# RETHINKING SURROGATE CRITERIA FOR RLVR ROLLOUT FILTERING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Reinforcement learning with verifiable rewards improves language-model reasoning, but it remains unclear whether common criteria for filtering generated responses preserve the learning signal needed for policy updates. We investigate when directional agreement preserves the gradient computed from all responses and when magnitude mismatch changes which subset should be preferred. We develop a controlled audit that separates gradient error into direction and scale, compares all subsets at the same retention budget, and evaluates selected subsets in the original trainable parameter space. Our theoretical analysis shows that perfect directional alignment can still yield large reconstruction error, motivating a selection criterion that jointly matches direction and magnitude. Across eight model–dataset pairs covering GSM8K, MATH, and TACO, retaining eight of sixteen responses with our criterion reduces mean squared relative gradient error by 66.8–75.7% compared with reward-variance selection and 42.9–53.7% compared with direction-only selection. These results identify scale mismatch as a limitation of common selection criteria and provide a framework for evaluating gradient preservation at a fixed policy.

## 1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) offers a scalable training signal for language-model reasoning on tasks such as mathematics and code generation. Group-based methods such as GRPO learn from multiple responses to each prompt without a separate learned critic (Shao et al., 2024), but processing these responses for policy optimization can be costly. Rollout filtering addresses this cost by updating on a smaller subset of generated responses, making the quality of the retained learning signal central to efficient post-training.

Existing approaches make this selection practical through surrogate indicators of training utility. CPPO prunes responses with small absolute advantages (Lin et al., 2025), while PODS retains groups with high reward variance (Xu et al., 2026b); $D^3S$ combines sample and token selection with a dynamic schedule (Wang et al., 2026). These methods demonstrate the value of allocating optimization effort selectively. Reward and advantage statistics summarize response outcomes; gradient fidelity also depends on how each response contributes in parameter space. Isolating this fidelity requires comparing subset and full-group gradients at the same policy state.

Gradient-aware criteria incorporate the learning signal into selection with different objectives. GradAlign, for example, selects training problems whose gradients align with a validation gradient (Yang et al., 2026) to target downstream utility. More generally, directional agreement leaves gradient magnitude unconstrained: a subset can point in the same direction while changing the strength of the learning signal. Although gradient reconstruction is established in data subset selection (Killamsetty et al., 2021), group-relative RL adds a complication: removing responses can change the advantages of those that remain. The unresolved issue is therefore when a surrogate's notion of a good subset is consistent with preserving the original gradient, and when it favors a different learning signal.

We ask: *when is directional agreement sufficient for rollout filtering to preserve a full-group gradient, and when does scale mismatch change which subset should be preferred?* As illustrated in Figure 1, our approach treats gradient fidelity as a joint requirement on direction and magnitude, comparing directional selection with a joint-error criterion under the same retention budget. A



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

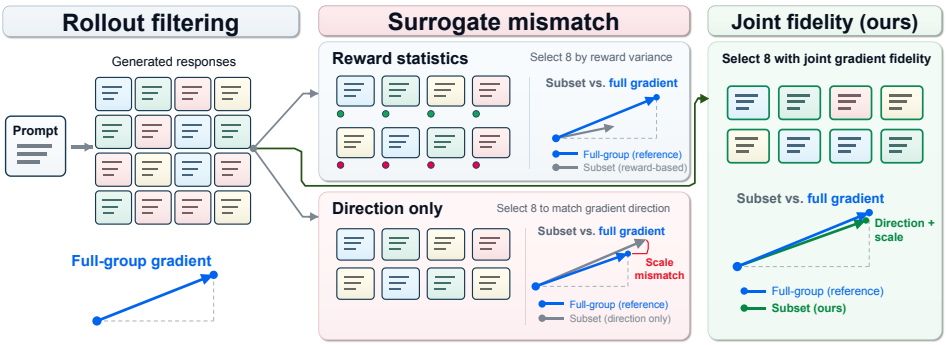



Figure 1: **Rollout filtering should preserve both gradient direction and scale.** Each criterion retains 8 of the same 16 responses. Reward-based selection can distort both direction and scale; direction-only selection can still change gradient magnitude. Our joint-error criterion favors subsets that match both properties of the full-group gradient (blue). Arrows are schematic.

controlled audit of frozen response groups separates objective mismatch from search, gradient approximation, and advantage normalization effects. Theory and empirical analyses then identify why the criteria disagree and when directional agreement becomes informative.

We make three contributions:

- **We reveal a direction–scale mismatch.** We establish realizable cases in which optimal directional alignment yields poor gradient reconstruction, together with norm conditions that make alignment reliable. Empirical analyses associate larger gaps with heterogeneous response contributions and stronger cancellation within groups.
- **We introduce a controlled audit of rollout criteria.** Matched subset comparisons and independent evaluation in the uncompressed trainable parameter space distinguish limitations of the selection objective from gradient approximation and normalization effects.
- **Our joint-error selector improves gradient fidelity.** Across eight backbone–dataset pairs spanning three models, it reduces mean squared relative gradient error by $66.8$–$75.7\%$ versus PODS variance and $42.9$–$53.7\%$ versus cosine selection, although cosine achieves the smallest mean angle (Table 1).

## 2 Related Work

**Reward-guided rollout filtering.** GRPO uses group-relative rewards to estimate advantages without a learned critic (Shao et al., 2024), motivating selective updates on generated responses. CPPO prunes completions with small absolute advantages (Lin et al., 2025); PODS maximizes retained reward variance (Xu et al., 2026b). D$^3$S combines advantage-variance selection across groups with token filtering and a dynamic schedule (Wang et al., 2026). These methods improve training efficiency using reward and advantage statistics, while retained-gradient geometry also depends on response score gradients. Our audit compares their applicable selection rules on identical completed groups with a fixed retention budget, isolating gradient fidelity from the other components of these training systems.

**Gradient matching for subset selection.** CRAIG constructs weighted coresets through a bound on gradient approximation error (Mirzasoleiman et al., 2020); GRAD-MATCH directly optimizes gradient matching using orthogonal matching pursuit (Killamsetty et al., 2021). This line establishes gradient reconstruction as a selection objective. Its standard formulation fixes per-example gradients during selection, whereas recomputing GRPO advantages makes each response's contribution depend on the retained group. We examine unweighted, fixed-size response subsets and use exhaustive search with separate proxy-selection and exact-evaluation geometries to distinguish objective mismatch from search and representation errors.

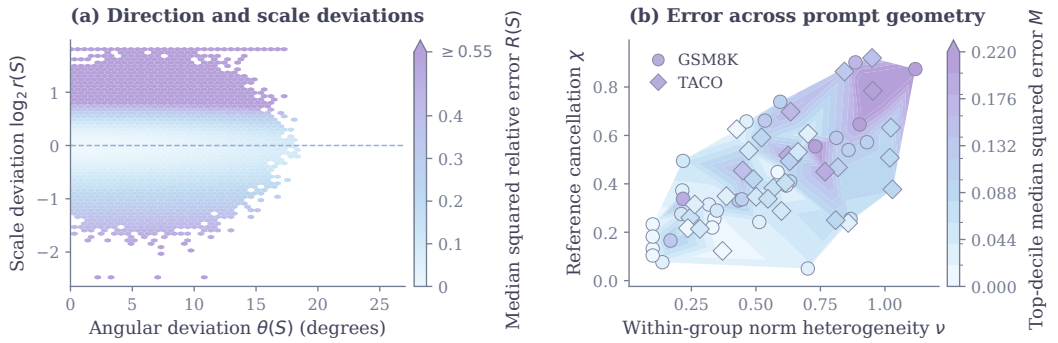



Figure 2: **Directional agreement can conceal scale mismatch. (a)** Angle versus log norm ratio for 131,072 sampled subsets; hexagon color shows median squared relative error $R$. **(b)** Prompt-level error $M_x$ across contribution-norm heterogeneity $\nu$ and cancellation $\chi$. Shapes identify datasets; background colors linearly interpolate observed prompt values. The panels use different color scales (Appendix B.7.3).

**Directional and geometric criteria.** GradAlign selects training problems by cosine similarity to a validation gradient (Yang et al., 2026), while GeoAlign replaces rollout outliers using directional consensus in a learned latent space (Zhou et al., 2026). SALT further uses gradient geometry to reweight advantages and mitigate cancellation (Chang et al., 2026). These methods target downstream improvement or update stability. We study reconstruction of a fixed, realized full-group gradient under an unweighted subset constraint. In this setting, directional agreement leaves scale unconstrained. We characterize when cosine selection loses fidelity, compare it with joint direction–scale matching, and separate unscaled error from error after reference-informed rescaling.

## 3 Preliminary Study: Directional Agreement Can Conceal Update Distortion

Does a subset pointing in nearly the same direction as the full-group gradient preserve that update? We examine frozen rollout groups from 32 GSM8K and 32 TACO development prompts, each containing 16 responses. For each prompt, we uniformly sample 2,048 distinct size-8 subsets without optimizing a selection objective or rescaling their gradients. We measure the squared relative error $R(S) = \|g_S - g_F\|^2/\|g_F\|^2$, where $g_S$ and $g_F$ are the subset and full-group reward gradients. To examine residual mismatch despite directional agreement, we retain the highest-cosine 10% of sampled subsets within each prompt for descriptive analysis; their prompt-level median error is denoted $M_x$. Appendix B.7 gives metric definitions and aggregation rules; Appendix C.5 reports additional results.

Figure 2(a) shows substantial scale variation even at small angles. Pooling the highest-cosine subsets across prompts gives a median angle of $0.82°$, yet a median $R$ of $0.054$; $21.6\%$ have $R > 0.25$. Figure 2(b) associates larger residual mismatch with more heterogeneous contribution norms and stronger cancellation within the full group. The nine prompts jointly in the lowest quartiles of these two descriptors have median $M_x = 0.011$, compared with $0.150$ for the nine jointly in the highest quartiles. The descriptors co-vary, and the descriptive association includes prompt-level exceptions.

**Nearly the same direction, a different update**

Directional agreement alone can leave substantial scale distortion. This pilot motivates evaluating direction and scale jointly.

We next compare selection criteria using a common candidate set and evaluation geometry.

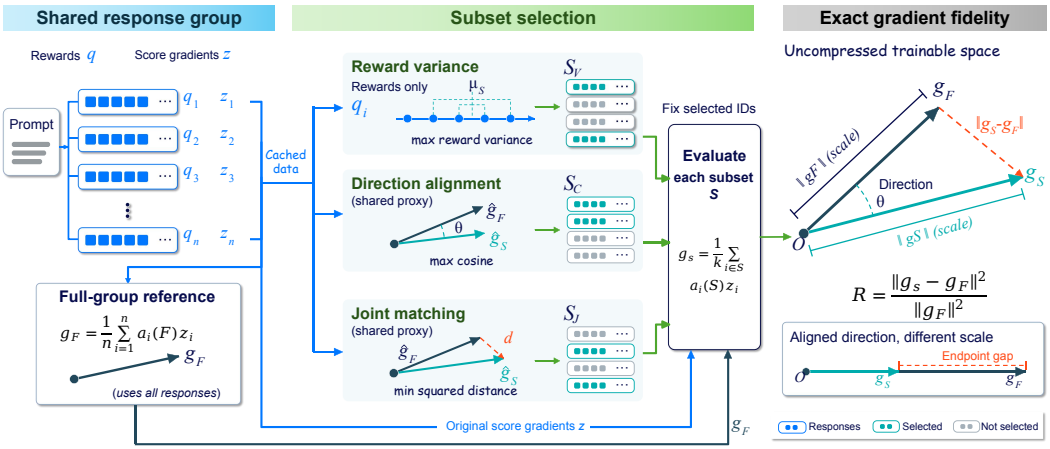



Figure 3: **Rollout-selection audit on a shared response group.** At a frozen policy, cached rewards $q_i$ and score gradients $z_i$ define the full-group reference $g_F$ for one prompt (left). Each criterion selects a size-$k$ subset: reward variance uses rewards alone; direction alignment and joint matching optimize cosine similarity and squared distance in the same proxy $\widehat{g} = Pg$ (middle). Advantages $a_i(S)$ are recomputed within each candidate subset. Once selected, the response IDs are fixed for evaluation against $g_F$ using the uncompressed trainable-space Gram matrix $K = Z^\top Z$ (right). Relative error $R$ depends on direction and scale, so aligned arrows can still have different endpoints. Response blocks, selections, and vectors are schematic.

## 4 METHOD

We compare selection criteria on the same response group, with each retaining the same number of responses. We then evaluate each subset gradient against the full-group gradient in the trainable parameter space. Exhaustive search removes search approximation; proxy and exact-geometry controls distinguish objective mismatch from representation error. Figure 3 shows the primary comparison with recomputed advantages, and Algorithm 1 in Appendix A.4 gives the per-group procedure.

### 4.1 FULL-GROUP REFERENCE AND SUBSET GRADIENTS

For a prompt $x$, let $F = \{1, \dots, n\}$ index responses generated at fixed parameters $\theta_0$. Response $y_i$ has verifier reward $q_i \in [0, 1]$ and $T_i \geq 1$ response tokens. Let $\theta \in \mathbb{R}^d$ denote the trainable parameters and $\pi_\theta$ the policy used in the loss. The response score gradient averages token log-probability gradients:

$$z_i = \frac{1}{T_i} \sum_{t=1}^{T_i} \nabla_\theta \log \pi_\theta(y_{i,t} \mid x, y_{i,<t}) \Big|_{\theta=\theta_0} \in \mathbb{R}^d. \tag{1}$$

We exclude prompt and padding tokens but include a generated end-of-sequence token. For a response truncated at the length cap, $T_i$ counts only observed response tokens. All subsets share the cached responses, token masks, rewards, and score gradients.

For a nonempty set $U \subseteq F$, define the population reward mean $\mu_U = |U|^{-1} \sum_{i \in U} q_i$ and standard deviation $\sigma_U = (|U|^{-1} \sum_{i \in U} (q_i - \mu_U)^2)^{1/2}$. Advantages can be recomputed within a subset $S$ or fixed using the full group:

$$a_i^{\mathrm{rec}}(S) = \frac{q_i - \mu_S}{\sigma_S + \varepsilon_a}, \qquad a_i^{\mathrm{fix}} = \frac{q_i - \mu_F}{\sigma_F + \varepsilon_a}, \qquad \varepsilon_a = 10^{-6}. \tag{2}$$

The corresponding gradients are

$$g_S^{\mathrm{rec}} = \frac{1}{|S|} \sum_{i \in S} a_i^{\mathrm{rec}}(S) z_i, \qquad g_S^{\mathrm{fix}} = \frac{1}{|S|} \sum_{i \in S} a_i^{\mathrm{fix}} z_i, \qquad g_F = g_F^{\mathrm{rec}} = g_F^{\mathrm{fix}}. \tag{3}$$

The primary comparison recomputes advantages, as in the PODS subset update (Xu et al., 2026b); fixed advantages provide a control. We write $g_S$ when a statement holds under either convention, used consistently within that comparison.

Treating advantages as constants, $g_S$ is the ascent direction of the response-averaged reward surrogate. It equals the reward-gradient component of Group Relative Policy Optimization (GRPO) (Shao et al., 2024) at the initial step, when probability ratios equal one, excluding KL regularization and optimizer transformations. Generation temperature affects the sampled responses; the loss uses untempered policy log probabilities. Thus $g_F$ is the reward-gradient reference for this realized group.

### 4.2 Separating direction and scale

For nonzero $g_F$ and $g_S$, define the norm ratio $r(S) = \|g_S\|/\|g_F\|$ and cosine similarity $c(S) = g_S^\mathsf{T} g_F/(\|g_S\|\|g_F\|)$. The squared relative error separates into scale and directional terms:

$$R(S) = \frac{\|g_S - g_F\|^2}{\|g_F\|^2} = \underbrace{(r(S)-1)^2}_{B(S)} + \underbrace{2r(S)(1-c(S))}_{A(S)}. \tag{4}$$

$B(S)$ is the squared norm mismatch; $A(S)$ is the directional contribution weighted by the norm ratio. We report the angle $\arccos c(S)$ separately because a smaller angle need not yield a smaller $A(S)$ when norms differ. Perfect alignment leaves error $(r-1)^2$, while matching the norm leaves error $2(1-c)$. Small joint error requires control of both terms. Appendix D.1.1 proves the decomposition and gives sufficient conditions and converse bounds.

For $g_S = 0$, we set $r = 0$, $R = B = 1$, and $A = 0$; the angle is undefined. Groups with numerically unresolved reference norms are excluded from relative-error summaries, but their absolute errors and exclusion counts are retained. Appendix A.3 defines the thresholds and handles small nonzero gradients using their actual errors.

### 4.3 Selection over a common candidate set

All criteria select from $\mathcal{S}_k = \{S \subseteq F : |S| = k\}$. The gradient-based criteria share a fixed linear map $P \in \mathbb{R}^{m\times d}$, with $\widehat{z}_i = Pz_i$ and $\widehat{g}_S = Pg_S$, and use the same features and advantages. The primary proxy is a CountSketch with $m = 128$ (Appendix A.2). Reward-variance selection uses only rewards:

$$S_V \in \arg\max_{S\in\mathcal{S}_k} \sigma_S^2. \tag{5}$$

With fixed $\varepsilon_a > 0$, maximizing reward variance is equivalent to maximizing advantage variance under either convention (Appendix D.1.4).

When $\|\widehat{g}_F\| > \tau_P$, cosine selection uses candidates $\mathcal{C}_P = \{S \in \mathcal{S}_k : \|\widehat{g}_S\| > \tau_P\}$, where $\tau_P$ is the numerical threshold in Appendix A.3:

$$S_C \in \arg\max_{S\in\mathcal{C}_P} \frac{\widehat{g}_S^\mathsf{T}\widehat{g}_F}{\|\widehat{g}_S\|\|\widehat{g}_F\|}. \tag{6}$$

Joint-error selection matches both direction and scale:

$$S_J \in \arg\min_{S\in\mathcal{S}_k} \frac{\|\widehat{g}_S - \widehat{g}_F\|^2}{\|\widehat{g}_F\|^2}. \tag{7}$$

The denominator is constant within a group, so we minimize the numerator. For $n = 16$ and $k = 8$, we enumerate all $12{,}870$ subsets. A shared seeded ordering breaks ties in computed objective values; near-equal values keep their numerical ordering. If the full proxy norm does not exceed $\tau_P$, both gradient criteria return the first subset in that ordering. Cosine selection uses the same fallback when $\mathcal{C}_P$ is empty. Each fallback is flagged (Appendix A.4).

We evaluate uniform selection by averaging over all subsets. Exact-geometry controls replace sketch inner products with trainable-space inner products: exact joint selection gives a lower bound on $R$, while exact cosine selection isolates directional optimization from proxy distortion. This exact geometry is unavailable to the main selectors.

### 4.4 EXACT EVALUATION IN THE TRAINABLE PARAMETER SPACE

After selection, we evaluate the retained gradients using the Gram matrix $K = Z^\mathsf{T} Z \in \mathbb{R}^{n \times n}$, where $Z = [z_1, \ldots, z_n] \in \mathbb{R}^{d \times n}$. For a size-$k$ subset, write $a_i(S)$ for the chosen advantages. Set $(w_S)_i = a_i(S)/k$ on $S$ and zero elsewhere, and $(w_F)_i = a_i(F)/n$. Then $g_S = Z w_S$, and

$$\begin{aligned} \|g_S\|^2 &= w_S^\mathsf{T} K w_S, & g_S^\mathsf{T} g_F &= w_S^\mathsf{T} K w_F, \\ R(S) &= \frac{(w_S - w_F)^\mathsf{T} K (w_S - w_F)}{w_F^\mathsf{T} K w_F}. \end{aligned} \tag{8}$$

These expressions use the same uncompressed reference for every subset without reconstructing dense candidate gradients. Gram entries are accumulated in float64 from cached response gradients: "exact" means uncompressed trainable space, subject to floating-point arithmetic. Appendix A.1 describes blockwise construction and direct-gradient checks.

After precomputing $K w_F$ and $w_F^\mathsf{T} K w_F$, each subset norm costs $\mathcal{O}(k^2)$ and its reference inner product costs $\mathcal{O}(k)$. Exhaustive evaluation costs $\mathcal{O}(\binom{n}{k} k^2)$, excluding generation, gradient extraction, and $\mathcal{O}(dn^2)$ Gram construction. These are the computational costs of the offline audit.

### 4.5 CONTROLS FOR OBJECTIVE, NORMALIZATION, AND PROXY EFFECTS

**Post-selection scale adjustment.** To measure how much error scale adjustment alone can remove, fix a nonzero selected gradient and choose the best nonnegative scalar:

$$\alpha^*(S) = \max\left(0, \frac{g_S^\mathsf{T} g_F}{\|g_S\|^2}\right), \qquad R^*(S) = \min_{\alpha \geq 0} \frac{\|\alpha g_S - g_F\|^2}{\|g_F\|^2} = 1 - \max(c(S), 0)^2. \tag{9}$$

Positive scaling preserves direction; when $c(S) \leq 0$, the optimum $\alpha^* = 0$ suppresses the update. For $g_S = 0$, we set $\alpha^* = 0$ and $R^* = 1$. The removable error $R(S) - R^*(S)$ uses the full reference and serves only as an offline diagnostic.

On a common set of nonzero candidates in the same geometry, maximizing cosine also minimizes $R^*$ because $1 - \max(c, 0)^2$ is nonincreasing in $c$. The Coder/GSM8K decomposition reports $R$, its $B$ and $A$ contributions, and the removable error. Appendix D.1.2 proves this ordering and explains why removable error differs from $B$.

