invitational mathematics examination - aime. In *American Invitational Mathematics Examination - AIME 2025*, 2025.

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.

Aviv Navon, Aviv Shamsian, Idan Achituve, Haggai Maron, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. Multi-task learning as a bargaining game. *arXiv preprint arXiv:2202.01017*, 2022.

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.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018.

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.

Sijun Tan, Michael Luo, Justin Wong, Colin Cai, Xiaoxiang Shi, William Yuan Tang, Manan Roongta, Tianjun Zhang, Li Erran Li, Raluca Ada Popa, et al. DeepScaleR: Effective RL scaling of reasoning models via iterative context lengthening, 2025.

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.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33: 5824–5836, 2020.

Jianguo Zhang, Tian Lan, Ming Zhu, Zuxin Liu, Thai Hoang, Shirley Kokane, Weiran Yao, Juntao Tan, Akshara Prabhakar, Haolin Chen, Zhiwei Liu, Yihao Feng, Tulika Awalgaonkar, Rithesh Murthy, Eric Hu, Zeyuan Chen, Ran Xu, Juan Carlos Niebles, Shelby Heinecke, Huan Wang, Silvio Savarese, and Caiming Xiong. xlam: A family of large action models to empower ai agent systems, 2024. URL https://arxiv.org/abs/2409.03215.

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

The appendix is organized as follows.

- Appendix A extends the related work.
- Appendix B gives the full theoretical results and proofs, including the power-mean characterization (Appendix B.1.1) and the analysis of Sum aggregation (Appendix B.3).
- Appendix C gives the tool-calling setup and additional math results (Table 3).
- Appendix D discusses limitations and future directions.
- Appendix E reports compute resources.
- Appendix F contains the AI use statement.

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 tail-focused approach 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 also enter the learning signal and can interact 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. GEOMA connects group-based RLHF (Shao et al., 2024) with multi-objective RL and welfare economics. Prior multi-objective alignment work uses regularization (Gupta et al., 2024) or a plain sum; GEOMA treats aggregation as a statistical and econometric problem. GDPO (Liu et al., 2026) is the diagonal case of its covariance-aware preconditioner. Risk-sensitive RL focuses on tail outcomes; GEOMA instead uses Nash and SoftMin welfare objectives, which penalize low values of any single reward, within the same design space as GRPO and GDPO.

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

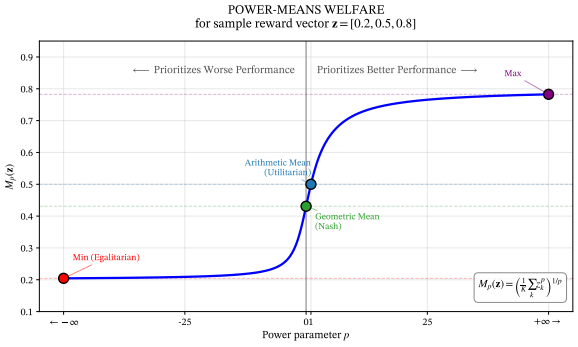


Figure 6: **Decreasing p moves the power mean from the average toward the minimum.** Power mean $M_p(\mathbf{z})$ of the fixed utility vector $\mathbf{z} = [0.2, 0.5, 0.8]$ as a function of p. $p = 1$ gives the arithmetic mean (Sum), $p \to 0$ the geometric mean (Nash), and $p \to -\infty$ the minimum (egalitarian).

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 $p < 0$,

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

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

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

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 is an 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*

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

Smaller (worse) coordinates therefore 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 $T > 0$,*

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

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

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 $\sum_k e^{-a_k/T} \geq e^{-m/T}$ and thus

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

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

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

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* $\nu_1, \ldots, \nu_K \geq 0$. *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

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

Hence

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

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 SUM AGGREGATION USES THE PRECONDITIONER ONLY THROUGH REWARD WEIGHTS

This subsection proves the statements in Remark 4.1. The aggregators are those of Table 1 with $\psi = \sigma$, and $\mathrm{Norm}_g(x^{(j)}) = (x^{(j)} - \bar{x})/(\mathrm{std}_g(x) + \varepsilon)$ standardizes a scalar over the G responses of a group.