**Advantage normalization.** To isolate normalization effects, we hold each selected subset fixed and evaluate both advantage conventions on Coder/GSM8K and Coder/TACO. Writing $\bar{z}_S = |S|^{-1} \sum_{i \in S} z_i$, Equation (3) gives

$$g_S^{\mathrm{rec}} = \frac{\sigma_F + \varepsilon_a}{\sigma_S + \varepsilon_a} g_S^{\mathrm{fix}} + \frac{\mu_F - \mu_S}{\sigma_S + \varepsilon_a} \bar{z}_S. \tag{10}$$

Recomputing advantages changes both scale and centering; the second term can also change direction. Testing all four selection/evaluation convention pairs separates these effects from changes in subset membership (Appendix B.6.1).

**Proxy and synthetic controls.** On Coder/GSM8K and Coder/TACO, we vary the selection geometry from sketches with $m \in \{32, 128, 512\}$ to exact geometry, keeping responses, rewards, and $K$ fixed. A paired synthetic control varies the angle and norm ratio between modes while fixing group balance and reward coupling; paired error gaps are decomposed into $B$ and $A$ (Appendix B.6.3). Appendix C reports both controls. Paired mean intervals use prompts or synthetic replicate IDs as independent units, with aggregation defined in Appendix B.5.2.

### 4.6 THEORETICAL SUPPORT FOR JOINT GRADIENT FIDELITY

For a realized group at a fixed policy, let $C$ maximize exact cosine over nonzero size-$k$ subset gradients, and let $S_\star$ minimize unscaled error $R$ over all size-$k$ subsets. With recomputed advantages and detached rewards, directional selection can be arbitrarily worse than joint matching:

**Theorem 1** (Directional agreement does not ensure reconstruction)**.** *For every $\varepsilon_a > 0$ and $0 < t \leq 1$, a one-token linear-softmax policy admits a positive-probability group with $n = 16$, balanced*

*binary rewards, $k = 8$, and $g_F = e_1$. For each fixed $\varepsilon_a$, its score-gradient norms are bounded independently of $t$. The cosine maximizer $C$ is unique, with $R(C) = 4$ and $0 < R(S_\star) \le t^2/16$. Hence $R(C)/R(S_\star) \ge 64/t^2$: no finite constant uniformly bounds cosine selection's error relative to the optimum.*

**Proof sketch.** Construct two reward-balanced blocks with gradients $3e_1$ and $-e_1$. Orthogonal score perturbations make the first block the unique cosine maximizer. A balanced subset spanning both blocks has error $t^2/16$, while no subset has zero error. Centered scores realize this construction through a uniform softmax policy (Appendix D.2).

Controlling candidate norms restores a guarantee for directional selection. Let $\mathcal{D}$ be a common finite, nonempty family of nonzero candidates, with $g_F \neq 0$ and $|r(S) - 1| \le \delta < 1$ for every $S \in \mathcal{D}$. If $\rho = \min_{S \in \mathcal{D}} R(S) \le 2\delta - \delta^2$, an exact cosine maximizer on $\mathcal{D}$ satisfies $R(C) - \rho \le (1 + \delta - \sqrt{1-\rho})^2$ (Theorem 6). Appendix D provides the full calibration theorem, reward-only minimax analysis, and proofs. These guarantees use exact geometry; extending them to sketches requires distortion control.

# 5 Experiments

## 5.1 Experimental setup

We audit frozen-policy reward gradients on GSM8K, MATH, and TACO (Cobbe et al., 2021; Hendrycks et al., 2021; Li et al., 2023) using Qwen2.5-Coder-3B-Instruct, Qwen2.5-Math-7B, and Llama-3.1-8B-Instruct (Hui et al., 2024; Yang et al., 2024; Grattafiori et al., 2024), with rank-16 LoRA evaluation and a 128-dimensional CountSketch. Squared relative error $R$ and angle measure full-group gradient fidelity; $B$, $A$, and $R^*$ diagnose error composition and reference-informed rescaling. Each dataset supplies 120 development prompts and 480 held-out audit prompts. Prompt-level metrics average 25 sketch/tie seed combinations; paired intervals use 10,000 prompt-bootstrap resamples.

We compare uniform, percentile, highest-reward, and PODS variance selection (Xu et al., 2026b), CPPO pruning (Lin et al., 2025), and cosine and joint-error selection. Each retains eight of the same sixteen responses, with full-group GRPO as the reference (Shao et al., 2024); gradient criteria share a proxy. Appendices B and C give the matched protocol, implementation details, paired uncertainty estimates, and controls.

## 5.2 Q1: Which selection criterion best preserves the gradient?

Joint-error selection has the lowest mean $R$ among non-privileged selectors in all eight backbone–dataset pairs (Table 1): 66.8–75.7% below variance and 42.9–53.7% below cosine. Cosine achieves the smallest mean angle, highlighting the trade-off between directional alignment and fidelity to the complete update. Paired pointwise intervals support the error rankings (Appendix C.1).

## 5.3 Q2: Why does directional alignment leave gradient error?

**Error composition.** On Coder/GSM8K, norm mismatch accounts for 97.6% of cosine's mean error ($B = 0.123$ of $R = 0.126$; Table 2). Joint selection reduces $B$ to 0.051 and $R$ to 0.061, while $A$ rises from 0.003 to 0.010: the scale improvement outweighs the directional trade-off.

**Approximation and normalization controls.** Exact selection on Coder/GSM8K gives $R = 0.119$ for cosine and 0.041 for joint. The cosine–joint gap remains positive under exact geometry and all four tested selection/evaluation normalization pairs on both Coder/GSM8K and Coder/TACO (Appendices C.2–C.3).

**Reference-informed rescaling.** Optimal nonnegative rescaling reverses the ranking: cosine has lower residual error ($R^\star = 0.002$ versus 0.008). All nine paired synthetic settings reproduce this direction–scale trade-off and ranking reversal (Appendix C.4). Rescaling and exact selection use privileged full-reference geometry.

Table 1: **Matched comparison of rollout-selection criteria.** Mean squared relative error $R$ and angle $\theta$ (degrees) under $16 \to 8$ retention, recomputed advantages, and uncompressed LoRA evaluation. Purple rows are ours; green subscripts show percentage reductions from the strongest baseline (PODS variance). Bold/underlined: best/second-best non-privileged results. Qwen2.5-Math-7B covers mathematics only.

| | | GSM8K | | MATH | | TACO | |
|---|---|---|---|---|---|---|---|
| Backbone | Selection criterion | $R \downarrow$ | $\theta \downarrow$ | $R \downarrow$ | $\theta \downarrow$ | $R \downarrow$ | $\theta \downarrow$ |
| *Qwen2.5-Coder-3B-Instruct* | Uniform (Xu et al., 2026b) | 0.468 | 23.6 | 0.551 | 27.9 | 0.637 | 31.4 |
| | Percentile (Xu et al., 2026b) | 0.311 | 18.8 | 0.372 | 21.7 | 0.426 | 24.6 |
| | Highest reward (Xu et al., 2026b) | 0.794 | 26.8 | 0.748 | 30.1 | 0.606 | 28.9 |
| | PODS variance (Xu et al., 2026b) | 0.238 | 15.7 | 0.289 | 18.9 | 0.341 | 22.2 |
| | CPPO (adapted) (Lin et al., 2025) | 0.367 | 19.1 | 0.412 | 22.4 | 0.463 | 24.4 |
| | Cosine (this work) | <u>0.126</u> +47.1% | **2.3** +85.4% | <u>0.158</u> +45.3% | **3.1** +83.6% | <u>0.196</u> +42.5% | **3.9** +82.4% |
| | Joint error (this work) | **0.061** +74.4% | <u>5.0</u> +68.2% | **0.083** +71.3% | <u>6.2</u> +67.2% | **0.112** +67.2% | <u>7.3</u> +67.1% |
| *Qwen2.5-Math-7B* | Uniform (Xu et al., 2026b) | 0.434 | 21.9 | 0.503 | 25.1 | *Not evaluated* | |
| | Percentile (Xu et al., 2026b) | 0.282 | 17.0 | 0.327 | 19.6 | | |
| | Highest reward (Xu et al., 2026b) | 0.761 | 25.6 | 0.724 | 28.4 | | |
| | PODS variance (Xu et al., 2026b) | 0.214 | 14.3 | 0.251 | 16.6 | | |
| | CPPO (adapted) (Lin et al., 2025) | 0.336 | 17.7 | 0.379 | 20.2 | | |
| | Cosine (this work) | <u>0.109</u> +49.1% | **2.0** +86.0% | <u>0.141</u> +43.8% | **2.7** +83.7% | | |
| | Joint error (this work) | **0.052** +75.7% | <u>4.4</u> +69.2% | **0.071** +71.7% | <u>5.5</u> +66.9% | | |
| *Llama-3.1-8B-Instruct* | Uniform (Xu et al., 2026b) | 0.529 | 26.4 | 0.601 | 29.7 | 0.694 | 34.1 |
| | Percentile (Xu et al., 2026b) | 0.346 | 20.9 | 0.394 | 23.4 | 0.468 | 27.3 |
| | Highest reward (Xu et al., 2026b) | 0.822 | 29.5 | 0.779 | 32.3 | 0.658 | 31.8 |
| | PODS variance (Xu et al., 2026b) | 0.267 | 17.5 | 0.309 | 20.6 | 0.374 | 24.9 |
| | CPPO (adapted) (Lin et al., 2025) | 0.398 | 21.2 | 0.441 | 23.9 | 0.501 | 27.0 |
| | Cosine (this work) | <u>0.149</u> +44.2% | **2.9** +83.4% | <u>0.176</u> +43.0% | **3.4** +83.5% | <u>0.219</u> +41.4% | **4.3** +82.7% |
| | Joint error (this work) | **0.069** +74.2% | <u>5.7</u> +67.4% | **0.091** +70.6% | <u>6.7</u> +67.5% | **0.124** +66.8% | <u>8.2</u> +67.1% |

**Distributional evidence.** Figure 4 shows the same trade-off in a separate Coder/GSM8K distributional sample, whose aggregates differ from Table 2. Cosine concentrates at small $A$ and elevated $B$; joint selection lowers $B$ enough to reduce mean $R$ by 57.3% (0.1883 to 0.0805). PODS variance and CPPO have larger means in both components. Appendices B.8 and C.6 give sample counts, density estimation, and component summaries.

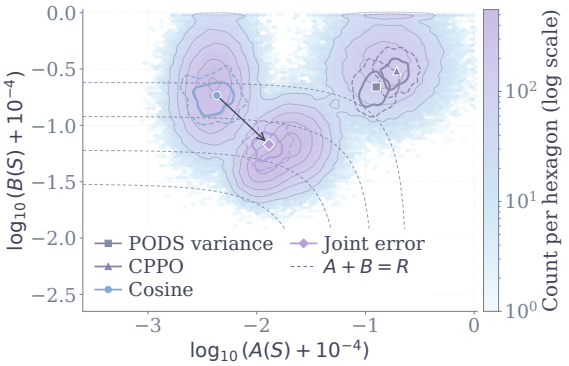



Figure 4: **Direction–scale distributions on Coder/GSM8K.** Hexagons show log-scaled counts; contours enclose 50% (solid) and 85% (dashed) of fitted log-coordinate density. Markers show transformed arithmetic means, with an arrow from cosine to joint. Gray curves satisfy $A + B = R$.

### 5.4 Q3: When are gains larger?

Across 3,486 instances from eight backbone–dataset pairs, pair-adjusted joint-matching gains are larger when norm heterogeneity $\nu_x$ and cancellation $\chi_x$ are jointly high (Figure 5). Median gains are 0.1298 in the joint upper-quartile group and 0.0480 in the joint lower-quartile group (difference: 0.0818). This joint association involves covarying descriptors; Appendices B.9 and C.7 detail averaging over 25 matched seed combinations, pair adjustment, and uncertainty.

### 5.5 Q4: Can response interactions predict objective mismatch?

Across 3,486 instances, we ask how the exact-selection gap $\Delta R_{\text{obj}} = R(S_C^{\text{ex}}) - R(S_J^{\text{ex}})$ varies with the contribution Gram matrix's entropy effective rank $r_{\text{eff}}$ and antagonism $\zeta$, the negative share of pairwise interaction magnitude (Appendix B.10).

Table 2: **Joint-error decomposition and privileged controls.** Qwen2.5-Coder-3B-Instruct on GSM8K with matched $16 \to 8$ retention. Columns report original-update fidelity, error components, and optimal nonnegative rescaling. Purple rows are ours; gray rows use privileged exact geometry. Bold/underlined: best/second-best non-privileged error metrics.

| | Update fidelity | | Error decomposition | | | Optimal rescaling | | |
|---|---|---|---|---|---|---|---|---|
| Selection criterion | Mean $R \downarrow$ | Median $R \downarrow$ | Norm $\bar{B} \downarrow$ | Direction $\bar{A} \downarrow$ | Norm share (%) | Residual $\bar{R}^\star \downarrow$ | Removable $\Delta\bar{R}$ | Removed (%) |
| *Baseline selectors* | | | | | | | | |
| Uniform | 0.468 | 0.356 | 0.278 | 0.190 | 59.4 | 0.166 | 0.302 | 64.5 |
| Percentile | 0.311 | 0.236 | 0.178 | 0.133 | 57.2 | 0.108 | 0.203 | 65.3 |
| Highest reward | 0.794 | 0.812 | 0.666 | 0.128 | 83.9 | 0.558 | 0.236 | 29.7 |
| PODS variance | 0.238 | 0.174 | 0.145 | 0.093 | 60.9 | 0.078 | 0.160 | 67.2 |
| CPPO (adapted) | 0.367 | 0.281 | 0.219 | 0.148 | 59.7 | 0.112 | 0.255 | 69.5 |
| *Proxy-gradient selectors* | | | | | | | | |
| Cosine (this work) | <u>0.126</u> | <u>0.086</u> | <u>0.123</u> | **0.003** | 97.6 | **0.002** | 0.124 | 98.4 |
| Joint error (this work) | **0.061** | **0.043** | **0.051** | <u>0.010</u> | 83.6 | <u>0.008</u> | 0.053 | 86.9 |
| *Privileged exact-geometry controls* | | | | | | | | |
| Exact cosine (control) | 0.119 | 0.079 | 0.118 | 0.001 | 99.2 | 0.001 | 0.118 | 99.2 |
| Exact joint error (oracle) | 0.041 | 0.029 | 0.033 | 0.008 | 80.5 | 0.006 | 0.035 | 85.4 |

The low-rank, high-antagonism quartile has mean gap 0.1458, versus 0.0524 in the opposite regime (2.78×; Figure 6). Matched quadratic ridge models with prompt-grouped cross-validation yield out-of-fold $R^2 = 0.821$ using $(r_{\text{eff}}, \zeta)$ versus 0.739 using $(\nu, \chi)$.

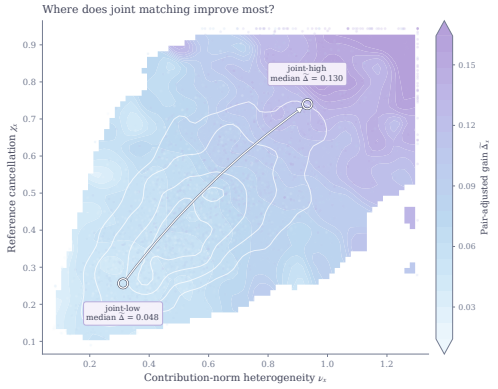



Figure 5: **Joint-matching gains across rollout-group geometry.** Points show pair-adjusted gains $\widetilde{\Delta}_x$. Circles mark coordinate-wise medians of joint-quartile groups, labeled by median gain. Appendix B.9 details the shaded means and density contours.

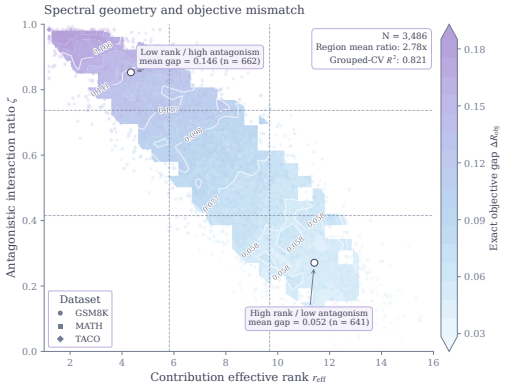



Figure 6: **Contribution spectrum and exact-geometry objective mismatch.** Points show $\Delta R_{\text{obj}}$ by dataset (shape); dashed lines mark marginal quartiles. Callouts give regime means and counts; the inset reports grouped out-of-fold $R^2$ (Appendix B.10).

## 6 Conclusion

Our fixed-group audit and theory identify scale mismatch as a source of gradient reconstruction error and separate objective mismatch from search and approximation effects. Across the evaluated models and tasks, joint matching reduced gradient error while direction-only selection achieved closer alignment. Joint-matching gains are larger in groups with both heterogeneous contribution norms and strong reference cancellation. These findings support evaluating both gradient magnitude and direction when comparing rollout-filtering criteria.

### AI USE STATEMENT

We used generative AI tools to assist with checking mathematical proofs, reviewing related literature, analyzing and interpreting results, and polishing the language of the manuscript. The authors take responsibility for the correctness of the proofs, the accuracy of the reported results and citations, and the final content of the paper.

## REFERENCES

Powei Chang, Jinpeng Zhang, Chaoqun Sun, MiniWell Tsao, Lianrui Li, Jianxiang Xiang, Chenyu Wang, Yukang Gao, and Dongying Kong. SALT: When more rollouts don't help in group-based policy optimization and how to make them matter. *arXiv preprint arXiv:2606.05800*, 2026. URL https://arxiv.org/abs/2606.05800.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

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

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, et al. Qwen2.5-Coder technical report. *arXiv preprint arXiv:2409.12186*, 2024. URL https://arxiv.org/abs/2409.12186.

Krishnateja Killamsetty, Durga S, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. GRAD-MATCH: Gradient matching based data subset selection for efficient deep model training. In *Proceedings of the International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5464–5474, 2021. URL https://proceedings.mlr.press/v139/killamsetty21a.html.

Rongao Li, Jie Fu, Bo-Wen Zhang, Tao Huang, Zhihong Sun, Chen Lyu, Guang Liu, Zhi Jin, and Ge Li. TACO: Topics in algorithmic COde generation dataset. *arXiv preprint arXiv:2312.14852*, 2023. URL https://arxiv.org/abs/2312.14852.

Zhihang Lin, Mingbao Lin, Yuan Xie, and Rongrong Ji. CPPO: Accelerating the training of group relative policy optimization-based reasoning models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 61043–61068, 2025. doi: 10.52202/085713-2040. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/57ec371599f3d4123c9b054ffa3ef0ca-Abstract-Conference.html.

Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6950–6960. PMLR, 2020. URL https://proceedings.mlr.press/v119/mirzasoleiman20a.html.

Zhihong Shao, Peiyi Wang, Qihao Zhu, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Chao Wang, Tao Yang, Hongtao Tian, Yunsheng Shi, Qiyao Ma, Xiaotao Liu, Ting Yao, and Wenbo Ding. Learning more with less: A dynamic dual-level down-sampling framework for efficient policy optimization. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/9c3156498c31f8c46cb47cffb22df683-Abstract-Conference.html.

Haobo Xu, Sirui Chen, Ruizhong Qiu, Yuchen Yan, Chen Luo, Monica Xiao Cheng, Jingrui He, and Hanghang Tong. Prune as you generate: Online rollout pruning for faster and better RLVR. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13876–13893, 2026a. doi: 10.18653/v1/2026.acl-long.632. URL https://aclanthology.org/2026.acl-long.632/.

Yixuan Even Xu, Yash Savani, Fei Fang, and J. Zico Kolter. Not all rollouts are useful: Down-sampling rollouts in LLM reinforcement learning. *Transactions on Machine Learning Research*, 2026b. URL https://arxiv.org/abs/2504.13818. arXiv:2504.13818, version 5.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, et al. Qwen2.5-Math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024. URL https://arxiv.org/abs/2409.12122.

Ningyuan Yang, Weihua Du, Weiwei Sun, Sean Welleck, and Yiming Yang. GradAlign: Gradient-aligned data selection for LLM reinforcement learning. In *Conference on Language Modeling*, 2026. URL https://arxiv.org/abs/2602.21492. arXiv:2602.21492, version 2.

Ting Zhou, Zhenqing Ling, Yiyang Zhao, Ying Shen, and Daoyuan Chen. GeoAlign: Geometric rollout curation for robust LLM reinforcement learning. In *International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/60634. arXiv:2606.26917.

# Appendix

## Contents

SUPPLEMENTARY MATERIAL

PAGE

**A Technical Details** . . . 13
- A.1 Score-gradient cache and Gram construction . . . 13
- A.2 CountSketch and seed conventions . . . 13
- A.3 Degenerate references, small gradients, and invalid records . . . 13
- A.4 Enumeration, tie resolution, and returned records . . . 14
- A.5 Computational requirements . . . 15

**B Experimental Details** . . . 16
- B.1 Benchmarks, splits, and verification . . . 16
- B.2 Metrics and reporting units . . . 16
- B.3 Comparison methods and fairness . . . 16
- B.4 Backbones, generation, and numerical implementation . . . 17
- B.5 Audit inputs and statistical summaries . . . 18
- B.6 Controls for normalization and representation . . . 19
- B.7 Preliminary-study protocol . . . 20
- B.8 Direction–scale distribution protocol . . . 21
- B.9 Geometry-stratified gain protocol . . . 21
- B.10 Spectral geometry and prediction protocol . . . 22

**C Additional Experimental Results** . . . 24
- C.1 Paired uncertainty in the main comparisons . . . 24
- C.2 Selection and evaluation normalization . . . 25
- C.3 Proxy distortion and exact-geometry controls . . . 26
- C.4 Paired angle–norm intervention . . . 27
- C.5 Preliminary-study numerical summaries . . . 27
- C.6 Direction–scale distribution results . . . 28
- C.7 Geometry-stratified joint-matching gains . . . 28
- C.8 Spectral regimes and objective mismatch . . . 28

**D Proofs** . . . 30
- D.1 Properties of the gradient comparison . . . 30
- D.2 Proof of the directional obstruction at the audit budget . . . 31
- D.3 The exact information cost of reward-only selection . . . 34
- D.4 Sharp calibration under norm control . . . 39
- D.5 Additional structure of the complete rollout family . . . 42
- D.6 Complementary subsets at a half-group budget . . . 44

# A TECHNICAL DETAILS

## A.1 SCORE-GRADIENT CACHE AND GRAM CONSTRUCTION

The gradient cache is indexed by prompt and response ID. Each entry contains response tokens, the response mask, reward, and the gradient in Equation (1). Parameter values remain fixed while automatic differentiation is enabled for the trainable coordinates. The loss used for extraction is negative mean response-token log probability; its gradient is negated to obtain $z_i$. Rewards and advantages are detached. Gradients from different responses are cleared between backward passes.

Trainable tensors are ordered lexicographically by their fully qualified parameter names and flattened in row-major order. The model forward and backward computations use float32 with stochastic layers disabled. Each response gradient is written to a CPU-backed float32 array; feature aggregation and inner products use float64. For a coordinate block $b$, let $Z_b$ contain the corresponding entries of all $n$ gradients. Accumulating $K = \sum_b Z_b^\mathsf{T} Z_b$ constructs the uncompressed Gram matrix while keeping a single block in working memory. The cache requires $\mathcal{O}(nd)$ storage; a block with $b_d$ coordinates requires $\mathcal{O}(nb_d + n^2)$ working memory, excluding model execution. The reference weights always divide by $n$, and subset weights always divide by $k$.

For evaluation, cache $v = Kw_F$ and $q_F = w_F^\mathsf{T} v$. For each subset, compute $u_S = w_S^\mathsf{T} K w_S$ and $b_S = w_S^\mathsf{T} v$. These determine the norm ratio, normalized signed projection $p(S) = b_S/q_F$, and error $R = (u_S + q_F - 2b_S)/q_F$. The equivalent form $A = 2r - 2p$ avoids division by the subset norm. Angle is measured in degrees as $(180/\pi)\arccos c$ after clipping roundoff to $[-1, 1]$. For $u_S > 0$, the scalar residual is evaluated as $R^* = 1 - \max(b_S, 0)^2/(u_S q_F)$ without multiplying the gradient by a potentially large $\alpha^*$.

The numerical check reconstructs selected gradients directly from the cached response vectors, including each criterion's selected subset and the exact joint minimizer. Direct and Gram squared norms and cross inner products must agree within $10^{-8}L^2$, where $L$ is the sum of the absolute-coefficient-weighted response norms in the checked linear combinations, with a floor of $10^{-12}$. A failed check invalidates the group.

## A.2 COUNTSKETCH AND SEED CONVENTIONS

For coordinate $j \in \{0, \ldots, d-1\}$, assign a bucket $h_s(j) \in \{0, \ldots, m-1\}$ and a sign $\xi_s(j) \in \{-1, 1\}$. The linear map is

$$\widehat{z}_{i,b} = \sum_{j : h_s(j) = b} \xi_s(j) z_{i,j}, \qquad b = 0, \ldots, m-1. \tag{A.1}$$

The practical deterministic assignment uses SHA-256 on the UTF-8 strings `bucket:s:j` and `sign:s:j`, with decimal integers substituted for $s, j$ and no spaces. Interpret the first eight digest bytes as an unsigned little-endian integer. Its remainder modulo $m$ defines the bucket; an even sign integer gives $+1$, and an odd integer gives $-1$. The two prefixes separate bucket and sign assignments. No scaling factor is applied. Since the map is linear, sketching before advantage aggregation gives exactly the same mathematical vector as sketching $g_S$.

Sketch seeds are $s \in \{0, 1, 2, 3, 4\}$. The same digest is reused across all dimensions, so the bucket partitions are nested for $m \in \{32, 128, 512\}$. All prompts reuse these seeds. Proxy inner products are computed from $\widehat{K} = \widehat{Z}^\mathsf{T} \widehat{Z}$, keeping selection independent of $K$. Numerical verification compares $\widehat{Z} w_S$ with the sketch of an explicitly aggregated selected gradient.

## A.3 DEGENERATE REFERENCES, SMALL GRADIENTS, AND INVALID RECORDS

For a Gram matrix $H$ representing the geometry currently used, define the reference scale and threshold by

$$s_H = \frac{1}{n} \sum_{i=1}^{n} \sqrt{\max(H_{ii}, 0)}, \qquad \tau_H = 10^{-12} + 10^{-8} s_H. \tag{A.2}$$

The absolute term is in the gradient units of that representation. Use $H = \widehat{K}$ for proxy selection and $H = K$ for exact evaluation; thus $\tau_P = \tau_{\widehat{K}}$. A reference is eligible for normalized quantities

exactly when its norm exceeds its own threshold. Thresholds are fixed constants, with no tuning on held-out errors. The same thresholds are used in the main comparison and focused controls.

For a selected gradient below threshold, retain its actual squared norm, absolute error, and relative error whenever the reference is eligible. Suppress its angle from directional summaries. An exactly zero gradient receives $R = B = R^* = 1$ and $A = 0$. Exact-zero and below-threshold events are logged separately. For small nonzero gradients, the scalar residual remains a numerical diagnostic and is flagged with the same threshold status; $\alpha^*$ is not applied as an update.

Cancellation-prone quadratic forms require a direct check. If a computed squared norm is at most $\tau_H^2$, recompute the linear combination explicitly in the relevant representation before deciding its status. Negative squared norms, negative squared errors, and violations of the Cauchy–Schwarz bound also trigger direct recomputation. Use float64 blockwise accumulation from the cache for exact geometry and direct feature aggregation for the proxy. Clip only residual floating-point excursions after this check. Nonfinite gradients, rewards outside $[0, 1]$, and responses without an observed token invalidate the group. Constant-reward groups remain in absolute-error and count summaries; they have zero reference gradients and cannot define relative errors.

### A.4 Enumeration, tie resolution, and returned records

For tie seed $t \in \{0, 1, 2, 3, 4\}$, order response IDs by the lexicographic bytes of SHA-256 applied to `tie:t:i`; resolve a digest collision by ascending response ID. Sort each subset's response ranks, then order the resulting rank tuples lexicographically. This induces one reward-independent ordering of all subsets. Objective values are computed in float64 with responses accumulated in ascending response ID. Equal computed objective values use the common subset ordering; there is no positive tie tolerance. Exhaustive search returns the optimum of the computed objective over its admissible candidates, subject to floating-point accuracy.

Algorithm 1 gives one per-group comparison. The main results use recomputed advantages. Both conventions are used only in the focused Coder/GSM8K and Coder/TACO normalization control; the proxy control uses recomputed advantages. Within each specified condition, the procedure is:

1. Load the immutable tokens, masks, rewards, score gradients, and parameter metadata. Construct $K$ and each requested $\widehat{K}$ from the same gradient cache. Compute full-group advantages and references.
2. Enumerate $\mathcal{S}_k$ once. Cache each subset's reward mean and population variance. Form the weights required by that condition.
3. Evaluate the three selection objectives in their permitted representations. For an unresolved full proxy, return the first ordered subset for both gradient criteria. With a resolved reference but no admissible cosine candidate, apply that fallback only to cosine selection. Record each fallback reason.
4. Save selected response IDs before accessing exact-geometry evaluation. Evaluate these subsets with $K$, retaining eligibility and degeneracy flags. Run exact-geometry cosine and joint controls separately with their own labels.
5. Accumulate the uniform distribution from every subset. For undefined angles, retain their probability mass and compute the angular mean conditional on a defined angle. Other defined metrics retain all candidates.
6. In the normalization control, evaluate each saved subset under both advantage conventions. Repeat the natural-data comparison across the five sketch seeds and five tie seeds, taking their Cartesian product. Return paired metric records, selected IDs, and exclusion flags.

Exact cosine uses the threshold associated with $K$; exact joint selection includes all subsets. Its relative-error minimum is a lower bound for every evaluated selector on a reference-eligible group. The rescaling ordering in Appendix D.1.2 is checked only on a common admissible candidate set. Fallback subsets remain in error summaries and retain their proxy-degeneracy flags in the audit records.

**Algorithm 1** Rollout-selection audit for one frozen response group

**Input:** Prompt $x$, checkpoint $\theta_0$, cached responses $\{(y_i, T_i, q_i)\}_{i=1}^n$;
subset size $k$, sketch $P$, advantage convention $b \in \{\mathrm{rec}, \mathrm{fix}\}$,
seeded subset ordering $\prec$ (Appendix A.4).
**Output:** Selected subsets, paired evaluation records, uniform reference, exact controls, and flags.

Compute $z_i$ by Equation (1); set $Z = [z_1, \ldots, z_n]$, $\widehat{Z} = PZ$.
Form $(w_F)_i = a_i(F)/n$; set $\widehat{g}_F = \widehat{Z} w_F$ and $\widehat{K} = \widehat{Z}^\mathsf{T} \widehat{Z}$.
Compute $\tau_P$ from $\widehat{K}$ using Appendix A.3; initialize an empty flag set.
Enumerate $\mathcal{S}_k = \{S \subseteq F : |S| = k\}$ in order $\prec$; let $S_0$ be its first subset.
**for** each $S \in \mathcal{S}_k$ **do**
    Compute $\sigma_S^2$ and advantages $a_i(S)$ under convention $b$ (Equation (2)).
    Form $w_S$ with entries $a_i(S)/k$ on $S$ and zeros elsewhere; cache $\widehat{g}_S = \widehat{Z} w_S$.
**end for**
Choose $S_V$ as the first maximizer of $\sigma_S^2$ under $\prec$; set $S_C = S_J = S_0$.
**if** $\|\widehat{g}_F\| > \tau_P$ **then**
    Choose $S_J$ by Equation (7), breaking equal values with $\prec$.
    Set $\mathcal{C}_P = \{S \in \mathcal{S}_k : \|\widehat{g}_S\| > \tau_P\}$.
    **if** $\mathcal{C}_P \neq \varnothing$ **then**
        Choose $S_C$ by Equation (6), breaking equal values with $\prec$.
    **else**
        Flag the empty cosine candidate set; retain $S_C = S_0$.
    **end if**
**else**
    Flag the degenerate full proxy; retain $S_C = S_J = S_0$.
**end if**
Freeze the IDs of $S_V, S_C, S_J$; construct $K = Z^\mathsf{T} Z$ and the exact reference norm.
Record exact-reference eligibility using the threshold from $K$ (Appendix A.3).
**for** each evaluation convention $b' \in \{\mathrm{rec}, \mathrm{fix}\}$ **do**
    Recompute weights for the same selected subsets under $b'$ without reselection.
    Evaluate absolute errors through $K$; compute $R, r, B, A, R^*$ on eligible references.
    Record defined angles and zero/small-gradient flags using Appendix A.3.
**end for**
Under $b$, evaluate the complete uniform distribution and the exact-geometry controls.
**return** Selected IDs, paired metrics, uniform and control records, and all flags.

### A.5 Computational requirements

Building $K$ costs $\mathcal{O}(dn^2)$; CountSketch extraction costs $\mathcal{O}(dn)$ for each seed, followed by $\mathcal{O}(mn^2)$ for its Gram matrix. Weight construction costs $\mathcal{O}(k)$ per subset. After full-reference precomputation, the leading subset cost is $\mathcal{O}(k^2)$, apart from direct recomputation of numerically unstable linear combinations. These are analytic costs; all $n$ response gradients are required even though only $k$ responses are retained.

# B EXPERIMENTAL DETAILS

This appendix details the frozen-policy audit in Section 5.1. The preliminary study and its data provenance are described separately in Appendix B.7.

## B.1 BENCHMARKS, SPLITS, AND VERIFICATION

**Core audit.** The core audit covers GSM8K and TACO, together with MATH for more difficult mathematical reasoning. For each dataset, we select 600 eligible training examples without replacement before generating responses: 120 development prompts and 480 held-out audit prompts. Eligibility requires a nonempty prompt and reference answer, or at least one executable test for TACO; TACO is restricted to standard-input/output problems. We order native IDs by SHA-256 of `split:0:dataset:id`, taking the first 120 and next 480, respectively. The literal dataset tag, source revision, and IDs are archived. Models share the same prompts. These held-out training examples are the audit's evaluation split; they are distinct from official task-test sets. Prompts with constant sampled rewards, incorrect answers, or truncated responses remain in the audit.

**Rewards and verification.** Mathematical rewards are binary correctness indicators obtained by normalized numeric/symbolic answer comparison. An absent or unparsable answer receives zero. TACO rewards are the fraction of available tests passed, with no formatting bonus. Code extraction, Python version, test-harness commit, per-test timeout, memory limit, and test-bundle hash are fixed and shared across methods. The TACO sandbox limit is 10 seconds and 2 GiB per test. Compilation errors, runtime errors, and timed-out tests fail; infrastructure failures are logged and rerun before scoring.

## B.2 METRICS AND REPORTING UNITS

**Gradient fidelity.** The main outcomes are $R = \|g_S - g_F\|^2/\|g_F\|^2$ and angle in degrees. The Coder/GSM8K decomposition also reports $B = (r-1)^2$, $A = 2r(1-c)$, $R^*$, and $R - R^*$. The supplementary controls report the quantities identified in Tables 4–7. All tolerances and zero-vector conventions follow Appendix A.3. "Exact" evaluation uses the uncompressed LoRA parameter space.

**Paired summaries.** Each of the eight backbone–dataset pairs is reported separately. Within a prompt, metrics are averaged over the 25 sketch/tie seed combinations; uniform selection uses its exact expectation over subsets. The supplementary inference table reports three paired mean differences with 95% intervals from 10,000 prompt-bootstrap resamples (Appendix B.5.2). Eligible prompt counts and angular-pair coverage accompany these differences. These intervals describe uncertainty for each contrast separately. Absolute errors and degeneracy, invalidity, and truncation flags are retained in the underlying records to make denominators auditable.

**Focused supplementary controls.** Normalization and proxy controls use only Qwen2.5-Coder-3B-Instruct on the same 480 held-out GSM8K prompts and 480 held-out TACO prompts as the main audit. Normalization crosses selection and evaluation in a $2 \times 2$ design at $m = 128$. The proxy control uses recomputed advantages, $m \in \{32, 128, 512\}$, and exact cosine/joint controls. Both retain $n = 16$, $k = 8$ and the five sketch and five tie seeds. There are 436 eligible GSM8K prompts and 419 eligible TACO prompts. A separate synthetic control uses nine angle–norm settings at fixed balance and reward coupling, with 400 paired replicates (Appendix B.6.3). Tables 4–7 report the results and their effective sample sizes.

## B.3 COMPARISON METHODS AND FAIRNESS

**Matched selection rules.** PODS (TMLR 2026) compares uniform, percentile, highest-reward, and max-variance down-sampling (Xu et al., 2026b); we include all four. Uniform is the exact average over size-$k$ subsets. Highest-reward retains the top $k$ rewards. Percentile selection sorts rewards and retains one-based ranks $\lfloor (j - \frac{1}{2})n/k \rfloor + 1$, $j = 1, \ldots, k$, implementing evenly spaced empirical reward quantiles. Ties in reward sorting use a seeded ordering of response IDs. PODS maximizes subset reward variance. The within-group sample criterion in D$^3$S (ICLR 2026) maximizes advantage variance (Wang et al., 2026); under our stated normalization and fixed $k$, it has the same maximizers as PODS (Appendix D.1.4). We report this shared criterion as "PODS variance".

Table 3: Selection criteria and information access.

| Rule | Fixed-size objective | Role |
|---|---|---|
| Uniform | Uniform sample of 8 from 16 | Reference distribution |
| Reward variance | Maximize reward variance within the subset | PODS criterion |
| Cosine | Maximize cosine to the full proxy gradient | Directional diagnostic |
| Joint error | Minimize squared error to the full proxy gradient | Joint diagnostic |
| Evaluation oracle | Minimize error using exact LoRA geometry | Offline lower bound only |

CPPO (NeurIPS 2025) retains the $k$ largest full-group absolute advantages (Lin et al., 2025). Its fixed-advantage result in the focused normalization control is the native-normalization component control; applying the same selected IDs with recomputed advantages is explicitly an adaptation. For all three reward-based ranked rules, equal scores use the same response-ID ordering. The cosine and joint-error selectors use the same CountSketch and exhaustive subset family defined in Section 4.3. These two criteria target directional alignment and gradient reconstruction, respectively. The full-group GRPO reward component (Shao et al., 2024) retains all $n$ responses and defines the reference; it is labeled separately from size-$k$ competitors. Exact cosine, exact joint selection, and optimal nonnegative rescaling access privileged full-reference geometry and are diagnostic controls only.

Table 3 summarizes the core objectives and their roles.

**Shared comparison conditions.** Within a backbone–prompt pair, every comparison fixes the parameter snapshot, response tokens, verifier rewards, trainable coordinates, loss token averaging, subset size, and evaluation geometry. All comparisons retain $n = 16$, $k = 8$ and exhaustive enumeration. The focused normalization control crosses selection and evaluation in a $2 \times 2$ design with selected IDs held fixed for cross-evaluation. Gradient selectors share $P$, while reward-only selectors use only rewards. The Gram matrix $K$ is inaccessible to ordinary selectors. Gradient criteria additionally require features for all responses, so computational cost varies across selectors at the same retention budget.

**Scope of component comparisons.** The original systems differ in reward shaping, rollout counts, model families, and training schedules. In particular, PODS and CPPO use correctness/format reward designs, whereas this audit uses the stated correctness or test-fraction rewards. Our common protocol evaluates their selection rules under these shared conditions. D$^3$S additionally performs cross-group and token-level selection with a schedule, and CPPO reallocates capacity to new prompts. The single-group audit isolates subset selection from these cross-group and token-level mechanisms.

Other approaches change the response pool or gradient coefficients. ARROL (ACL 2026) prunes partially generated responses with an online quality predictor (Xu et al., 2026a); it operates during generation. SALT (2026 preprint) changes gradient coefficients using mini-batch geometry (Chang et al., 2026), and GRAD-MATCH optimizes weighted subset reconstruction (Killamsetty et al., 2021). These methods address different optimization problems from selecting an unweighted subset of completed responses under a prescribed advantage rule. The matched ranking therefore covers the selection rules applicable to that setting.

### B.4 Backbones, generation, and numerical implementation

**Model coverage.** Qwen2.5-Coder-3B-Instruct is the primary backbone (Hui et al., 2024). Qwen2.5-Math-7B and Llama-3.1-8B-Instruct (Yang et al., 2024; Grattafiori et al., 2024) provide a mathematics specialist and a second model family; both occur in D$^3$S's evaluation. PODS and CPPO also use the Qwen2.5 family, with the general Qwen2.5-3B checkpoint; our primary backbone is its Coder variant. All three models run mathematics tasks; the Coder and Llama instruction models run programming tasks. Qwen2.5-Math-7B is the base checkpoint. It receives a fixed zero-shot plain-text question/answer prompt; instruction models use their released chat templates. Within each task, the requested answer format and task content are shared. Literal prompts and templates are archived. Cross-model comparisons reflect differences in specialization, parameter count, and prompting.

**Frozen trainable coordinates.** For every model, LoRA targets `q_proj`, `k_proj`, `v_proj`, and `o_proj` in every attention layer, with rank 16, alpha 32, and zero dropout. The released base weights are frozen. The default adapter is initialized once with Kaiming-uniform input factors and zero output factors (seed 0), so insertion initially preserves the released policy; its tensors are then serialized and shared across all criteria. Gradients are computed with respect to the adapter matrices, without applying updates. The audit measures reward gradients at this initial adapter parameterization, before optimization and without a KL penalty.

**Sampling and token budgets.** The audit uses $n = 16$, temperature 1.1, and top-$p$ 1.0, with response caps of 384 tokens on GSM8K, 1,024 on TACO, and 2,048 on MATH. A requested cap is reduced to the remaining model context when necessary, and the effective cap is recorded. Prompt overflow is flagged, preserving the full task text. Generation seeds derive from model, dataset, prompt ID, and sample index through a recorded SHA-256 mapping. EOS stops generation; capped responses retain their observed tokens and verifier outcomes. The focused controls reuse these same responses and caps.

**Numerics and reproducibility.** The main CountSketch has $m = 128$; the focused proxy control adds $m = 32$ and $m = 512$. Sketch and tie seeds each range from 0 to 4. We retain $\varepsilon_a = 10^{-6}$ and the fixed numerical thresholds in Appendix A.3. Score gradients use untempered log probabilities, response-token means, float32 forward/backward computation, and disabled stochastic layers. Gram reductions use float64. The first 32 development prompts per core dataset support the direct-gradient checks in Appendix A.1; these are implementation checks, with no tuning on held-out errors. Model/dataset revisions, adapter hashes, parameter order, code and verifier versions, literal prompts, response masks, seeds, rewards, and failure flags identify each audit record. The run manifest records the execution environment; analytic computational costs appear in Appendix A.5.