Proposition B.16 (Sum aggregation reweights the rewards). *Let P be constant within a group and let $\mathbf{w} = P^\top \mathbf{1}$. Then $\mathrm{Norm}_g(\mathbf{1}^\top P \mathbf{u}^{(j)}) = \mathrm{Norm}_g(\mathbf{w}^\top \mathbf{r}^{(j)})$ for every response j, so the $(P, \textit{Sum})$ cell is GRPO applied to the scalar reward $\mathbf{w}^\top \mathbf{r}$. In particular, (i) Batch-Covariance-Sum is GRPO with the batch-adaptive weights $\mathbf{w} = (\Sigma_{\textit{batch}} + \varepsilon I)^{-1/2}\mathbf{1}$, and (ii) with group statistics, Covariance-Sum and Diagonal-Sum (GDPO) give the same advantages when the rewards are uncorrelated within the group, and, for $K = 2$, also when the two rewards have equal within-group variance, in which case both equal GRPO. For $K = 2$, $\varepsilon = 0$ and $|\rho| < 1$, the Covariance-Sum and GDPO weight vectors are parallel only in these cases.*

Proof. Since P is constant within the group, $\mathbf{1}^\top P \mathbf{u}^{(j)} = \mathbf{w}^\top \mathbf{u}^{(j)} = \mathbf{w}^\top \mathbf{r}^{(j)} - \mathbf{w}^\top \boldsymbol{\mu}$, and $\mathbf{w}^\top \boldsymbol{\mu}$ is the same for every response in the group. Norm_g is invariant to adding a constant, which proves the identity; up to its ε, it is also invariant to multiplication by a positive constant, so only the direction of $\mathbf{w}$ matters. Part (i) is the identity for $P = (\Sigma_{\text{batch}} + \varepsilon I)^{-1/2}$, which is symmetric, so $P^\top \mathbf{1} = P\mathbf{1}$. For part (ii), uncorrelated rewards give $\Sigma = D$, hence $P_{\text{Cov}} = P_{\text{Diag}}$. If $K = 2$ and both variances equal σ^2, then $\Sigma\mathbf{1} = \sigma^2(1 + \rho)\mathbf{1}$, so $\mathbf{1}$ is an eigenvector of both $\Sigma + \varepsilon I$ and $D + \varepsilon I$, and both weight vectors are positive multiples of $\mathbf{1}$, the GRPO weights. Conversely, let $K = 2$ and $\varepsilon = 0$, and write $t = \sqrt{\det \Sigma} = \sigma_1\sigma_2\sqrt{1 - \rho^2}$ and $\mathbf{v} = D^{-1/2}\mathbf{1}$. Every 2×2 positive definite matrix satisfies $\Sigma^{1/2} = (\Sigma + tI)/\sqrt{\mathrm{tr}\,\Sigma + 2t}$, so $\Sigma^{-1/2}\mathbf{1}$ is parallel to $\mathbf{v}$ if and only if $(\Sigma + tI)\mathbf{v}$ is parallel to $\mathbf{1}$. A direct computation gives $(\Sigma + tI)\mathbf{v} = \big((1 + \rho)\sigma_1 + t/\sigma_1,\ (1 + \rho)\sigma_2 + t/\sigma_2\big)^\top$, whose entries differ by $(\sigma_1 - \sigma_2)\big(1 + \rho - \sqrt{1 - \rho^2}\big)$. This vanishes if and only if $\sigma_1 = \sigma_2$ or $\rho = 0$. □

The sum ignores every direction orthogonal to 1. For any $\mathbf{v}$ with $\mathbf{1}^\top \mathbf{v} = 0$, $\mathcal{G}_{\text{sum}}(\mathbf{a} + \mathbf{v}) = \mathcal{G}_{\text{sum}}(\mathbf{a})$, while MNW and SoftMin are not functions of $\mathbf{1}^\top \mathbf{a}$ alone. The values below are for preconditioned vectors with $a_1 + a_2 = 4$ that differ along the trade-off direction $(1, -1)$. The sum is 4 for each, and MNW and SoftMin rank the balanced vector highest.

$\mathbf{a}$	$\mathcal{G}_{\text{sum}}(\mathbf{a})$	$\mathcal{G}_{\text{MNW}}(\mathbf{a})$	$\mathcal{G}_{\text{softmin}}(\mathbf{a})$
$(2,\ 2)$	4	-0.25	2.00
$(3,\ 1)$	4	-0.36	1.57
$(5,\ -1)$	4	-1.32	-0.31

For $K = 2$, MNW and SoftMin penalize the trade-off component. For $K = 2$, let $s = (a_1 + a_2)/\sqrt{2}$ and $d = (a_1 - a_2)/\sqrt{2}$ be the components of $\mathbf{a}$ along the agreement and trade-off directions. Then

$$\mathcal{G}_{\text{sum}}(\mathbf{a}) = \sqrt{2}\, s, \qquad \mathcal{G}_{\text{softmin}}(\mathbf{a}) = \frac{s}{\sqrt{2}} - \log\cosh\frac{d}{\sqrt{2}}. \tag{24}$$