### B.5 Audit inputs and statistical summaries

#### B.5.1 Frozen policy and rollout inputs

The primary RLVR instantiation uses Qwen2.5-Coder-3B-Instruct with attention-projection LoRA (rank 16, alpha 32, dropout zero), 16 responses per prompt, and the original GSM8K/TACO token caps. Appendix B specifies the core backbone/dataset coverage, the 120/480 development/evaluation split, immutable response and verifier records, and the focused supplementary controls. All selectors share serialized model and adapter tensors; a different parameter snapshot defines a separate audit. The first 32 development prompts per core dataset support direct-gradient checks. The provenance and scope of the preliminary CSV are documented separately in Appendix B.7.

#### B.5.2 Prompt-level estimates and uncertainty

The independent sampling unit in the natural audit is the prompt. For each reported condition, average a selector's metric over the 25 combinations of five sketch seeds and five tie seeds within each prompt. Reward-only and exact controls repeat identically across sketch seeds, which gives them no additional statistical weight. Uniform metrics are expectations over the complete subset distribution. Form paired prompt differences after matching repetitions; their mean is the reported effect.

Relative-error comparisons use prompts with resolved exact references and finite features. Angular differences use only matched repetitions in which both selected angles are defined; a prompt contributes its mean over such pairs and is excluded from that angular contrast if no pair is valid. Table 4 reports $n_R$, $n_\theta$, and angular-pair coverage, defined as the valid angular pairs divided by $25n_R$. Error contrasts retain proxy fallbacks. The underlying audit records keep absolute errors for reference-degenerate groups and separate zero-update, small-gradient, invalidity, and truncation flags.

Draw 10,000 paired bootstrap samples of eligible prompts using seed 0 and recompute the mean for each sample. The 2.5th and 97.5th percentiles give a pointwise 95% interval. At $m = 128$ with recomputed advantages, the main contrasts are cosine minus variance in $R$, cosine minus variance in angle, and cosine minus joint selection in $R$ (Table 4). The focused controls use the paired contrasts specified in their table headers. Intervals describe uncertainty for each contrast separately.

The five sketch and tie seeds are shared across prompts, so intervals are conditional on these fixed seed sets. For the synthetic control, first average the five tie seeds within each replicate; bootstrap the 400 replicate IDs while preserving pairing across all nine configurations. The independent units remain the replicate IDs across configurations and seed repetitions. The intervals are computed from paired records.

### B.6 Controls for normalization and representation

#### B.6.1 Selection-by-evaluation comparison

For PODS variance, CPPO, cosine, and joint selection on Coder/GSM8K and Coder/TACO, construct $S^{\mathrm{rec}}$ with recomputed advantages and $S^{\mathrm{fix}}$ with full-group advantages. Evaluate the four gradients $g^{\mathrm{rec}}_{S^{\mathrm{rec}}}$, $g^{\mathrm{fix}}_{S^{\mathrm{rec}}}$, $g^{\mathrm{rec}}_{S^{\mathrm{fix}}}$, and $g^{\mathrm{fix}}_{S^{\mathrm{fix}}}$ against their common reference $g_F$. Within a selected subset, changing the evaluation convention measures the combined centering and scaling change in Equation (10). Within an evaluation convention, changing the selected subset measures the effect of the selection convention on membership. The evaluation comparison holds membership fixed; the selection comparison allows membership to change. Reward-variance and CPPO membership are unchanged across conventions; their cross-evaluated gradients can still differ. The four combinations appear in Table 5.

#### B.6.2 Objective gaps and representation distortion

For two selectors evaluated on the same eligible groups, paired gaps satisfy $\Delta R = \Delta B + \Delta A$. The reported Coder/GSM8K decomposition and the paired synthetic control use this identity to distinguish scale and directional contributions.

The focused proxy control changes only the selection geometry: CountSketch dimensions $m \in \{32, 128, 512\}$ and privileged exact cosine/joint selection. It uses Coder/GSM8K and Coder/TACO with recomputed advantages. Table 6 reports exact-space $R$, the paired cosine-minus-joint error, and the joint selector's gap to the exact joint oracle. The exact row tests objective mismatch without sketch distortion. All natural-data metrics remain in the uncompressed LoRA evaluation geometry.

#### B.6.3 Paired synthetic construction

The synthetic control generates rewards and unweighted score vectors before applying the same selection procedure. Set $d = 512$, $n = 16$, and $k = 8$. Draw orthonormal directions $u, v$ by applying Gram–Schmidt to two independent standard Gaussian vectors. Draw a random permutation of the response IDs and assign the first $M$ responses label $\ell_i = 1$; the remaining responses have label 0. Define modal means $\nu_0 = u$ and $\nu_1 = \rho(\cos\phi\, u + \sin\phi\, v)$, then generate

$$z_i = \nu_{\ell_i} + \frac{0.1\|\nu_{\ell_i}\|}{\sqrt{d}}\epsilon_i, \qquad q_i = \frac{1}{1 + \exp[-\gamma(2\ell_i - 1) - e_i]}, \tag{B.1}$$

where $\epsilon_i \sim \mathcal{N}(0, I_d)$ and $e_i \sim \mathcal{N}(0, 1)$ are independent across responses and independent of the modal directions. The norm ratio $\rho = \|\nu_1\|/\|\nu_0\|$ concerns modal means; individual realized norms also include noise. The coefficient $\gamma$ specifies the association between mode label and reward.

The focused grid uses $\phi \in \{15°, 75°, 135°\}$ and $\rho \in \{0.5, 1, 2\}$, fixing $M = 8$ and $\gamma = 1$. Each of 400 independent replicates shares $u, v$, the response permutation, and all $\epsilon_i, e_i$ across these nine settings. Changing $\phi$ rotates the second modal mean; changing $\rho$ scales that mode and its noise. Rewards are identical across the nine paired settings, and recomputed subset advantages are applied after constructing each group. Even at modal ratio $\rho = 1$, feasible subset gradients can have different norms.

Replicate IDs $0, \ldots, 399$ serve as seeds for NumPy's PCG64 generator. Each seed draws, in order, the two Gaussian direction vectors, the $n \times d$ Gaussian noise array, the $n$ reward-noise values, and the response permutation. Both selectors optimize in exact synthetic geometry, removing sketch distortion from this intervention. The five tie seeds are averaged within each replicate; no sketch seeds are required. Table 7 reports cosine-minus-joint gaps in $R$, $B$, and $A$, together with each selector's rescaled residual. Replicate ID is the pairing unit. These controls probe the specified generative testbed.

### B.7 PRELIMINARY-STUDY PROTOCOL

#### B.7.1 SAMPLING AND DIAGNOSTIC QUANTITIES

The pilot in Section 3 uses 64 frozen development groups (32 per dataset), with $n = 16$ responses per group. For each prompt, 2,048 distinct subsets are sampled uniformly from the $\binom{16}{8} = 12{,}870$ size-$k = 8$ subsets, yielding 131,072 metric records. Uniform sampling precedes the cosine-based descriptive analysis, and the analysis uses the original, unscaled gradients.

For nonzero subset and full-group gradients $g_S$ and $g_F$, with detached reward weights, define

$$c(S) = \frac{g_S^\top g_F}{\|g_S\|\|g_F\|}, \qquad r(S) = \frac{\|g_S\|}{\|g_F\|}, \qquad R(S) = \frac{\|g_S - g_F\|^2}{\|g_F\|^2}. \tag{B.2}$$

The angular deviation is $\theta(S) = (180/\pi)\arccos c(S)$ in degrees. The error is *squared* relative error and satisfies $R = (r-1)^2 + 2r(1-c)$. Even near-collinear gradients can have a large scale term $(r-1)^2$.

The group descriptors are computed on the original full group. With response score gradient $z_i$, full-group advantage $a_i(F)$, and contribution $u_i = a_i(F)z_i$, write $g_F = n^{-1}\sum_i u_i$. The descriptors are

$$\nu_x = \mathrm{Std}_{i\in F}\big[\log(\|u_i\| + \epsilon)\big], \qquad \chi_x = 1 - \frac{\|g_F\|}{n^{-1}\sum_{i=1}^n \|u_i\| + \epsilon}, \tag{B.3}$$

where $\epsilon > 0$ stabilizes the norm summaries. Larger $\nu_x$ denotes more heterogeneous contribution norms; larger $\chi_x$ denotes greater cancellation relative to their average magnitude. The analysis uses the recorded descriptor values, with data coverage specified in Appendix B.7.4.

#### B.7.2 AGGREGATION OF THE HIGHEST-COSINE DECILE

Within each (dataset, prompt ID), sort records by decreasing cosine, breaking ties by increasing subset ID. The highest-cosine set $\mathcal{H}_x$ contains $\lceil 0.1 \times 2{,}048 \rceil = 205$ records, and

$$M_x = \mathrm{median}_{S\in\mathcal{H}_x} R(S). \tag{B.4}$$

Pooled summaries use all $64 \times 205 = 13{,}120$ retained records, whereas geometry-stratified summaries take the median of $M_x$ across prompts. Each prompt contributes the same number of records to the pool. The 13,120 subset records share responses within prompts, making the prompt the independent unit.

#### B.7.3 FIGURE CONSTRUCTION

Panel (a) bins all 131,072 pairs $(\theta, \log_2 r)$ on a hexagonal grid with `gridsize=(48,34)`. Cells containing at least eight records are colored by median $R$, with colors saturating at $0.55$; sparse-cell omission affects only the display; numerical summaries use all records. The dashed line $\log_2 r = 0$ denotes equal subset and reference norms.

Panel (b) places all 64 prompts at $(\nu_x, \chi_x)$, colors them by $M_x$, and uses circles for GSM8K and diamonds for TACO, with fixed marker size. Its background is piecewise linear interpolation on a Delaunay triangulation; thin boundary triangles are masked using `min_circle_ratio=0.01`, with no extrapolation beyond the remaining triangles. Coincident coordinates, if present, are combined by median $M_x$ for interpolation only. The color range ends at $\max(0.22, Q_{0.95}(M)) = 0.22$, divided into 11 bands; larger values saturate. Interpolated colors provide a visual guide to the observed values. Each panel uses its own color range.

#### B.7.4 DATA COVERAGE AND REPRODUCIBILITY

The pilot artifact contains derived subset metrics, group descriptors, and analysis code that reproduce the figure and numerical summaries. The script checks unique record IDs, finite metrics, prompt-metadata consistency, and the angle/error identities in Equation (B.2). All exported records pass these checks and have positive norm ratios, confirming the internal consistency of the metric records.

The artifact supports reproduction from recorded metrics; raw gradients, rewards, subset memberships, and upstream configuration metadata are not included. The main audit's configuration is documented separately in Appendix B.5.1.

### B.8 DIRECTION–SCALE DISTRIBUTION PROTOCOL

#### B.8.1 MEASUREMENTS AND AGGREGATION

The distributional analysis uses Qwen2.5-Coder-3B-Instruct/GSM8K $16 \to 8$ measurements. Figure 4 aggregates 239,800 background measurements, comprising 550 records for each of 436 prompt identifiers, and 436 measurements for each of four selectors: PODS variance, CPPO, cosine, and joint error. The background summarizes the empirical distribution of this sample.

Table 9 reports arithmetic means over the 436 records for each selector, with $R$ computed as $A + B$. The summaries are calculated before the logarithmic display transform. In particular, a mean marker is located at $(\log_{10}(\bar{A} + \epsilon), \log_{10}(\bar{B} + \epsilon))$, where $\epsilon = 10^{-4}$. The arrow in Figure 4 connects the cosine and joint-error mean markers.

#### B.8.2 DENSITY ESTIMATION AND VISUALIZATION

Both coordinates use $\log_{10}(\cdot + 10^{-4})$. Background counts are displayed on a hexagonal grid with horizontal grid size 70 and logarithmic color normalization. Thin background outlines use a $140 \times 140$ histogram smoothed with a Gaussian filter of width 1.2 bins; their levels are quantiles of positive smoothed cell values. For each selector, a Gaussian kernel density estimate in the transformed coordinates uses bandwidth factor 0.28. Density thresholds on a $240 \times 240$ grid enclose approximately 50% and 85% of fitted mass. The grid extends five kernel standard deviations beyond the observed coordinate range to include kernel tails. These contours summarize the fitted distributions.

The log transform preserves coordinate order but changes geometric distances and density. Constant-error curves are therefore drawn by transforming $A + B = R$ for $R \in \{0.03, 0.06, 0.12, 0.24\}$; the displayed log coordinates represent transformed error components. The horizontal coordinate measures the directional contribution $A = 2r(1 - c)$, which depends jointly on the angle and norm ratio; the vertical coordinate measures the scale contribution $B = (r - 1)^2$.

### B.9 GEOMETRY-STRATIFIED GAIN PROTOCOL

#### B.9.1 PROMPT-LEVEL AGGREGATION AND GEOMETRY

We analyze 87,150 paired measurements from 3,486 prompt–backbone instances across the eight backbone–dataset combinations. Each instance has all 25 combinations of five sketch seeds and five tie seeds. Geometry is defined on the full response group: with $u_{x,i} = a_i(F)z_i$ and $g_{F,x} = n^{-1}\sum_i u_{x,i}$, the descriptors in Equation (B.3) are

$$\nu_x = \mathrm{Std}_{i \in F}[\log(\|u_{x,i}\|_2 + \epsilon)], \qquad \chi_x = 1 - \frac{\|g_{F,x}\|_2}{n^{-1}\sum_i \|u_{x,i}\|_2 + \epsilon}. \tag{B.5}$$

Larger $\nu_x$ denotes greater heterogeneity in contribution norms; larger $\chi_x$ denotes stronger cancellation relative to their mean magnitude. Each descriptor is fixed across the 25 seed combinations for a given instance. The seed-averaged reconstruction gain is

$$\Delta_x = \frac{1}{25}\sum_{s=0}^{4}\sum_{t=0}^{4}\left[R_x(S_C^{(s,t)}) - R_x(S_J^{(s,t)})\right]. \tag{B.6}$$

Positive $\Delta_x$ favors joint matching. Each prompt–backbone instance receives equal weight after seed averaging; a prompt evaluated under multiple backbones contributes one instance per backbone.

To compare geometry across backbone–dataset pairs, let $\bar{\Delta}_{b,d}$ denote the mean within pair $(b, d)$ and $\bar{\Delta}$ the mean over all instances. We use

$$\widetilde{\Delta}_x = \Delta_x - \bar{\Delta}_{b(x),d(x)} + \bar{\Delta}. \tag{B.7}$$

This transformation preserves the pooled mean and equalizes pair means. It removes additive pair offsets while retaining within-pair variation and geometry-dependent pair effects.

### B.9.2 LANDSCAPE CONSTRUCTION

Figure 5 aggregates the instances onto a $72 \times 60$ grid in $(\nu, \chi)$ coordinates. The count and adjusted-gain sum in each cell are smoothed separately with a Gaussian kernel of standard deviation 1.7 bins, using zero padding. Their ratio gives the local mean adjusted gain. Cells outside the convex hull of the observed coordinates or below the 18th percentile of positive smoothed counts are masked. All instances are plotted as points; the support mask applies only to the smoothed background.

The color range spans the 3rd–97th percentiles of $\widetilde{\Delta}_x$, with values beyond this range shown in the end colors. White contours use the 62nd, 78th, 90th, and 96.5th percentiles of supported smoothed cell counts. They indicate relative density. Circled locations use the coordinate-wise medians of the joint-low and joint-high groups; their labels report group median adjusted gains. The connecting arrow indicates the group comparison. Smoothing provides a visual summary of the observations.

## B.10 SPECTRAL GEOMETRY AND PREDICTION PROTOCOL

### B.10.1 CONTRIBUTION GEOMETRY AND SELECTION PROTOCOL

Let $u_i = a_i(F) z_i$ be the full-group contribution of response $i$, $U = [u_1, \ldots, u_n]$, and $D_a = \mathrm{diag}(a_1(F), \ldots, a_n(F))$. The contribution Gram matrix is

$$G_u = U^\mathsf{T} U = D_a K D_a, \qquad K = Z^\mathsf{T} Z. \tag{B.8}$$

Here $z_i$ is the uncompressed LoRA-space score gradient and $a_i(F)$ is the full-group advantage. For nonzero $G_u$ with eigenvalues $\lambda_j \geq 0$, define

$$p_j = \frac{\lambda_j}{\sum_{\ell=1}^{n} \lambda_\ell}, \qquad r_{\mathrm{eff}} = \exp\left( - \sum_{j : p_j > 0} p_j \log p_j \right). \tag{B.9}$$

Small $r_{\mathrm{eff}} \in [1, n]$, with $n = 16$, indicates concentration in a few modes. Define $M = \sum_{i<j} |(G_u)_{ij}|$ and

$$\zeta = \begin{cases} M^{-1} \sum_{i<j} [-(G_u)_{ij}]_+, & M > 0, \\ 0, & M = 0, \end{cases} \qquad [v]_+ = \max(v, 0). \tag{B.10}$$

The ratio $\zeta \in [0, 1]$ measures the negative share of pairwise interaction magnitude. Both descriptors are computed before subset selection. We symmetrize $G_u$, reject negative eigenvalues beyond relative tolerance $10^{-12}$, and clip remaining negative values to zero. The audit's nonzero-reference condition excludes the zero matrix, whose entropy rank is undefined.

Both selectors enumerate $\binom{16}{8} = 12{,}870$ subsets in exact LoRA geometry with recomputed subset advantages. Cosine selection uses the admissibility rule in Appendix A.3; joint-error selection minimizes $R$ over all size-eight subsets. For selections $S_C^{\mathrm{ex},t}$ and $S_J^{\mathrm{ex},t}$ under tie seed $t$, we report

$$\Delta R_{\mathrm{obj},x} = \frac{1}{5} \sum_{t=0}^{4} \left[ R_x(S_C^{\mathrm{ex},t}) - R_x(S_J^{\mathrm{ex},t}) \right] \geq 0. \tag{B.11}$$

Tie resolution follows Appendix A.4. Each cosine-selected subset, including the fallback, is a candidate for the joint minimum, ensuring nonnegativity.

### B.10.2 GROUPED PREDICTIVE EVALUATION

All predictive models use the same five `GroupKFold` splits by dataset–prompt ID, keeping each prompt's backbone instances together. Continuous features are expanded to degree two, including cross terms, and standardized within training folds. Every model includes one-hot backbone–dataset indicators and an unpenalized intercept. Ridge uses $\alpha = 1$ on all other coefficients, including pair indicators, with no hyperparameter search. Table 12 reports $R^2 = 1 - \sum_x (y_x - \hat{y}_x)^2 / \sum_x (y_x - \bar{y})^2$ over all held-out predictions, where $y_x = \Delta R_{\mathrm{obj},x}$ and $\bar{y}$ is the mean across all instances.

### B.10.3 LANDSCAPE CONSTRUCTION

Figure 6 uses a $29 \times 26$ uniform grid on $[1, 16] \times [0, 1]$. Bin medians with at least six instances are linearly interpolated from bin centers onto a $320 \times 260$ display grid. We mask bins below six instances and locations outside interpolation support or the observed convex hull; all records remain visible as points. This surface summarizes local medians, separately from ridge prediction.

Colors saturate at the 2nd and 98th gap percentiles. White contours mark the 20th, 40th, 60th, 75th, 88th, and 95th gap percentiles, rounded to three decimals and retained when supported by the surface. Dashed lines mark marginal quartiles. Callouts report regime means and counts, pointing to records nearest the coordinate-wise medians in Euclidean distance, with each coordinate scaled by its within-regime range.

# C Additional Experimental Results

Tables 1 and 2 contain the reported matched comparison and the Coder/GSM8K decomposition. Means of $R$, $B$, and $A$ use the same resolved-reference groups; angle means use the defined-angle subset. The paired intervals below quantify uncertainty in their mean differences.

The first four tables below report paired uncertainty, normalization effects, proxy distortion, and a controlled angle–norm intervention. Appendices B.5.2 and B specify their sampling units and denominators. All intervals are pointwise and conditional on the fixed seed sets.

## C.1 Paired uncertainty in the main comparisons

Let $V$, $C$, and $J$ denote PODS variance, cosine, and joint selection. Write $\Delta L_{C-V} = L(S_C) - L(S_V)$ and $\Delta L_{C-J} = L(S_C) - L(S_J)$ for metric $L$. Each effect is computed within prompts before aggregation.

Across the eight comparisons, the pointwise intervals for cosine minus variance are below zero in both error and angle; those for cosine minus joint error are above zero. Eligible error counts range from 401 to 468 prompts, and angular-pair coverage is 99.0–99.9%. These paired estimates support the reported mean rankings under the stated sampling protocol.

Table 4: **Paired main-comparison estimates.** Main $16 \to 8$ setting, recomputed advantages, and $m = 128$. Effect cells contain the mean paired difference and its pointwise 95% bootstrap interval.

| Backbone | Dataset | $\Delta R_{C-V}$ | $\Delta\theta_{C-V}$ | $\Delta R_{C-J}$ | $n_R$ | $n_\theta$ | Pairs (%) |
|---|---|---|---|---|---|---|---|
| Coder-3B | GSM8K | $-0.112$ $[-0.124, -0.100]$ | $-13.4$ $[-14.2, -12.7]$ | $0.065$ $[0.057, 0.073]$ | 436 | 436 | 99.8 |
| | MATH | $-0.131$ $[-0.145, -0.118]$ | $-15.8$ $[-16.8, -14.9]$ | $0.075$ $[0.066, 0.084]$ | 452 | 451 | 99.6 |
| | TACO | $-0.145$ $[-0.161, -0.129]$ | $-18.3$ $[-19.4, -17.2]$ | $0.084$ $[0.073, 0.095]$ | 419 | 418 | 99.4 |
| Math-7B | GSM8K | $-0.105$ $[-0.116, -0.094]$ | $-12.3$ $[-13.0, -11.6]$ | $0.057$ $[0.050, 0.064]$ | 421 | 421 | 99.9 |
| | MATH | $-0.110$ $[-0.122, -0.098]$ | $-13.9$ $[-14.7, -13.1]$ | $0.070$ $[0.062, 0.078]$ | 468 | 468 | 99.8 |
| Llama-8B | GSM8K | $-0.118$ $[-0.131, -0.105]$ | $-14.6$ $[-15.5, -13.7]$ | $0.080$ $[0.071, 0.089]$ | 446 | 445 | 99.6 |
| | MATH | $-0.133$ $[-0.148, -0.119]$ | $-17.2$ $[-18.2, -16.2]$ | $0.085$ $[0.075, 0.095]$ | 443 | 442 | 99.5 |
| | TACO | $-0.155$ $[-0.173, -0.138]$ | $-20.6$ $[-21.9, -19.3]$ | $0.095$ $[0.082, 0.108]$ | 401 | 399 | 99.0 |

Backbone abbreviations follow Table 1. Angles are in degrees. $n_R$ counts eligible prompts for the error contrasts; $n_\theta$ counts prompts with at least one valid $C$–$V$ angular pair. Pair coverage is $100N_{\text{valid angular pairs}}/(25n_R)$. Negative $C - V$ effects favor cosine; a positive $C - J$ error favors joint selection. Intervals are pointwise.

### C.2 SELECTION AND EVALUATION NORMALIZATION

The primary Coder backbone supplies one mathematics and one programming audit. Each row changes only the selection/evaluation conventions; responses, $K$, $m = 128$, and seed sets stay fixed. Columns include the native CPPO normalization control.

Joint selection has lower mean error than cosine, variance, and CPPO in all four convention pairs on both tasks. The cosine–joint gap varies from 0.041 to 0.065 on GSM8K and from 0.051 to 0.084 on TACO; every pointwise interval is above zero. Normalization changes the magnitude of the discrepancy without removing it in these comparisons.

Table 5: **Advantage-normalization control.** Mean exact-space errors for PODS variance ($V$), CPPO, cosine ($C$), and joint selection ($J$), with a paired cosine-minus-joint interval.

| Select | Evaluate | $\bar{R}_V$ | $\bar{R}_{\mathrm{CPPO}}$ | $\bar{R}_C$ | $\bar{R}_J$ | $\Delta R_{C-J}$ mean [95% CI] | $n_R$ |
|---|---|---|---|---|---|---|---|
| *Qwen2.5-Coder-3B-Instruct / GSM8K* | | | | | | | |
| Rec | Rec | 0.238 | 0.367 | 0.126 | 0.061 | $0.065$ $[0.057,\ 0.073]$ | 436 |
| Rec | Fix | 0.221 | 0.286 | 0.112 | 0.071 | $0.041$ $[0.033,\ 0.049]$ | 436 |
| Fix | Rec | 0.238 | 0.367 | 0.143 | 0.080 | $0.063$ $[0.053,\ 0.073]$ | 436 |
| Fix | Fix | 0.221 | 0.286 | 0.101 | 0.052 | $0.049$ $[0.041,\ 0.057]$ | 436 |
| *Qwen2.5-Coder-3B-Instruct / TACO* | | | | | | | |
| Rec | Rec | 0.341 | 0.463 | 0.196 | 0.112 | $0.084$ $[0.073,\ 0.095]$ | 419 |
| Rec | Fix | 0.326 | 0.382 | 0.176 | 0.125 | $0.051$ $[0.040,\ 0.063]$ | 419 |
| Fix | Rec | 0.341 | 0.463 | 0.219 | 0.143 | $0.076$ $[0.064,\ 0.088]$ | 419 |
| Fix | Fix | 0.326 | 0.382 | 0.161 | 0.096 | $0.065$ $[0.054,\ 0.077]$ | 419 |

Rec/Fix denote recomputed/full-group-fixed advantages. Within a selection convention, both evaluation rows use identical selected IDs. Changing the selection convention reruns cosine/joint selection. PODS and CPPO IDs are invariant to this switch. CPPO Fix/Fix preserves its native full-group advantages; Rec/Rec is the adapted setting in Table 1. All means in a row use the same eligible prompts.

### C.3 PROXY DISTORTION AND EXACT-GEOMETRY CONTROLS

Only the representation used to select $C$ and $J$ changes. Evaluation always uses the same uncompressed LoRA Gram matrix and recomputed advantages. Let $J_{\text{ex}}$ be the exact joint minimizer and define $\Gamma_J = R(S_J) - R(S_{J_{\text{ex}}})$.

Increasing $m$ from 32 to 512 reduces the joint oracle gap from 0.042 to 0.007 on GSM8K and from 0.069 to 0.008 on TACO. With exact selection geometry, cosine still exceeds joint error by 0.078 and 0.105, respectively, with positive pointwise intervals. The remaining gap under exact geometry measures objective mismatch above the joint-error lower bound. At the primary $m = 128$ on Coder/GSM8K, the joint selector's $R = 0.061$ exceeds the exact oracle's $R = 0.041$ by $0.020$.

Table 6: **Proxy versus exact selection.** The exact row removes sketch distortion. $\bar{\Gamma}_J$ measures excess error above the exact joint oracle; it is zero by definition for exact joint selection.

| Selection geometry | $\bar{R}_C$ | $\bar{R}_J$ | $\Delta R_{C-J}$ mean [95% CI] | $\bar{\Gamma}_J$ | $n_R$ |
|---|---|---|---|---|---|
| *Qwen2.5-Coder-3B-Instruct / GSM8K* | | | | | |
| CountSketch, $m = 32$ | 0.151 | 0.083 | 0.068 [0.058, 0.078] | 0.042 | 436 |
| CountSketch, $m = 128$ | 0.126 | 0.061 | 0.065 [0.057, 0.073] | 0.020 | 436 |
| CountSketch, $m = 512$ | 0.121 | 0.048 | 0.073 [0.065, 0.081] | 0.007 | 436 |
| Exact (privileged) | 0.119 | 0.041 | 0.078 [0.069, 0.087] | 0.000 | 436 |
| *Qwen2.5-Coder-3B-Instruct / TACO* | | | | | |
| CountSketch, $m = 32$ | 0.228 | 0.148 | 0.080 [0.066, 0.094] | 0.069 | 419 |
| CountSketch, $m = 128$ | 0.196 | 0.112 | 0.084 [0.073, 0.095] | 0.033 | 419 |
| CountSketch, $m = 512$ | 0.187 | 0.087 | 0.100 [0.087, 0.113] | 0.008 | 419 |
| Exact (privileged) | 0.184 | 0.079 | 0.105 [0.091, 0.120] | 0.000 | 419 |

All geometries use the same eligible prompts, retaining proxy fallbacks. Proxy rows average five sketch seeds and five tie seeds within prompts. Exact controls average the five tie seeds; repeating them across sketch seeds changes no estimate. All rows evaluate against the same full-group reference.

### C.4 PAIRED ANGLE–NORM INTERVENTION

The synthetic protocol fixes $M = 8$, $\gamma = 1$, $d = 512$, and $16 \to 8$ retention. The nine settings share 400 replicate IDs and use exact cosine/joint selection with recomputed advantages. Rewards and underlying random draws are paired across the grid.

The mean cosine–joint error gap is 0.0932–0.1130, with all nine pointwise intervals above zero. In every setting, $\Delta\bar{B} > 0$ and $\Delta\bar{A} < 0$: cosine's smaller directional contribution is outweighed by its larger scale error. Cosine has the smaller residual after optimal rescaling throughout. The ranking is consistent across the grid, supporting the scale-based explanation within these paired configurations. Rescaled residuals measure fidelity after correcting update magnitude using the full reference, a different evaluation target from the original update.

Table 7: **Controlled angle–norm intervention.** $\phi$ is the modal angle and $\rho$ the modal norm ratio. Differences are cosine minus joint; residual columns use optimal nonnegative rescaling.

| $\phi$ | $\rho$ | $\Delta R_{C-J}$ mean [95% CI] | $\Delta\bar{B}_{C-J}$ | $\Delta\bar{A}_{C-J}$ | $\bar{R}^*_C$ | $\bar{R}^*_J$ | $n_R$ |
|---|---|---|---|---|---|---|---|
| 15° | 0.5 | 0.0964 [0.0864, 0.1069] | 0.0987 | −0.0023 | 0.0019 | 0.0046 | 400 |
| | 1.0 | 0.0932 [0.0845, 0.1018] | 0.1068 | −0.0136 | 0.0135 | 0.0294 | 400 |
| | 2.0 | 0.0947 [0.0849, 0.1050] | 0.0971 | −0.0024 | 0.0018 | 0.0046 | 400 |
| 75° | 0.5 | 0.0983 [0.0880, 0.1090] | 0.0991 | −0.0008 | 0.0006 | 0.0015 | 400 |
| | 1.0 | 0.1124 [0.1021, 0.1228] | 0.1134 | −0.0010 | 0.0007 | 0.0019 | 400 |
| | 2.0 | 0.0965 [0.0868, 0.1072] | 0.0973 | −0.0008 | 0.0005 | 0.0014 | 400 |
| 135° | 0.5 | 0.0987 [0.0885, 0.1095] | 0.0991 | −0.0004 | 0.0003 | 0.0008 | 400 |
| | 1.0 | 0.1130 [0.1028, 0.1235] | 0.1134 | −0.0004 | 0.0003 | 0.0009 | 400 |
| | 2.0 | 0.0962 [0.0863, 0.1061] | 0.0966 | −0.0004 | 0.0003 | 0.0008 | 400 |

For every eligible replicate, $\Delta R = \Delta B + \Delta A$; means use identical records. Five tie seeds are averaged before bootstrapping replicate IDs. All 400 independent replicate IDs are eligible in every row. The 3,600 generated groups are configurations paired by replicate ID. Even with equal modal norms ($\rho = 1$), subset-gradient norms can vary.

### C.5 PRELIMINARY-STUDY NUMERICAL SUMMARIES

Appendix B.7 defines the sampled subsets, diagnostic quantities, and aggregation used in Table 8.

Table 8: Descriptive summaries of the highest-cosine decile. Error statistics use squared relative error.

| Summary | Count or denominator | Value |
|---|---|---|
| Pooled median angle | 13,120 subsets | 0.82° |
| Pooled median $R$ | 13,120 subsets | 0.054 |
| Pooled fraction with $R > 0.25$ | 2,831/13,120 | 21.6% |
| Pooled fraction with $R > 0.50$ | 1,562/13,120 | 11.9% |
| Joint lower-quartile median $M_x$ | 9 prompts | 0.011 |
| Joint upper-quartile median $M_x$ | 9 prompts | 0.150 |

Quartiles are computed across all 64 prompts, pooling datasets, using linear interpolation between adjacent ordered values. The first and third quartiles are 0.337112 and 0.778208 for $\nu$, and 0.258915 and 0.559286 for $\chi$ (rounded here; comparisons use full precision). The lower group is the intersec-

tion $\nu \le Q_1(\nu)$ and $\chi \le Q_1(\chi)$; the upper group uses both corresponding $\ge Q_3$ conditions. Each intersection contains nine prompts.

The groups overlap: $M_x$ ranges from $0.005$ to $0.120$ in the lower group and from $0.087$ to $0.316$ in the upper group. The comparison characterizes the joint descriptive association of heterogeneity and cancellation with error among uniformly sampled subsets.

### C.6 Direction–scale distribution results

Table 9 summarizes the distributional sample and aggregation protocol in Appendix B.8.

Table 9: **Direction–scale summary statistics.** Each row summarizes 436 measurements. Values are means in the original coordinates, rounded independently to six decimals; unrounded values satisfy $\bar{R} = \bar{A} + \bar{B}$.

| Selector | $\bar{A}$ | $\bar{B}$ | $\bar{R}$ |
|---|---|---|---|
| PODS variance | 0.125463 | 0.219100 | 0.344563 |
| CPPO | 0.193804 | 0.304622 | 0.498426 |
| Cosine | 0.004186 | 0.184162 | 0.188348 |
| Joint error | 0.012869 | 0.067619 | 0.080488 |

Taking joint minus cosine, the differences in marginal means are $\Delta\bar{A} = 0.008683$, $\Delta\bar{B} = -0.116543$, and $\Delta\bar{R} = -0.107860$. Thus, mean $B$ decreases by $63.3\%$ and mean $R$ by $57.3\%$; $100\bar{B}/\bar{R} = 97.8\%$ for cosine. The increase in $A$ is small relative to the reduction in $B$; relative to cosine's own $A$, the joint mean is 3.07 times as large. Percent reductions are computed from marginal means in the original coordinates.

### C.7 Geometry-stratified joint-matching gains

Appendix B.9 defines the prompt-level gains, geometry descriptors, and pair adjustment used here.

Quartiles are computed across all instances using linear interpolation. For $\nu$, $(Q_1, Q_3) = (0.426247, 0.742749)$; for $\chi$, $(Q_1, Q_3) = (0.343929, 0.593207)$. The joint-low group satisfies $\nu_x \le Q_1(\nu)$ and $\chi_x \le Q_1(\chi)$; the joint-high group satisfies $\nu_x \ge Q_3(\nu)$ and $\chi_x \ge Q_3(\chi)$. These intersections contain 530 and 496 instances, respectively. Table 10 reports medians over these fixed groups.

Table 10: **Joint-matching gains by geometric regime.** Raw and pair-adjusted gaps are computed after averaging the 25 paired seed combinations within each instance.

| Regime | Instances | Median $\Delta_x$ | Median $\widetilde{\Delta}_x$ |
|---|---|---|---|
| Joint-low | 530 | 0.0411 | 0.0480 |
| Joint-high | 496 | 0.1361 | 0.1298 |

The median raw gain is 3.31 times as large in the joint-high group as in the joint-low group (Table 10). The adjusted high-minus-low median difference is 0.0818. The 2.5th and 97.5th percentiles of 5,000 bootstrap contrasts are 0.0763 and 0.0873. This joint association is consistent with the directional criterion's insensitivity to update magnitude.

### C.8 Spectral regimes and objective mismatch

The descriptors, exact-selection protocol, and grouped predictive evaluation are specified in Appendix B.10.

The 3,486 instances span eight backbone–dataset pairs and 1,333 distinct dataset–prompt IDs. Pooled quartiles use linear interpolation. For $r_{\text{eff}}$, $(Q_1, Q_3) = (5.8230, 9.6859)$; for $\zeta$, $(Q_1, Q_3) = (0.4153, 0.7368)$. Table 11 summarizes the two extreme regimes using unadjusted gaps. Their mean-gap ratio is 2.78; the same ordering holds within all eight backbone–dataset pairs under these pooled thresholds.

Table 11: **Exact objective gaps by spectral regime.**

| Regime | Conditions | Instances | Mean $\Delta R_{\mathrm{obj}}$ |
|---|---|---|---|
| Low rank, high antagonism | $r_{\mathrm{eff}} \le Q_1(r_{\mathrm{eff}}),\ \zeta \ge Q_3(\zeta)$ | 662 | 0.1458 |
| High rank, low antagonism | $r_{\mathrm{eff}} \ge Q_3(r_{\mathrm{eff}}),\ \zeta \le Q_1(\zeta)$ | 641 | 0.0524 |

Table 12: **Grouped out-of-fold prediction of objective mismatch.** All models share the pair indicators, folds, and ridge penalty.

| Continuous descriptors | Out-of-fold $R^2$ |
|---|---|
| None (pair indicators only) | 0.112 |
| Norm heterogeneity and cancellation $(\nu, \chi)$ | 0.739 |
| Effective rank and antagonism $(r_{\mathrm{eff}}, \zeta)$ | 0.821 |

The spectral descriptors improve prediction within this model class. Rank and antagonism covary strongly, so the result concerns their joint association with mismatch.

# D PROOFS

After the basic gradient identities, this appendix proves the directional obstruction in Section 4.6 and develops the additional reward-only minimax and geometric calibration results. All results use the score and advantage definitions in Section 4.1 and condition on a realized group at a fixed policy, with detached rewards and advantages. The impossibility constructions use recomputed advantages and exact Euclidean geometry. We distinguish exact values for a specified construction, minimax values over a bounded-score class, and sharp bounds over general vector families. The complete rollout family obeys additional identities, treated separately below.

## D.1 PROPERTIES OF THE GRADIENT COMPARISON

### D.1.1 ERROR DECOMPOSITION AND BOUNDS

Fix an advantage convention and a nonzero reference $g_F$. For $g_S \neq 0$, expanding the squared difference and dividing by $\|g_F\|^2$ yields

$$R(S) = r(S)^2 + 1 - 2r(S)c(S) = (r(S) - 1)^2 + 2r(S)(1 - c(S)). \tag{D.1}$$

Both terms are nonnegative. For $\delta, \eta \geq 0$, the conditions $|r - 1| \leq \delta$ and $1 - c \leq \eta$ imply $R \leq \delta^2 + 2(1 + \delta)\eta$. Conversely, $R \leq \rho$ with $\rho \geq 0$ implies $|r - 1| \leq \sqrt{\rho}$. If $r \geq r_{\min} > 0$, it also implies $1 - c \leq \rho/(2r_{\min})$. The lower bound on $r$ is needed because a vanishing subset gradient contributes little directional error even when its direction is poorly aligned.

For two selected subsets, the ranking condition is exactly $R(S_C) > R(S_V)$ if and only if $B(S_C) - B(S_V) > A(S_V) - A(S_C)$. This compares contributions in the same decomposition, with $A$ incorporating both angle and norm ratio. With $c = 1$, the error is $(r - 1)^2$, unbounded over unconstrained vectors. With $r = 1$, it lies in $[0, 4]$. These bounds apply to general vector geometry; attainability depends on the candidate family.

For a stabilized diagnostic $R_{\varepsilon_R} = \|g_S - g_F\|^2/(\|g_F\|^2 + \varepsilon_R)$, the correct relation is $R_{\varepsilon_R} = \kappa R$, where $\kappa = \|g_F\|^2/(\|g_F\|^2 + \varepsilon_R)$. Here $\varepsilon_R$ has units of squared gradient norm; it is distinct from the reward-scale constant $\varepsilon_a$. The primary analysis uses the unstabilized $R$ on eligible references.

### D.1.2 NONNEGATIVE SCALING AND THE COSINE OPTIMUM

For $g_S \neq 0$, minimizing the rescaled error is the one-dimensional convex problem

$$f_S(\alpha) = \alpha^2 r(S)^2 - 2\alpha r(S)c(S) + 1, \qquad \alpha \geq 0. \tag{D.2}$$

The unconstrained minimizer is $c/r$. Projecting it onto the nonnegative half-line gives Equation (9). For positive cosine, the residual equals $1 - c^2$; for nonpositive cosine, the minimum equals 1. In particular, the removable error is

$$R - R^* = \begin{cases} (r - c)^2, & c > 0, \\ r^2 - 2rc, & c \leq 0. \end{cases} \tag{D.3}$$

This quantity differs in general from $B = (r - 1)^2$ because optimal scaling also changes the scale-weighted directional contribution. We therefore keep $B$, $A$, and $R - R^*$ as distinct diagnostics. When $g_S = 0$, every scalar leaves the gradient zero; choosing $\alpha^* = 0$ fixes a unique reported value and gives $R^* = 1$.

Let $\mathcal{D}$ be any common set of nonzero subset gradients with the same nonzero reference. Every maximizer of $c$ over $\mathcal{D}$ minimizes $R^*$ over $\mathcal{D}$. The converse can fail when all cosines are nonpositive, since every residual then equals 1. This ordering result requires cosine and residual error to use the same geometry and candidate set, including matched numerical exclusions of small candidates. The full exact joint-error lower bound includes every size-$k$ subset; it remains a lower bound even when a cosine control has a smaller admissible set.

### D.1.3 RECOMPUTED AND FIXED ADVANTAGES

For $i \in S$, add and subtract $\mu_F$ in the recomputed advantage:

$$a_i^{\mathrm{rec}}(S) = \frac{\sigma_F + \varepsilon_a}{\sigma_S + \varepsilon_a} a_i^{\mathrm{fix}} + \frac{\mu_F - \mu_S}{\sigma_S + \varepsilon_a}. \tag{D.4}$$

Averaging its product with $z_i$ proves Equation (10). The centering term vanishes when $\mu_S = \mu_F$ and also when $\bar{z}_S = 0$. In general it is not parallel to $g_S^{\mathrm{fix}}$. If a subset has constant rewards, its recomputed gradient is zero, whereas its fixed-advantage gradient can be nonzero. If the full group has constant rewards, every advantage under both conventions is zero.

#### D.1.4 Reward and advantage variance

Using population variance throughout gives

$$\mathrm{Var}_{i\in S}(a_i^{\mathrm{fix}}) = \frac{\sigma_S^2}{(\sigma_F+\varepsilon_a)^2}, \qquad \mathrm{Var}_{i\in S}(a_i^{\mathrm{rec}}(S)) = \frac{\sigma_S^2}{(\sigma_S+\varepsilon_a)^2}. \tag{D.5}$$

The first denominator is independent of the subset. For fixed $\varepsilon_a > 0$, the second expression is strictly increasing in $\sigma_S \geq 0$. Both objectives therefore have the same maximizing subset set as reward variance. We evaluate this baseline directly using $\sigma_S^2$, avoiding floating-point differences from a redundant transformation. In the limiting convention $\varepsilon_a = 0$, recomputed variance equals 1 for every nonconstant subset; a constant subset has an undefined normalization in that convention.