The MNW aggregator with $\psi = \sigma$ is concave (Theorem B.6) and symmetric in (a_1, a_2), so for fixed s it is largest at $d = 0$. SoftMin and MNW therefore penalize the trade-off component d, and the

sum does not depend on it. With $\varepsilon = 0$ and two rewards of equal variance and correlation ρ, diagonal sphering gives $\mathrm{Var}(s)/\mathrm{Var}(d) = (1+\rho)/(1-\rho)$ (Proposition B.15), and covariance sphering gives $\mathrm{Var}(s) = \mathrm{Var}(d)$ (Proposition B.11). Relative to diagonal sphering, covariance sphering thus multiplies the scale of s relative to d by $\sqrt{(1-\rho)/(1+\rho)}$, which is 3 at $\rho = -0.8$ and 0.23 at $\rho = 0.9$. For MNW and SoftMin, this enlarges joint improvement relative to the imbalance they penalize when the rewards are anti-correlated, and shrinks it when they are positively correlated. For the sum, it changes only the scale of s, which group normalization removes.

With unequal variances, covariance sphering changes the Sum weights. When the variances differ, $\mathbf{w}$ depends on the preconditioner, and Covariance-Sum is a different weighted sum from GDPO. The table below gives examples with $\varepsilon = 0$ and weights scaled so that $w_2 = 1$. Over standard-deviation ratios from 1 to 20, the angle between $\Sigma^{-1/2}\mathbf{1}$ and $D^{-1/2}\mathbf{1}$ is at most 3.2° for every negative correlation; for positive correlation, its maximum is 25° over $0 < \rho \leq 0.9$ and 49° over $0 < \rho \leq 0.99$.

(σ_1, σ_2)	ρ	Covariance-Sum w_1	GDPO w_1	Angle
$(1, 1)$	-0.8	1.00	1.00	0°
$(1, 1)$	0.9	1.00	1.00	0°
$(1, 2)$	-0.8	1.79	2.00	2.6°
$(1, 2)$	0.9	42.8	2.00	25.2°

Under Sum, covariance sphering reduces double counting without removing it. Let $r_1 = r_2$ with within-group standard deviation σ_1, and let r_3 be uncorrelated with them, with standard deviation σ_3. With the pseudo-inverse square root of Definition B.13, $\mathbf{1}^\top \Sigma^{\dagger -1/2}\mathbf{u} = \sqrt{2}\, u_1/\sigma_1 + u_3/\sigma_3$, and diagonal sphering gives $2u_1/\sigma_1 + u_3/\sigma_3$. Relative to the independent reward, covariance sphering therefore lowers the weight of the duplicated reward from 2 to $\sqrt{2}$; counting it once would give 1. Precision weights $\mathbf{w} \propto \Sigma^\dagger \mathbf{1}$ give $\mathbf{w}^\top \mathbf{u} \propto u_1/\sigma_1^2 + u_3/\sigma_3^2$, the same advantage as without the duplicate.

C Additional Experimental Results

Tool-Calling Training Setup. 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.

D Limitations and Future Directions

GEOMA takes the rubrics as given. GEOMA addresses how to aggregate multiple reward signals. Rubric design, which covers the choice of reward dimensions, the training of reward models and the calibration of their scales, is a complementary problem. As with any aggregation method, a rubric that rewards the wrong behavior passes that behavior on to the policy, and correcting it requires changing the rubric.

The cost of covariance sphering grows with the number of rubrics. Covariance sphering needs one symmetric eigendecomposition of a $K \times K$ matrix per group or batch. For the small K of current multi-rubric pipelines, this cost is negligible next to the forward and backward passes of the policy. Pipelines with hundreds of rubrics would call for low-rank or structured covariance estimates, which the framework accommodates without other changes.

Table 3: Accuracy (%) and mean response length (tokens) on four math benchmarks for every trained configuration after 200 training steps, grouped by preconditioner. Accuracy is the mean ± standard deviation over 5 generations per question; each row is one training run.

		MATH-500		AIME-25		AMC-24		Minerva	
Preconditioning	**Aggregator**	**Acc**	**Avg Tok**	**Acc**	**Avg Tok**	**Acc**	**Avg Tok**	**Acc**	**Avg Tok**
Baseline	-	74.6 ± 1.5	7340	22.6 ± 3.2	21360	38.7 ± 5.7	16110	22.3 ± 0.3	9961
Identity	Sum	85.2 ± 0.9	3410	27.3 ± 1.3	11990	61.8 ± 3.2	6903	26.5 ± 0.6	5087
	Nash	82.3 ± 1.0	1523	18.7 ± 1.6	2705	52.9 ± 5.5	2502	24.8 ± 1.2	1632
	Softmin	83.7 ± 0.3	3426	24.6 ± 1.6	11580	55.1 ± 4.3	7212	26.2 ± 1.8	4281
Diag. Sphering	Sum	83.2 ± 1.1	2063	28.0 ± 6.0	6767	52.4 ± 6.0	5111	23.5 ± 0.9	1646
	Nash	82.2 ± 0.0	1711	23.3 ± 2.0	3792	60.0 ± 3.0	3260	24.4 ± 0.9	1446
	Softmin	83.5 ± 0.8	1667	26.7 ± 3.0	4023	60.0 ± 5.0	3170	23.5 ± 0.5	1561
Cov. Sphering	Sum	84.0 ± 0.9	1713	24.6 ± 5.0	3137	59.1 ± 6.0	3221	24.7 ± 1.7	1881
	Nash	83.8 ± 0.6	1693	21.9 ± 5.0	3293	52.8 ± 4.0	2921	23.6 ± 0.9	1770
	Softmin	83.5 ± 0.9	1712	24.0 ± 3.0	3590	57.3 ± 6.0	2956	23.8 ± 1.2	1506
Batch Cov. Sph.	Sum	83.9 ± 1.0	1706	28.0 ± 4.0	3475	57.7 ± 4.5	3113	23.9 ± 0.9	1778
	Softmin	84.3 ± 0.7	1656	23.4 ± 4.2	2956	55.1 ± 6.0	2751	26.3 ± 1.3	1955

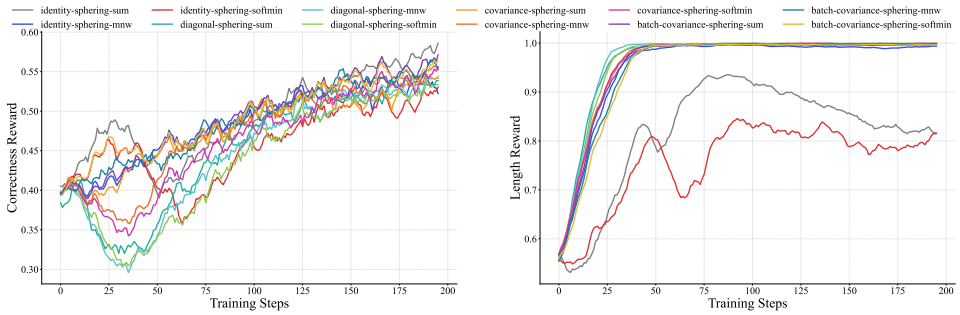


Figure 7: **All configurations except Identity-Sum and Identity-SoftMin reach the length limit.** Training correctness (left) and length reward (right) for all twelve configurations on DeepScaleR-Preview with DeepSeek-R1-Distill-Qwen-1.5B over 200 training steps. The other ten configurations reach a length reward near 1.0 by about step 50.

The aggregator encodes a value judgment. Nash and SoftMin encode the judgment that no rubric should collapse. When a rubric may be traded away freely, the Sum column of GEOMA is the appropriate choice, and the framework leaves this choice to the practitioner. Each run uses one aggregator throughout training.

Curriculum over the welfare spectrum. Sum, Nash and SoftMin correspond to different values of p in the power-mean family (Section 6.1). This suggests aggregators that change during training, for example starting from Sum and annealing toward Nash or SoftMin. The appropriate schedule is an open question.

Reward geometry as a training diagnostic. The eigenspectrum of Σ summarizes the dependence between rubrics. A large condition number indicates strongly correlated or anti-correlated rewards, and near-zero eigenvalues indicate redundant rubrics. We hypothesize that tracking the spectrum during training could flag reward hacking early, for example through a sudden drop in the smallest eigenvalue; testing this is future work.

Connection to natural gradient methods. Covariance preconditioning in reward space is analogous to Fisher preconditioning in parameter space, since both rescale by an inverse second-moment matrix. A framework that preconditions rewards by $\Sigma^{-1/2}$ and policy gradients by F^{-1} could combine the two.

E COMPUTE RESOURCES

All experiments were conducted on a cluster equipped with NVIDIA A100 (80GB) GPUs. The tool calling experiments (Qwen2.5-1.5B-Instruct, 100 steps, batch size 512, 4 rollouts) required approximately 4 GPU-hours per configuration. The mathematical reasoning experiments (DeepSeek-R1-Distill-Qwen-1.5B, 200 steps, batch size 512, 8 rollouts, 8192 max context) required approximately 16 GPU-hours per configuration. Including all ablation variants across both the geometric and econometric axes, the total compute for the experiments reported in this paper is approximately 200 GPU-hours. Preliminary experiments and hyperparameter exploration required an additional estimated 300 GPU-hours beyond what is reported in the final results.

F AI USE STATEMENT

We used AI assistants to edit the manuscript text, to check that the numbers in the text match the tables and figures, and to convert the paper to the ICLR template. AI tools were not used to write code, run experiments, or search the literature. The authors take full responsibility for the content of the paper.