### D.2 Proof of the directional obstruction at the audit budget

Theorem 1 respects the shared-response and advantage-normalization constraints that couple the subset gradients of one rollout group. The construction below uses fourteen trainable dimensions and satisfies $\max_i \|z_i\| \leq a_\varepsilon^{-1}\sqrt{9+13t^2/16}$, where $a_\varepsilon = (1+2\varepsilon_a)^{-1}$. Thus the score bound is uniform in $0 < t \leq 1$ for every fixed positive stabilizer. Under $\varepsilon_a \leq 1/16$, Proposition 2 further proves the construction-specific equality $R(S_\star) = t^2/16$. The theorem concerns unscaled reconstruction on realized groups.

*Proof of Theorem 1.* Fix $\varepsilon = \varepsilon_a > 0$ and $0 < t \leq 1$, and write $a = (1+2\varepsilon)^{-1}$. Partition the sixteen response indices into two blocks $C$ and $D$, each of size eight and containing four reward-one and four reward-zero responses. Let $s_i = 2q_i - 1$ and $s = (s_1, \ldots, s_{16})^\mathsf{T}$. Define $s_C$ and $s_D$ by restricting $s$ to the indicated block and setting its other entries to zero. The vectors $\mathbf{1}, s_C, s_D \in \mathbb{R}^{16}$ are mutually orthogonal, with squared norms $16, 8, 8$. Hence

$$\begin{gathered} H = I_{16} - \frac{\mathbf{1}\mathbf{1}^\mathsf{T}}{16} - \frac{s_C s_C^\mathsf{T}}{8} - \frac{s_D s_D^\mathsf{T}}{8}, \\ H^\mathsf{T} = H, \qquad H^2 = H, \qquad \mathrm{rank}(H) = 13, \qquad H_{ii} = \frac{13}{16}. \end{gathered} \tag{D.6}$$

Choose $B \in \mathbb{R}^{13\times 16}$ whose rows form an orthonormal basis of the range of $H$, so that $B^\mathsf{T}B = H$ and $BB^\mathsf{T} = I_{13}$. Set $x_i = 3$ on $C$ and $x_i = -1$ on $D$. With $e_i$ denoting the $i$th coordinate vector in $\mathbb{R}^{16}$, define

$$\begin{gathered} z_i = \frac{1}{a}\begin{pmatrix} s_i x_i \\ tBe_i \end{pmatrix} \in \mathbb{R}^{14}, \qquad \sum_{i=1}^{16} z_i = 0, \\ \|z_i\|^2 = \frac{x_i^2 + t^2 H_{ii}}{a^2}, \qquad \max_i \|z_i\| = \frac{\sqrt{9+13t^2/16}}{a}. \end{gathered} \tag{D.7}$$

The zero-sum identity follows from the reward balance within each block and $B\mathbf{1} = 0$. In particular, for each fixed $\varepsilon > 0$, the score norms are bounded uniformly over $0 < t \leq 1$.

**Policy realization.** Take sixteen distinct one-token responses and define a linear-softmax policy with trainable parameter $\theta \in \mathbb{R}^{14}$ by

$$\begin{gathered} \pi_\theta(i) = \frac{\exp(\theta^\mathsf{T} z_i)}{\sum_{j=1}^{16}\exp(\theta^\mathsf{T} z_j)}, \qquad \theta_0 = 0, \\ \nabla_\theta \log \pi_\theta(i)|_{\theta=0} = z_i - \frac{1}{16}\sum_{j=1}^{16} z_j = z_i, \quad \pi_0(i) = \frac{1}{16}. \end{gathered} \tag{D.8}$$

Thus the constructed vectors are actual response score gradients. Under sixteen independent draws, the event that every response appears exactly once has probability $16!/16^{16} > 0$. Condition on this

event and index the group by response identity. Then $\mu_F = \sigma_F = 1/2$ and the full-group advantages are $as_i$. Writing $e_1 \in \mathbb{R}^{14}$ for the first coordinate vector gives

$$g_F = \frac{1}{16}\sum_{i=1}^{16} as_i z_i = \begin{pmatrix} (8\cdot 3 - 8)/16 \\ tBs/16 \end{pmatrix} = e_1, \qquad g_C = 3e_1, \qquad g_D = -e_1, \qquad R(C) = R(D) = 4. \tag{D.9}$$

Both blocks use recomputed advantages identical to the full-group advantages, since their reward distributions are balanced.

**Uniqueness and strictly positive oracle error.** For a size-eight subset $S$, let $w_S$ be its coefficient vector, with $(w_S)_i = a_i^{\mathrm{rec}}(S)/8$ on $S$ and zero elsewhere, and set $b = w_S/a$. A constant-reward subset has $b = 0$, $g_S = 0$, and $R(S) = 1$. For every mixed-reward subset, all eight selected coefficients are nonzero and $\mathbf{1}^\mathsf{T} b = 0$. Its gradient therefore has the representation

$$g_S = \begin{pmatrix} X \\ tBb \end{pmatrix}, \qquad X = 3s_C^\mathsf{T} b - s_D^\mathsf{T} b, \qquad Bb = 0 \iff b \in \operatorname{span}\{\mathbf{1}, s_C, s_D\} \iff b \in \operatorname{span}\{s_C, s_D\}. \tag{D.10}$$

The last equivalence uses $\mathbf{1}^\mathsf{T} b = 0$ and the orthogonality of the three defining vectors. Since $s_C$ and $s_D$ have disjoint supports, a nonzero vector $\beta s_C + \gamma s_D$ has support $C$, $D$, or all sixteen indices. Its support can equal the size-eight set $S$ only when $S = C$ or $S = D$. These two gradients have cosine one and minus one, respectively. Every other mixed-reward subset has a nonzero transverse component and hence cosine strictly less than one. Therefore $C$ is the unique exact cosine maximizer. Moreover, no subset gradient equals $g_F$: constant-reward gradients vanish, the only nonzero collinear gradients are $3g_F$ and $-g_F$, and every remaining gradient has a nonzero transverse component. The family of subsets is finite, so its minimum reconstruction error is strictly positive.

**A reconstruction witness.** Choose $S_0$ with two successes and two failures from each block. Then $b_i = s_i/8$ on $S_0$ and zero elsewhere. Its two block masses are $s_C^\mathsf{T} b = s_D^\mathsf{T} b = 1/2$, and its scalar component is $X = 1$. Consequently,

$$\|b\|^2 = \frac{1}{8}, \qquad \|Bb\|^2 = \frac{1}{8} - \frac{(1/2)^2 + (1/2)^2}{8} = \frac{1}{16}, \qquad R(S_0) = \frac{t^2}{16}, \qquad 0 < R(S_\star) \le \frac{t^2}{16}, \qquad \frac{R(C)}{R(S_\star)} \ge \frac{4}{t^2/16} = \frac{64}{t^2}. \tag{D.11}$$

Taking $t \downarrow 0$ proves that no finite approximation factor can hold uniformly over these realizable groups. The argument applies to every fixed positive stabilizer $\varepsilon_a$ and keeps the reference norm equal to one and the score norms bounded independently of $t$. □

### D.2.1 An exact oracle value for the constructed family

The preceding impossibility result uses an upper bound on the oracle error. The next proposition shows that this bound is attained within the constructed family for the stabilizer used in the audit.

**Proposition 2** (Exact oracle error in the projection construction)**.** *For the score group constructed in the proof of Theorem 1, if* $0 < \varepsilon_a \le 1/16$ *and* $0 < t \le 1$*, then*

$$\min_{|S|=8} R(S) = \frac{t^2}{16}, \qquad \frac{R(C)}{R(S_\star)} = \frac{64}{t^2}. \tag{D.12}$$

*The minimizing subsets are exactly those with four successes and four failures overall and four selected indices from each of $C$ and $D$.*

*Proof.* Retain the preceding notation, write $\varepsilon = \varepsilon_a$, and keep $k = 8$ explicit in the algebra. Consider a mixed-reward subset with $h \in \{1, \dots, 7\}$ successes and $\ell = k - h$ failures. Since its reward

standard deviation is $\sqrt{h\ell}/k$, putting $d_h = \sqrt{h\ell} + k\varepsilon$ gives

$$b_i = \begin{cases} \ell/(kad_h), & i \in S,\ q_i = 1, \\ -h/(kad_h), & i \in S,\ q_i = 0, \\ 0, & i \notin S, \end{cases} \qquad \mathbf{1}^\mathsf{T} b = 0, \qquad \|b\|^2 = \frac{h\ell^2 + \ell h^2}{k^2 a^2 d_h^2} = \frac{h\ell}{ka^2 d_h^2}. \tag{D.13}$$

Define the reward-balance parameter $u = 2\sqrt{h\ell}/k$, the signed mass $M = s^\mathsf{T} b$, and the block masses $m_C = s_C^\mathsf{T} b$ and $m_D = s_D^\mathsf{T} b$. Direct substitution yields

$$\frac{1}{2} \le \frac{\sqrt{7}}{4} \le u \le 1, \qquad M = \frac{2h\ell}{kad_h} = \frac{u^2(1+2\varepsilon)}{u+2\varepsilon}, \qquad \|b\|^2 = \frac{M^2}{ku^2}, \\ M = m_C + m_D, \qquad X = 3m_C - m_D, \qquad m_C = \frac{X+M}{4}, \qquad m_D = \frac{3M-X}{4}. \tag{D.14}$$

The bound on $u$ follows from $h\ell \ge 7$. The projection identity then expresses the transverse error using only $u$, $M$, and $X$:

$$\begin{aligned} \|Bb\|^2 = b^\mathsf{T} H b &= \|b\|^2 - \frac{(\mathbf{1}^\mathsf{T} b)^2}{16} - \frac{m_C^2 + m_D^2}{k} \\ &= \frac{M^2}{ku^2} - \frac{(X+M)^2 + (3M-X)^2}{16k} \\ &= \frac{M^2}{k}\left(\frac{1}{u^2} - \frac{1}{2}\right) - \frac{(X-M)^2}{8k}. \end{aligned} \tag{D.15}$$

Let $\delta = 1 - u \in [0, 1/2]$. The normalization factor obeys

$$\begin{aligned} 0 \le u - M &= \frac{2\varepsilon u(1-u)}{u+2\varepsilon} \le 2\varepsilon\delta, \\ 0 \le 1 - M &\le (1+2\varepsilon)\delta, \\ \frac{M}{u} &= 1 - \frac{2\varepsilon\delta}{u+2\varepsilon} \ge 1 - 4\varepsilon\delta > 0, \\ \frac{M^2}{u^2} &\ge (1-4\varepsilon\delta)^2 \ge 1 - 8\varepsilon\delta. \end{aligned} \tag{D.16}$$

Using $(X-M)^2 \le 2(X-1)^2 + 2(1-M)^2$ and $M \le u$ gives the following lower bound. The balanced witness constructed above attains the final bound.

$$\begin{aligned} k\|Bb\|^2 &= \frac{M^2}{u^2} - \frac{M^2}{2} - \frac{(X-M)^2}{8} \\ &\ge 1 - 8\varepsilon\delta - \frac{u^2}{2} - \frac{(X-1)^2}{4} - \frac{(1-M)^2}{4} \\ &\ge 1 - 8\varepsilon\delta - \frac{u^2}{2} - \frac{(X-1)^2}{4} - \frac{(1+2\varepsilon)^2\delta^2}{4} \\ &= \frac{1}{2} - \frac{(X-1)^2}{4} + \delta\left[\frac{1+u}{2} - 8\varepsilon - \frac{(1+2\varepsilon)^2\delta}{4}\right] \\ &\ge \frac{1}{2} - \frac{(X-1)^2}{4} + \delta\left[\frac{3}{4} - \frac{1}{2} - \frac{81}{512}\right] \\ &= \frac{1}{2} - \frac{(X-1)^2}{4} + \frac{47}{512}\delta. \end{aligned} \tag{D.17}$$

The penultimate line uses $\varepsilon \le 1/16$, $u \ge 1/2$, and $\delta \le 1/2$, so that $8\varepsilon \le 1/2$ and $(1+2\varepsilon)^2\delta/4 \le (9/8)^2/8 = 81/512$. Since the reference has unit norm, every mixed-reward subset satisfies

$$\begin{aligned} R(S) &= (X-1)^2 + t^2\|Bb\|^2 \\ &\ge \frac{t^2}{2k} + \left(1 - \frac{t^2}{4k}\right)(X-1)^2 + \frac{47t^2}{512k}(1-u) \\ &\ge \frac{t^2}{2k} = \frac{t^2}{16}. \end{aligned} \tag{D.18}$$

Constant-reward subsets have error one, strictly greater than $t^2/16$. The witness in Equation (D.11) attains the lower bound, proving the exact minimum and ratio.

Both residual coefficients in Equation (D.18) are strictly positive. A minimizing mixed-reward subset must therefore satisfy $u = 1$ and $X = 1$. The first condition is equivalent to $h = \ell = 4$, in which case $b_i = s_i/k$ on $S$. If $j = |S \cap C|$, its scalar component is $X = 3j/k - (k-j)/k = 4j/k - 1$, so $X = 1$ is equivalent to $j = 4$. Conversely, every subset with these two balance properties has $m_C = m_D = 1/2$ and $\|Bb\|^2 = 1/16$, and thus attains the minimum. This proves the claimed characterization. □

**Exact selection and the angular margin.** The gap between the two largest cosine values can shrink as the perturbation decreases. A reward-balanced subset $S_1$ with seven indices from $C$ and one from $D$ has

$$X = \frac{5}{2}, \qquad \|Bb\|^2 = \frac{7}{256}, \qquad c(S_1) = \left(1 + \frac{7t^2}{1600}\right)^{-1/2}, \\ 0 < 1 - c(S_1) \le \frac{7t^2}{3200}. \tag{D.19}$$

For $f(x) = (1+x)^{-1/2}$, one has $f(0) = 1$, $f'(0) = -1/2$, and $f''(x) = 3/[4(1+x)^{5/2}] > 0$ on $x \ge 0$. Convexity therefore gives $f(x) \ge f(0) + f'(0)x = 1 - x/2$, proving the final inequality. Thus the largest-to-second-largest cosine gap is at most $7t^2/3200$. The theorem uses exact cosine maximization; resolving this shrinking gap requires correspondingly fine optimization tolerance and numerical precision as $t$ decreases.

### D.3 The exact information cost of reward-only selection

A reward-only selector is a randomized map from the indexed rewards to a subset. Its distribution may depend on known class parameters and response indices and is identical across all score assignments with those rewards. The scores remain unobserved. Let $\mathcal{G}^{(d)}_{m,G}$ contain all realizable $d$-dimensional score groups with $n = 4m$, $k = 2m$, rewards $q = (1^{2m}, 0^{2m})$, reference $g_F = e_1$, and $\max_i \|z_i\| \le G$. Advantages are recomputed within each retained subset, and $a_\varepsilon = (1 + 2\varepsilon_a)^{-1}$.

**Theorem 3** (Minimax risk over the bounded-score class)**.** *Fix $\varepsilon_a > 0$, an integer $m \ge 1$, $d \ge 2$, and $G \ge a_\varepsilon^{-1}$. Define $V_{m,G} = \min\{1, (a_\varepsilon^2 G^2 - 1)/(2(2m-1))\}$. Then $\inf_{\mathcal{A}} \sup_{Z \in \mathcal{G}^{(d)}_{m,G}} \mathbb{E}_{\mathcal{A}} R_Z(\mathcal{A}(q)) = V_{m,G}$, where the infimum is over reward-only selectors. The value is unchanged if the class is restricted to groups with zero joint-oracle error. When $a_\varepsilon^2 G^2 \le 4m - 1$, independently selecting $m$ responses uniformly from each reward class is minimax optimal; above that threshold, selecting a homogeneous-reward subset is minimax optimal.*

At $n = 16, k = 8$, the value is $\min\{1, (a_\varepsilon^2 G^2 - 1)/14\}$. The balanced minimax rule also maximizes reward variance, identifying the information limit of this reward-only class. Attainment is in expectation over selector randomness.

We prove the matching bounds over the entire class $\mathcal{G}^{(d)}_{m,G}$. Throughout the proof, write $a = a_\varepsilon$ and $\mathcal{S}_{2m} = \{S \subseteq \{1, \dots, 4m\} : |S| = 2m\}$. The subclass with a perfect reconstruction oracle is

$$\mathcal{G}^{(d),0}_{m,G} = \left\{ Z \in \mathcal{G}^{(d)}_{m,G} : \min_{S \in \mathcal{S}_{2m}} R_Z(S) = 0 \right\}. \tag{D.20}$$

A reward-only selector observes the indexed vector $q$ and induces a probability law $Q(S \mid q)$ on $\mathcal{S}_{2m}$. This law may use the response indices and the known problem parameters and is independent of the unobserved score assignment $Z$. Its auxiliary randomness is independent of that assignment. Consequently, the same $Q(S \mid q)$ must be used for every group in the class. All expectations below are conditional on the realized group unless an explicit distribution over constructed groups is introduced.

### D.3.1 AN EXACT VECTOR FINITE-POPULATION IDENTITY

**Lemma 4** (Mean and covariance of sampling without replacement)**.** *Let $v_1, \dots, v_N \in \mathbb{R}^d$, where $N \geq 2$, and let $I$ be uniform over the size-$h$ subsets of $\{1, \dots, N\}$, with $1 \leq h \leq N$. Define*

$$\mu = \frac{1}{N}\sum_{i=1}^{N} v_i, \qquad \Sigma = \frac{1}{N}\sum_{i=1}^{N}(v_i - \mu)(v_i - \mu)^{\mathsf{T}}, \qquad \overline{v}_I = \frac{1}{h}\sum_{i\in I} v_i. \tag{D.21}$$

*Then*

$$\begin{aligned} \mathbb{E}_I \overline{v}_I &= \mu, \\ \mathbb{E}_I\big[(\overline{v}_I - \mu)(\overline{v}_I - \mu)^{\mathsf{T}}\big] &= \frac{N-h}{h(N-1)}\Sigma, \\ \mathbb{E}_I\|\overline{v}_I - \mu\|^2 &= \frac{N-h}{h(N-1)}\frac{1}{N}\sum_{i=1}^{N}\|v_i - \mu\|^2. \end{aligned} \tag{D.22}$$

*Proof.* Let $d_i = v_i - \mu$ and $J_i = \mathbf{1}\{i \in I\}$. Uniform fixed-size sampling gives $\mathbb{E}J_i = h/N$ and, for distinct $i, j$, $\mathbb{E}J_iJ_j = h(h-1)/(N(N-1))$. Therefore

$$\mathbb{E}_I \overline{v}_I = \frac{1}{h}\sum_{i=1}^{N}\mathbb{E}J_i\, v_i = \frac{1}{h}\frac{h}{N}\sum_{i=1}^{N} v_i = \mu. \tag{D.23}$$

Since $\sum_i d_i = 0$, expanding its outer product gives

$$\sum_{i\neq j} d_i d_j^{\mathsf{T}} = \left(\sum_i d_i\right)\left(\sum_j d_j\right)^{\mathsf{T}} - \sum_i d_i d_i^{\mathsf{T}} = -\sum_i d_i d_i^{\mathsf{T}}. \tag{D.24}$$

The covariance follows without an independence approximation:

$$\begin{aligned} \mathbb{E}_I\big[(\overline{v}_I - \mu)(\overline{v}_I - \mu)^{\mathsf{T}}\big] &= \frac{1}{h^2}\left[\frac{h}{N}\sum_i d_i d_i^{\mathsf{T}} + \frac{h(h-1)}{N(N-1)}\sum_{i\neq j} d_i d_j^{\mathsf{T}}\right] \\ &= \frac{1}{h^2}\left[\frac{h}{N} - \frac{h(h-1)}{N(N-1)}\right]\sum_i d_i d_i^{\mathsf{T}} \\ &= \frac{N-h}{hN(N-1)}\sum_i d_i d_i^{\mathsf{T}} = \frac{N-h}{h(N-1)}\Sigma. \end{aligned} \tag{D.25}$$

Taking traces gives the squared-norm identity, including the endpoint $h = N$, where both sides vanish. □

We will also use this identity when a sign assignment is random and the sampled index set is fixed. The following consequence records why these two randomizations have the same moments.

**Lemma 5** (A fixed subset of a random balanced sign assignment)**.** *Let $N \geq 2$ be even, let $\omega \in \{-1, 1\}^N$ be uniform over vectors with exactly $N/2$ positive entries, and fix $I \subseteq \{1, \dots, N\}$ with $|I| = h \geq 1$. For $\overline{\omega}_I = h^{-1}\sum_{i\in I}\omega_i$,*

$$\mathbb{E}_\omega \overline{\omega}_I = 0, \qquad \mathbb{E}_\omega \overline{\omega}_I^2 = \frac{N-h}{h(N-1)}. \tag{D.26}$$

*Proof.* Let $K$ count positive signs in $I$. Direct counting, followed by cancellation of factorials, gives

$$\mathbb{P}(K = r) = \frac{\binom{h}{r}\binom{N-h}{N/2-r}}{\binom{N}{N/2}} = \frac{\binom{N/2}{r}\binom{N/2}{h-r}}{\binom{N}{h}}, \tag{D.27}$$

with binomial coefficients outside their allowable range interpreted as zero. The final expression is the number-of-positive-signs distribution obtained by sampling $h$ positions uniformly from any fixed balanced sign population. The corresponding sign averages are both $(2K - h)/h$. Applying Lemma 4 to that scalar population, whose mean is zero and population variance is one, proves both assertions. □

### D.3.2 UNIVERSAL UPPER BOUND AND A MATCHING REALIZABLE LOWER BOUND

*Proof of Theorem 3.* **A universal upper bound from reward-balanced sampling.** Fix an arbitrary $Z \in \mathcal{G}_{m,G}^{(d)}$. Define the two score means and their population central second moments by

$$\begin{aligned} \mu_+ &= \frac{1}{2m}\sum_{i:q_i=1} z_i, & \mu_- &= \frac{1}{2m}\sum_{i:q_i=0} z_i, \\ V_+ &= \frac{1}{2m}\sum_{i:q_i=1} \|z_i-\mu_+\|^2, & V_- &= \frac{1}{2m}\sum_{i:q_i=0} \|z_i-\mu_-\|^2. \end{aligned} \tag{D.28}$$

Since the full group is reward-balanced, its advantages are $a$ and $-a$, and the unit-reference constraint gives

$$g_F = \frac{a}{2}(\mu_+ - \mu_-) = e_1, \qquad \mu_+ - \mu_- = \frac{2}{a}e_1. \tag{D.29}$$

Independently choose $m$ indices uniformly from each reward class. Let $\widehat{\mu}_+, \widehat{\mu}_-$ be the retained class means. Every resulting subset remains reward-balanced, so recomputation leaves its positive and negative advantages equal to $a$ and $-a$. Therefore

$$g_S = \frac{a}{2}(\widehat{\mu}_+ - \widehat{\mu}_-), \qquad g_S - g_F = \frac{a}{2}\big[(\widehat{\mu}_+ - \mu_+) - (\widehat{\mu}_- - \mu_-)\big]. \tag{D.30}$$

The two centered sampling errors are independent and have zero mean. Lemma 4, with $N = 2m$ and $h = m$, thus gives the exact identity

$$\begin{aligned} \mathbb{E}_S R_Z(S) &= \mathbb{E}_S\|g_S - g_F\|^2 \\ &= \frac{a^2}{4}\left[\mathbb{E}\|\widehat{\mu}_+ - \mu_+\|^2 + \mathbb{E}\|\widehat{\mu}_- - \mu_-\|^2\right] \\ &= \frac{a^2(V_+ + V_-)}{4(2m-1)}. \end{aligned} \tag{D.31}$$

Expanding the two central second moments and using the parallelogram identity now yields a bound uniform over the entire class:

$$\begin{aligned} V_+ + V_- &= \frac{1}{2m}\sum_{i=1}^{4m}\|z_i\|^2 - \|\mu_+\|^2 - \|\mu_-\|^2 \\ &\le 2G^2 - \frac{1}{2}\|\mu_+ - \mu_-\|^2 - \frac{1}{2}\|\mu_+ + \mu_-\|^2 \\ &= 2G^2 - \frac{2}{a^2} - \frac{1}{2}\|\mu_+ + \mu_-\|^2 \\ &\le 2G^2 - \frac{2}{a^2}. \end{aligned} \tag{D.32}$$

Combining the preceding identities gives

$$\sup_{Z\in\mathcal{G}_{m,G}^{(d)}} \mathbb{E}_{\text{balanced}} R_Z(S) \le \frac{a^2G^2 - 1}{2(2m-1)}. \tag{D.33}$$

There is also a reward-only deterministic rule with risk exactly one on every group: retain all $2m$ reward-one responses. Their recomputed advantages are all zero because $\varepsilon_a > 0$, so the retained gradient is zero. Choosing whichever of these two rules has the smaller guaranteed risk proves

$$\begin{aligned} \inf_{\mathcal{A}} \sup_{Z\in\mathcal{G}_{m,G}^{(d)}} \mathbb{E}_{\mathcal{A}} R_Z(\mathcal{A}(q)) &\le V_{m,G}, \\ V_{m,G} &= \min\left\{1, \frac{a^2G^2 - 1}{2(2m-1)}\right\}. \end{aligned} \tag{D.34}$$

The infimum is throughout over reward-only selectors with the information restriction stated above. The same upper bound applies to the subclass $\mathcal{G}_{m,G}^{(d),0}$.

**A bounded hard family using the available variance.** Choose $B \geq 0$ such that

$$B^2 = \min\{a^2G^2 - 1,\, 2(2m-1)\}, \qquad A = \frac{B^2}{2m-1}, \qquad 0 \leq A \leq 2, \qquad \frac{A}{2} = V_{m,G}. \tag{D.35}$$

For indices in each reward class, use a sign vector with exactly $m$ positive and $m$ negative entries, and define scores in $\mathbb{R}^d$ by

$$az_i = \begin{cases} e_1 + B\xi_i e_2, & q_i = 1, \\ -e_1 + B\eta_i e_2, & q_i = 0. \end{cases} \qquad \sum_{i:q_i=1} \xi_i = 0, \quad \sum_{i:q_i=0} \eta_i = 0. \tag{D.36}$$

All coordinates after the first two vanish. Let $\mathcal{H}^{(d)}_{m,B}$ contain all these within-class sign assignments. Every assignment satisfies

$$\sum_{i=1}^{4m} z_i = 0, \qquad \mu_+ = \frac{1}{a}e_1, \qquad \mu_- = -\frac{1}{a}e_1, \qquad g_F = e_1, \qquad \|z_i\|^2 = \frac{1+B^2}{a^2} \leq G^2. \tag{D.37}$$

To verify realizability, assign the $4m$ distinct one-token outputs the fixed indexed rewards $q_i$ and define the $d$-parameter linear-softmax policy

$$\pi_\theta(i \mid x) = \frac{\exp(\theta^\mathsf{T} z_i)}{\sum_{j=1}^{4m} \exp(\theta^\mathsf{T} z_j)}, \qquad \theta_0 = 0, \qquad \nabla_\theta \log \pi_\theta(i \mid x)|_{\theta=0} = z_i - \frac{1}{4m}\sum_{j=1}^{4m} z_j = z_i. \tag{D.38}$$

All output probabilities at $\theta_0$ equal $1/(4m)$. The ordered realized group containing output $i$ in position $i$, for all $i$, has probability $(4m)^{-4m} > 0$. Thus the construction realizes conditional score-gradient groups under a linear-softmax policy.

Select the $m$ positive-sign responses from each reward class. The resulting subset contains $m$ responses of each reward, and its two retained score means share the same transverse component. It therefore satisfies

$$\begin{aligned} g_S &= \frac{a}{2}\left[\frac{e_1 + Be_2}{a} - \frac{-e_1 + Be_2}{a}\right] = e_1 = g_F, \\ 0 &\leq \min_{T \in \mathcal{S}_{2m}} R_Z(T) \leq R_Z(S) = 0. \end{aligned} \tag{D.39}$$

It follows that $\mathcal{H}^{(d)}_{m,B} \subseteq \mathcal{G}^{(d),0}_{m,G}$, so both natural classes are nonempty. If $B = 0$, then $G = a^{-1}$ and $V_{m,G} = 0$. Nonnegativity of squared error and Equation (D.34) already establish both minimax equalities in that boundary case. Henceforth assume $B > 0$; distinct sign assignments then give distinct score groups.

**Exact risk of every reward composition.** Draw $Z$ uniformly from $\mathcal{H}^{(d)}_{m,B}$ by drawing its two balanced sign vectors independently. Fix any $S \in \mathcal{S}_{2m}$ containing $j$ reward-one responses. If $j \in \{0, 2m\}$, its gradient vanishes and its risk is exactly one. Otherwise, put

$$p = \frac{j}{2m}, \qquad x = 2\sqrt{p(1-p)}, \qquad \delta = 2\varepsilon_a, \qquad \lambda = \frac{2p(1-p)}{a(\sqrt{p(1-p)} + \varepsilon_a)} = \frac{x^2(1+\delta)}{x+\delta}. \tag{D.40}$$

Let $U$ be the average of the $j$ selected positive-reward signs and $W$ the average of the $2m - j$ selected zero-reward signs. Direct substitution of the recomputed advantages gives

$$\begin{aligned} g_S &= \frac{p(1-p)}{a(\sqrt{p(1-p)} + \varepsilon_a)}\big[(e_1 + BUe_2) - (-e_1 + BWe_2)\big] \\ &= \lambda e_1 + \frac{\lambda B}{2}(U - W)e_2. \end{aligned} \tag{D.41}$$

The two sign assignments are independent. Applying Lemma 5 separately to the selected indices in each class yields

$$\begin{aligned} &\mathbb{E}U = \mathbb{E}W = 0, \qquad \mathbb{E}UW = 0, \\ &\mathbb{E}U^2 = \frac{2m-j}{j(2m-1)}, \quad \mathbb{E}W^2 = \frac{j}{(2m-j)(2m-1)}. \end{aligned} \tag{D.42}$$

Since the two displayed coordinate directions are orthogonal and $\|g_F\| = 1$, the exact assignment-averaged error is

$$\begin{aligned}\mathbb{E}_Z R_Z(S) &= (1-\lambda)^2 + \frac{\lambda^2 B^2}{4}\mathbb{E}_Z(U-W)^2 \\ &= (1-\lambda)^2 + \frac{\lambda^2 B^2}{4(2m-1)}\left(\frac{2m-j}{j} + \frac{j}{2m-j}\right) \\ &= (1-\lambda)^2 + A\lambda^2\left(\frac{1}{x^2} - \frac{1}{2}\right).\end{aligned} \tag{D.43}$$

The normalization factor obeys

$$x^2 \le \lambda = \frac{x^2(1+\delta)}{x+\delta} \le x \le 1, \qquad \frac{\lambda}{x^2} - 1 = \frac{1-x}{x+\delta} \ge 0. \tag{D.44}$$

The first inequality follows from $x+\delta \le 1+\delta$; the second follows from $x(1+\delta) \le x+\delta$. Including the full normalization loss gives the exact residual decomposition and sharp lower-bound chain

$$\begin{aligned}\mathbb{E}_Z R_Z(S) &= (1-\lambda)^2 + A\lambda - \frac{A}{2}\lambda^2 + A\lambda\left(\frac{\lambda}{x^2} - 1\right) \\ &= \frac{A}{2} + \left(1 - \frac{A}{2}\right)(1-\lambda)^2 + \frac{A\lambda(1-x)}{x+\delta} \\ &\ge \frac{A}{2} + \left(1 - \frac{A}{2}\right)(1-\lambda)^2 \\ &\ge \frac{A}{2} = V_{m,G}.\end{aligned} \tag{D.45}$$

Both discarded terms are nonnegative, including the saturation boundary $A = 2$. For homogeneous subsets, the same lower bound follows from $R_Z(S) = 1 \ge A/2$. Thus it holds for every fixed size-$2m$ subset. At balanced composition $j = m$, one has $x = \lambda = 1$, and both residual terms vanish. In the nonsaturated case $0 < A < 2$, an imbalanced mixed subset has $x < 1$ and strictly positive final residual term, while a homogeneous subset has risk $1 > A/2$. This identifies exactly where the composition-wise lower bound is attained.

**The information restriction and the minimax conclusion.** For an arbitrary randomized reward-only selector, the same law $Q(S \mid q)$ is used for every hidden assignment. Finite summation and the preceding pointwise-in-$S$ lower bound therefore imply

$$\begin{aligned}\sup_{Z \in \mathcal{G}_{m,G}^{(d),0}} \mathbb{E}_{\mathcal{A}} R_Z(\mathcal{A}(q)) &\ge \max_{Z \in \mathcal{H}_{m,B}^{(d)}} \mathbb{E}_{\mathcal{A}} R_Z(\mathcal{A}(q)) \\ &\ge \frac{1}{|\mathcal{H}_{m,B}^{(d)}|} \sum_{Z \in \mathcal{H}_{m,B}^{(d)}} \sum_{S \in \mathcal{S}_{2m}} Q(S \mid q) R_Z(S) \\ &= \sum_{S \in \mathcal{S}_{2m}} Q(S \mid q) \mathbb{E}_Z R_Z(S) \\ &\ge \sum_{S \in \mathcal{S}_{2m}} Q(S \mid q) V_{m,G} = V_{m,G}.\end{aligned} \tag{D.46}$$

Taking the infimum over selectors, using class inclusion, and invoking the universal upper bound yields

$$\begin{aligned}V_{m,G} &\le \inf_{\mathcal{A}} \sup_{Z \in \mathcal{G}_{m,G}^{(d),0}} \mathbb{E}_{\mathcal{A}} R_Z(\mathcal{A}(q)) \\ &\le \inf_{\mathcal{A}} \sup_{Z \in \mathcal{G}_{m,G}^{(d)}} \mathbb{E}_{\mathcal{A}} R_Z(\mathcal{A}(q)) \le V_{m,G}.\end{aligned} \tag{D.47}$$

All three bounds coincide. Independent uniform sampling within the two reward classes attains the nonsaturated upper bound; homogeneous selection attains the saturated value one. The proof handles the full natural class and its zero-oracle subclass simultaneously. □

**Randomization and scope.** The infimum in Theorem 3 includes deterministic selectors as special cases, so the lower bound applies to every fixed reward-only tie-breaking rule. Attainment of the nonsaturated upper bound uses independent uniform within-class sampling; a fixed seed yields a deterministic rule covered by the lower bound. The hidden-assignment distribution is a proof device for the conditional realized-group result. A zero retained gradient attains the saturated value one under the stated reconstruction objective.

### D.4 Sharp calibration under norm control

Let $\mathcal{D}$ be a finite, nonempty common family of nonzero candidate gradients with nonzero reference $g_F$. Suppose $|r(S) - 1| \le \delta < 1$ for every $S \in \mathcal{D}$. Write $C$ for an exact cosine maximizer on $\mathcal{D}$, put $\rho = \min_{S\in\mathcal{D}} R(S)$, and set $a = 1 - \delta > 0$ and $b = 1 + \delta$. Here these symbols denote the lower and upper geometric radii. Dividing every vector by $\|g_F\|$ and rotating the coordinates makes the reference $e_1$ without changing any radius ratio, cosine, or relative error. For a candidate of radius $r$ and cosine $c$, its error is

$$f(r, c) = r^2 + 1 - 2rc = (r - c)^2 + 1 - c^2, \qquad a \le r \le b, \quad -1 \le c \le 1. \tag{D.48}$$

**Theorem 6** (Sharp geometric calibration)**.** *Under the preceding assumptions, define*

$$\kappa_\delta(\rho) = \begin{cases} \sqrt{1-\rho}, & 0 \le \rho \le 1 - a^2, \\ (a^2 + 1 - \rho)/(2a), & 1 - a^2 \le \rho \le (a+1)^2, \\ -1, & (a+1)^2 \le \rho \le (b+1)^2. \end{cases} \tag{D.49}$$

*The exact cosine maximizer $C = S_C^{\mathcal{D}}$ obeys*

$$R(C) \le U_\delta(\rho) := 1 + b^2 - 2b\kappa_\delta(\rho). \tag{D.50}$$

*In particular, in the low-error regime $0 \le \rho \le 1 - a^2 = 2\delta - \delta^2$,*

$$0 \le R(C) - \rho \le \left(1 + \delta - \sqrt{1-\rho}\right)^2. \tag{D.51}$$

*For every admissible $(\delta, \rho)$, a family of at most two vectors attains the bound in Equation* (D.50) *under an admissible cosine tie resolution. When $\delta = 0$, all candidates have the reference norm, so cosine and reconstruction minimization are equivalent.*

The sharpness statement concerns general vector families. At $\rho = 0$ and $\delta > 0$, attaining the worst-case value $\delta^2$ requires a cosine tie; a unique maximizer has zero error. For $\delta > 0$ and $0 < \rho < (b+1)^2$, unique maximizers can approach the same envelope arbitrarily closely (Proposition 9). Complete rollout families admit additional guarantees: Appendix D.5 strengthens the recomputed-advantage bound through a weighted-barycenter identity, and Corollary 13 uses fixed-advantage complementarity at $n = 2k$ to give the excess-error bound $b^2 + 2 - 3b^{2/3}$, with leading term $4\delta^2/3$ as $\delta \to 0$. Sharpness is established for the general-vector envelope; the complete-family refinements provide upper bounds.

The following lemmas and proofs establish the exact geometric envelope, including its boundary cases.

**Lemma 7** (Extremal radius at a fixed cosine)**.** *For every $r \in [a, b]$ and $c \in [-1, 1]$, $f(r, c) \le f(b, c)$. Moreover, $f(b, c)$ is strictly decreasing in $c$.*

*Proof.* The symmetry $a + b = 2$ is essential to the endpoint comparison. Since $b - r \ge 0$ and $1 - c \ge 0$,

$$\begin{aligned} f(b, c) - f(r, c) &= (b - r)(b + r - 2c) \\ &\ge (b - r)(b + a - 2c) \\ &= 2(b - r)(1 - c) \ge 0. \end{aligned} \tag{D.52}$$

For $c_2 > c_1$, the difference $f(b, c_1) - f(b, c_2) = 2b(c_2 - c_1)$ is strictly positive because $b > 0$. This proves both claims. □

**Lemma 8** (Smallest feasible cosine at a prescribed error)**.** *If $f(r, c) = \rho$ with $r \in [a, b]$ and $c \in [-1, 1]$, then $0 \le \rho \le (b+1)^2$ and $c \ge \kappa_\delta(\rho)$. For every $\rho$ in this interval, equality is attained at an admissible radius.*

*Proof.* Nonnegativity follows from the squared-distance interpretation. Lemma 7, followed by $c \geq -1$, gives the upper endpoint of the possible error interval. Rearranging the error constraint therefore yields

$$2c = r + \frac{1-\rho}{r}, \qquad 0 \leq \rho = f(r,c) \leq f(b,c) \leq f(b,-1) = (b+1)^2. \tag{D.53}$$

We minimize the expression for $c$ while retaining both the radius restriction and the cosine range.

**Regime 1:** $0 \leq \rho \leq 1 - a^2$**.** Put $v = \sqrt{1-\rho}$, so $a \leq v \leq 1 \leq b$. Completing a square keeps the coupling between the two terms in Equation (D.53):

$$2c = r + \frac{v^2}{r} = 2v + \frac{(r-v)^2}{r} \geq 2v = 2\sqrt{1-\rho}. \tag{D.54}$$

The denominator is strictly positive. Equality is attained at $r = v$, $c = v$, for which $f(v,v) = 1 - v^2 = \rho$. Both $r$ and $c$ lie in their required intervals.

**Regime 2:** $1 - a^2 \leq \rho \leq (a+1)^2$**.** Here $1 - \rho \leq a^2$. Factoring the difference from the proposed minimum and using $r \geq a > 0$ gives

$$\begin{aligned} 2c - \left(a + \frac{1-\rho}{a}\right) &= (r-a)\left(1 - \frac{1-\rho}{ar}\right) \\ &\geq (r-a)\left(1 - \frac{a^2}{ar}\right) \\ &= \frac{(r-a)^2}{r} \geq 0. \end{aligned} \tag{D.55}$$

This argument remains valid if $1 - \rho < 0$. The proposed cosine at $r = a$ is feasible because

$$-1 = \frac{a^2 + 1 - (a+1)^2}{2a} \leq \frac{a^2 + 1 - \rho}{2a} \leq \frac{a^2 + 1 - (1-a^2)}{2a} = a \leq 1. \tag{D.56}$$

Thus $r = a$, $c = (a^2 + 1 - \rho)/(2a)$ attains the lower bound and has error exactly $\rho$.

**Regime 3:** $(a+1)^2 \leq \rho \leq (b+1)^2$**.** The range constraint $c \geq -1$ is attainable by choosing $r = \sqrt{\rho} - 1$:

$$a \leq \sqrt{\rho} - 1 \leq b, \qquad f(\sqrt{\rho} - 1, -1) = (\sqrt{\rho} - 1 + 1)^2 = \rho. \tag{D.57}$$

At $\rho = 1 - a^2$, both adjacent cosine formulas equal $a$ and both radii equal $a$. At $\rho = (a+1)^2$, both adjacent cosine formulas equal $-1$ and both radii again equal $a$. Hence the definitions agree at their shared endpoints. If $\delta = 0$, the first regime is the singleton $\rho = 0$ and the third is the singleton $\rho = 4$; the proof involves no division by $\delta$ and remains valid. □

*Proof of Theorem 6.* Let $J \in \mathcal{D}$ attain the minimum error $\rho$. Both optimizers exist because $\mathcal{D}$ is finite and nonempty. Exact cosine optimality gives $c(C) \geq c(J)$, and Lemma 8 gives $c(J) \geq \kappa_\delta(\rho)$. Applying the radius comparison and then cosine monotonicity yields

$$\begin{aligned} \rho \leq R(C) &= f(r(C), c(C)) \\ &\leq f(b, c(C)) \\ &\leq f(b, c(J)) \\ &\leq f(b, \kappa_\delta(\rho)) = 1 + b^2 - 2b\kappa_\delta(\rho). \end{aligned} \tag{D.58}$$

The first inequality follows from error optimality of $J$. In the first regime, subtraction of $\rho$ gives the exact simplification

$$\begin{aligned} 0 \leq R(C) - \rho &\leq 1 + b^2 - 2b\sqrt{1-\rho} - \rho \\ &= b^2 - 2b\sqrt{1-\rho} + (1-\rho) \\ &= (b - \sqrt{1-\rho})^2. \end{aligned} \tag{D.59}$$

This proves both bounds in the theorem.

For sharpness, fix any admissible $(\delta, \rho)$, let $\kappa = \kappa_\delta(\rho)$, and choose the equality radius from Lemma 8:

$$y = \begin{cases} \sqrt{1-\rho}, & 0 \le \rho \le 1-a^2, \\ a, & 1-a^2 \le \rho \le (a+1)^2, \\ \sqrt{\rho}-1, & (a+1)^2 \le \rho \le (b+1)^2, \end{cases} \tag{D.60}$$

$$v_J = y\left(\frac{\kappa}{\sqrt{1-\kappa^2}}\right), \qquad v_C = b\left(\frac{\kappa}{\sqrt{1-\kappa^2}}\right).$$

Both candidates are nonzero, have permitted radii, and share cosine $\kappa$ with the reference $e_1$. Their errors obey

$$R(v_J) = f(y, \kappa) = \rho \le f(b, \kappa) = R(v_C) = U_\delta(\rho). \tag{D.61}$$

Thus $J$ is an error minimizer, and choosing $C$ is an admissible cosine tie resolution. If the vectors coincide, the family can contain just one vector. This construction attains the bound at every admissible pair $(\delta, \rho)$ and proves that no smaller function of these two quantities bounds all candidate families in the theorem.

If $\delta = 0$, the middle formula is $\kappa_0(\rho) = 1 - \rho/2$ for $0 \le \rho \le 4$. Consequently Equation (D.58) reduces to $\rho \le R(C) \le \rho$. If $\rho = 0$ with arbitrary $\delta$, the matching vectors are $e_1$ and $be_1$, and their errors are zero and $\delta^2$. This last positive excess requires a cosine tie: when a zero-error candidate exists, it equals the reference and has cosine one, so a unique cosine maximizer must be that candidate. □

**Proposition 9** (Sharpness with a unique cosine maximizer at positive error)**.** *Fix $0 < \delta < 1$ and $0 < \rho < (b+1)^2$. Among finite vector families satisfying the hypotheses of Theorem 6 and having a unique cosine maximizer, the supremum of its error is still $U_\delta(\rho)$. The supremum is over general finite vector families.*

*Proof.* The upper bound already holds without assuming a tie. To approximate it from below, first note that $\kappa_\delta(\rho) < 1$ for $\rho > 0$, and that the excess in the extremal construction is strictly positive throughout the stated open interval. Direct substitution gives

$$U_\delta(\rho) - \rho = \begin{cases} (b - \sqrt{1-\rho})^2, & 0 \le \rho \le 1-a^2, \\ \dfrac{2\delta}{a}(\rho - \delta^2), & 1-a^2 \le \rho \le (a+1)^2, \\ (b+1)^2 - \rho, & (a+1)^2 \le \rho \le (b+1)^2. \end{cases} \tag{D.62}$$

The first expression is positive because $b > 1 \ge \sqrt{1-\rho}$. In the middle regime, $\rho \ge 1 - a^2 = 2\delta - \delta^2 > \delta^2$, where the strict inequality uses $0 < \delta < 1$. The third expression is positive below its upper endpoint.

Retain $v_J$ from Equation (D.60), and perturb only the second candidate's direction. For any

$$\begin{aligned} 0 < \tau < \min\left\{1 - \kappa_\delta(\rho), \frac{U_\delta(\rho) - \rho}{2b}\right\}, \\ v_C^{(\tau)} = b\left(\frac{\kappa_\delta(\rho) + \tau}{\sqrt{1 - (\kappa_\delta(\rho) + \tau)^2}}\right), \end{aligned} \tag{D.63}$$

the radius remains $b$ and the cosine is strictly greater than that of $J$. Thus $C$ is the unique cosine maximizer. Moreover,

$$R(v_C^{(\tau)}) = U_\delta(\rho) - 2b\tau > \rho = R(v_J), \qquad \lim_{\tau \downarrow 0} R(v_C^{(\tau)}) = U_\delta(\rho). \tag{D.64}$$

The oracle error is exactly $\rho$ throughout this perturbation, proving the supremum statement. The excluded endpoint $\rho = 0$ behaves differently for the reason established in the theorem's proof. □

**Corollary 10** (Approximate cosine optimization in the same geometry)**.** *Under the conditions of Theorem 6, suppose $\widetilde{C} \in \mathcal{D}$ satisfies $c(\widetilde{C}) \ge \max_{S \in \mathcal{D}} c(S) - \eta$ for $\eta \ge 0$. Then*

$$R(\widetilde{C}) \le 1 + b^2 - 2b \max\{-1, \kappa_\delta(\rho) - \eta\} \le U_\delta(\rho) + 2b\eta. \tag{D.65}$$

*The first bound is sharp for general finite vector families, for every admissible $\delta, \rho, \eta$.*

*Proof.* Let $J$ be an error minimizer and set $q = \max\{-1, \kappa_\delta(\rho) - \eta\}$. The cosine range, the approximation premise, and Lemma 8 imply

$$\begin{aligned} c(\widetilde{C}) &\geq \max\{-1, \max_{S \in \mathcal{D}} c(S) - \eta\} \\ &\geq \max\{-1, c(J) - \eta\} \\ &\geq q, \\ R(\widetilde{C}) &\leq f(b, c(\widetilde{C})) \leq f(b, q) \leq f(b, \kappa_\delta(\rho)) + 2b\eta. \end{aligned} \tag{D.66}$$

For equality in the first bound, retain $v_J$ from Equation (D.60) and replace $v_C$ by $b(q, \sqrt{1-q^2})^\mathsf{T}$. Since $q \leq \kappa$ and $\kappa - q \leq \eta$, this candidate is $\eta$-optimal in cosine. The oracle remains $J$ because

$$R(v_C) = f(b, q) \geq f(b, \kappa) \geq f(y, \kappa) = \rho = R(v_J). \tag{D.67}$$

The argument also covers $\eta \geq 2$: the cosine constraint then permits the saturated choice $q = -1$. The first bound is sharp; the final additive relaxation can exceed the maximal possible error. The approximation premise requires a cosine guarantee in the evaluation geometry; applying it to sketch-based selection requires control of sketch distortion. □

### D.5 Additional structure of the complete rollout family

The extremal families above establish sharpness for general vector geometry. A complete size-$k$ rollout family satisfies additional identities, and these can strengthen the bounds. Every expectation in this subsection is uniform over all $S \in \mathcal{S}_k$ and conditional on the realized rewards and score gradients. These exact identities retain every candidate, including small nonzero gradients.

**Proposition 11** (Weighted-barycenter cosine bounds with zero candidates retained)**.** *Suppose $n \geq 2$, $2 \leq k \leq n$, $\varepsilon_a > 0$, $g_F \neq 0$, and $\|g_S\| \leq b\|g_F\|$ for every $S \in \mathcal{S}_k$, where $b > 0$. Define*

$$\begin{aligned} \beta &= \frac{n(k-1)}{k(n-1)}, \qquad r(S) = \frac{\|g_S\|}{\|g_F\|}, \qquad t = \max_{S \in \mathcal{S}_k :\, g_S \neq 0} c(S), \\ w_S &= \sigma_S + \varepsilon_a, \qquad h = \beta(\sigma_F + \varepsilon_a), \qquad p = \mathbb{P}_S(\sigma_S > 0). \end{aligned} \tag{D.68}$$

*Under fixed advantages, $t \geq b^{-1}$. Under recomputed advantages, the nonzero family is nonempty, $p > 0$, every denominator below is positive, and*

$$\begin{aligned} t &\geq \frac{h}{\mathbb{E}_S[w_S r(S)]} \\ &\geq \frac{h}{b\,\mathbb{E}_S[w_S \mathbf{1}_{\{g_S \neq 0\}}]} \\ &\geq \frac{h}{b(\mathbb{E}_S \sigma_S + \varepsilon_a p)} \\ &\geq \frac{h}{b(\sqrt{p\beta}\,\sigma_F + \varepsilon_a p)} > 0. \end{aligned} \tag{D.69}$$

*The last two bounds require only reward statistics and the upper norm bound. Zero gradients remain in the uniform subset distribution and contribute zero to the radial moment in the first denominator.*

*Proof.* We first establish the exact finite-population identities. For a nonempty $U \subseteq F$, the vector covariance between reward and score has the pairwise representation

$$\begin{aligned} h_U &:= \frac{1}{|U|} \sum_{i \in U} (q_i - \mu_U) z_i \\ &= \frac{2|U| \sum_{i \in U} q_i z_i - 2(\sum_{i \in U} q_i)(\sum_{i \in U} z_i)}{2|U|^2} \\ &= \frac{1}{2|U|^2} \sum_{i,j \in U} (q_i - q_j)(z_i - z_j). \end{aligned} \tag{D.70}$$

The scalar $h$ in Equation (D.68) is distinct from these indexed vector covariances. Each off-diagonal ordered pair belongs to a uniform size-$k$ subset with probability $k(k-1)/(n(n-1))$. Diagonal summands vanish. Taking the finite expectation therefore gives

$$\begin{aligned}\mathbb{E}_S h_S &= \frac{1}{2k^2}\sum_{i\neq j}\mathbb{P}(i,j\in S)(q_i-q_j)(z_i-z_j)\\ &= \frac{k(k-1)}{2k^2n(n-1)}\sum_{i,j\in F}(q_i-q_j)(z_i-z_j)\\ &= \frac{n(k-1)}{k(n-1)}h_F = \beta h_F.\end{aligned} \tag{D.71}$$

Replacing $z_i$ by the scalar $q_i$ in this calculation proves the population-variance identity

$$\mathbb{E}_S\sigma_S^2 = \beta\sigma_F^2. \tag{D.72}$$

Under recomputed advantages, $h_U = (\sigma_U + \varepsilon_a)g_U$. Thus the exact weighted-barycenter identity is

$$\mathbb{E}_S[w_S g_S] = \beta(\sigma_F+\varepsilon_a)g_F = hg_F. \tag{D.73}$$

Let $e = g_F/\|g_F\|$. The inner product of the right side with $e$ is strictly positive. All weights $w_S$ are positive, so at least one candidate has positive inner product with $e$. Consequently the nonzero family is nonempty and $t > 0$. A subset with $\sigma_S = 0$ has all recomputed advantages equal to zero and therefore has $g_S = 0$. The existence of a nonzero candidate implies $p > 0$ as well. Every denominator in the proposition is now strictly positive: the first two contain a strictly positive contribution from a nonzero candidate, while the remaining two dominate the second up to the common factor $b$.

Taking the inner product in Equation (D.73) and normalizing by $\|g_F\|$ yields the first bound without replacing actual candidate radii by their maximum:

$$\begin{aligned}0 < h = \mathbb{E}_S\left[w_S\frac{\langle g_S, e\rangle}{\|g_F\|}\right] &\le t\,\mathbb{E}_S[w_S r(S)]\\ &\le bt\,\mathbb{E}_S[w_S\mathbf{1}_{\{g_S\neq 0\}}]\\ &\le bt\,\mathbb{E}_S[w_S\mathbf{1}_{\{\sigma_S>0\}}]\\ &= bt(\mathbb{E}_S\sigma_S + \varepsilon_a p).\end{aligned} \tag{D.74}$$

The first inequality uses the definition of $t$ on nonzero candidates; zero candidates contribute zero. The next uses $r(S) \le b$ only when $g_S \neq 0$. The final inequality uses the implication $g_S \neq 0 \Rightarrow \sigma_S > 0$, with the probability space retaining the zero-gradient subsets. To bound the remaining reward moment, Cauchy–Schwarz and Equation (D.72) give

$$\begin{aligned}\mathbb{E}_S\sigma_S &= \mathbb{E}_S[\sigma_S\mathbf{1}_{\{\sigma_S>0\}}]\\ &\le \sqrt{\mathbb{E}_S\sigma_S^2}\,\sqrt{\mathbb{E}_S[\mathbf{1}^2_{\{\sigma_S>0\}}]}\\ &= \sqrt{\beta\sigma_F^2}\,\sqrt{p} = \sqrt{p\beta}\,\sigma_F.\end{aligned} \tag{D.75}$$

Dividing the successive inequalities by their positive denominators proves Equation (D.69). Since $p \le 1$, these reward-only floors also dominate the corresponding bounds that charge $\varepsilon_a$ to every subset, including constant-reward subsets.

For fixed advantages, the inclusion probability $\mathbb{P}(i \in S) = k/n$ gives the unweighted barycenter identity. Taking its inner product with $e$ shows positivity of $t$ before applying the upper norm bound:

$$\begin{aligned}\mathbb{E}_S g_S^{\mathrm{fix}} = \frac{1}{k}\sum_{i=1}^n\mathbb{P}(i\in S)a_i^{\mathrm{fix}}z_i = \frac{1}{n}\sum_{i=1}^n a_i^{\mathrm{fix}}z_i = g_F,\\ 0 < \|g_F\| = \mathbb{E}_S\langle g_S^{\mathrm{fix}}, e\rangle \le t\,\mathbb{E}_S\|g_S^{\mathrm{fix}}\| \le bt\|g_F\|.\end{aligned} \tag{D.76}$$

Division by $b\|g_F\|$ proves $t \ge b^{-1}$. □

**Corollary 12** (A complete-family envelope for recomputed advantages)**.** *Under all conditions of Proposition 11, use recomputed advantages, set $b = 1+\delta$, and assume every nonzero subset gradient has radius at least $a = 1 - \delta$, where $0 \le \delta < 1$. Let $\mathcal{D} = \{S \in \mathcal{S}_k : g_S \neq 0\}$ and $\rho = \min_{S\in\mathcal{D}} R(S)$. Define*

$$\begin{aligned} \gamma_F &= \frac{\beta(\sigma_F + \varepsilon_a)}{\mathbb{E}_S[(\sigma_S + \varepsilon_a)r(S)]}, \\ \gamma_F^{\mathrm{rew}} &= \frac{\beta(\sigma_F + \varepsilon_a)}{b(\mathbb{E}_S\sigma_S + \varepsilon_a\mathbb{P}_S(\sigma_S > 0))}. \end{aligned} \tag{D.77}$$

*For the exact cosine maximizer $C$ on $\mathcal{D}$,*

$$\begin{aligned} R(C) &\le 1 + b^2 - 2b\max\{\kappa_\delta(\rho), \gamma_F\} \\ &\le 1 + b^2 - 2b\max\{\kappa_\delta(\rho), \gamma_F^{\mathrm{rew}}\}. \end{aligned} \tag{D.78}$$

*If the minimum over all subsets is less than one, or if no subset has zero gradient, that minimum equals $\rho$. The first bound then compares with the unrestricted joint oracle.*

*Proof.* Proposition 11 proves that $\mathcal{D}$ is nonempty, that $c(C) \ge \gamma_F > 0$, and that $\gamma_F \ge \gamma_F^{\mathrm{rew}}$. A minimizer $J \in \mathcal{D}$ also satisfies $c(J) \ge \kappa_\delta(\rho)$ by Lemma 8, so exact cosine optimality gives

$$\begin{aligned} c(C) &\ge \max\{c(J), \gamma_F\} \ge \max\{\kappa_\delta(\rho), \gamma_F\}, \\ R(C) &= f(r(C), c(C)) \le f(b, c(C)) \\ &\le f\big(b, \max\{\kappa_\delta(\rho), \gamma_F\}\big) \\ &\le f\big(b, \max\{\kappa_\delta(\rho), \gamma_F^{\mathrm{rew}}\}\big). \end{aligned} \tag{D.79}$$

The cosine floors cannot exceed one under the hypotheses, because each is bounded above by the actual optimal cosine. Hence every use of $f$ above has an admissible cosine argument. A zero gradient has error exactly one; it therefore cannot attain a full-family minimum below one. If there are no zero gradients, the two candidate families coincide. This proves both oracle comparisons. These inequalities strengthen the general-vector bound using rollout structure and reward moments. □

### D.6 Complementary subsets at a half-group budget

At the audit budget $n = 16, k = 8$, fixed advantages provide more structure than the unweighted barycenter alone: every candidate has a complementary candidate reflected through the reference. The next result uses this symmetry to improve the general full-family excess bound.

**Corollary 13** (Improved fixed-advantage bound at a half-group budget)**.** *Suppose $n = 2k$, $k \ge 2$, $g_F \neq 0$, and every size-$k$ subset uses fixed full-group advantages and satisfies $1-\delta \le r(S) \le 1+\delta$ for $0 \le \delta < 1$. Let $C$ maximize exact cosine and $J$ minimize exact relative error over all size-$k$ subsets. With $b = 1 + \delta$ and $\rho = R(J)$,*

$$\begin{aligned} 0 \le R(C) - \rho &\le b^2 + 1 - \frac{2b}{\sqrt{1+\rho}} - \rho \\ &\le b^2 + 2 - 3b^{2/3}. \end{aligned} \tag{D.80}$$

*For every $\delta > 0$, the final bound is strictly smaller than $(b - b^{-1})^2$, the bound obtained from the full-family cosine floor alone.*

*Proof.* Write $a = 1 - \delta$. The fixed advantages are identical for a subset and its complement. Since both have size $k$ and $n = 2k$, their gradients satisfy the exact reflection identity

$$g_S^{\mathrm{fix}} + g_{S^c}^{\mathrm{fix}} = \frac{1}{k}\sum_{i=1}^{2k} a_i^{\mathrm{fix}} z_i = 2g_F. \tag{D.81}$$

Let $e = g_F/\|g_F\|$ and write $g_J/\|g_F\| = e + x$. The complementary oracle candidate has normalized gradient $e - x$. Both have squared relative error $\|x\|^2 = \rho$, and both are nonzero by the lower radius bound. Their squared norms also show that

$$0 \le \rho = \|x\|^2, \qquad 1 + \rho = \frac{\|e + x\|^2 + \|e - x\|^2}{2} \le b^2, \qquad \rho \le b^2 - 1. \tag{D.82}$$

Set $v = |\langle e, x\rangle| \geq 0$, and choose from the pair the vector whose component along $e$ is $1 + v$. Its cosine is positive and obeys

$$\begin{aligned}\left(\frac{1+v}{\sqrt{1+2v+\rho}}\right)^2 - \frac{1}{1+\rho} &= \frac{(1+v)^2(1+\rho) - (1+2v+\rho)}{(1+2v+\rho)(1+\rho)} \\ &= \frac{2v\rho + v^2(1+\rho)}{(1+2v+\rho)(1+\rho)} \geq 0.\end{aligned} \tag{D.83}$$

Both sides before squaring are nonnegative. Exact cosine optimality of $C$ therefore yields

$$c(C) \geq \max\{c(J), c(J^c)\} \geq \frac{1+v}{\sqrt{1+2v+\rho}} \geq \frac{1}{\sqrt{1+\rho}}. \tag{D.84}$$

The radius interval remains the symmetric interval $[1-\delta, 1+\delta]$, so Lemma 7 applies. Using that lemma, the preceding cosine floor, and error optimality of $J$ gives the pointwise chain

$$\begin{aligned}0 \leq R(C) - \rho &= f(r(C), c(C)) - \rho \\ &\leq f(b, c(C)) - \rho \\ &\leq f\left(b, \frac{1}{\sqrt{1+\rho}}\right) - \rho \\ &= b^2 + 1 - \frac{2b}{\sqrt{1+\rho}} - \rho =: H_b(\rho).\end{aligned} \tag{D.85}$$

We maximize this one-variable bound with its positive and negative terms coupled. For $\rho \geq 0$,

$$H_b'(\rho) = \frac{b}{(1+\rho)^{3/2}} - 1, \qquad H_b''(\rho) = -\frac{3b}{2(1+\rho)^{5/2}} < 0. \tag{D.86}$$

Hence $H_b$ is strictly concave and is maximized at $\rho_* = b^{2/3} - 1$. This value lies in the feasible interval $[0, b^2 - 1]$ because $b \geq 1$. Substitution gives

$$\begin{aligned}H_b(\rho) &\leq H_b(\rho_*) \\ &= b^2 + 1 - \frac{2b}{b^{1/3}} - (b^{2/3} - 1) \\ &= b^2 + 2 - 3b^{2/3}.\end{aligned} \tag{D.87}$$

This proves Equation (D.80), including $b = 1$, when $\rho_* = 0$ and the final bound is zero.

To compare with the bound from the unweighted barycenter, Proposition 11 gives $t = c(C) \geq b^{-1}$. Since $ab = 1 - \delta^2 \leq 1$, this also gives $t \geq a$. Completing the square in $f$ then yields

$$\begin{aligned}R(J) = f(r(J), c(J)) &\geq f(r(J), t) \\ &\geq \min_{a \leq r \leq b}\big((r-t)^2 + 1 - t^2\big) = 1 - t^2, \\ R(C) - R(J) &\leq f(b, t) - (1 - t^2) = (b - t)^2 \leq (b - b^{-1})^2.\end{aligned} \tag{D.88}$$

Here $t \leq 1 \leq b$, so the constrained radial minimum is indeed attained at $r = t$. Finally, put $u = b^{2/3}$. Arithmetic–geometric mean applied to the four positive numbers $u, u, u, u^{-3}$ proves

$$\begin{aligned}(b - b^{-1})^2 - (b^2 + 2 - 3b^{2/3}) &= 3u + u^{-3} - 4 \\ &\geq 4\big(u \cdot u \cdot u \cdot u^{-3}\big)^{1/4} - 4 = 0.\end{aligned} \tag{D.89}$$

Equality in this arithmetic–geometric mean inequality requires $u = 1$, equivalently $\delta = 0$. Thus the complement bound is strictly smaller for $\delta > 0$. The reflection argument requires the complete candidate family with fixed full-group advantages. □