NASH CREDIT ROUTING FOR OPTIMIZING RESPONSES BEYOND GRPO

Anonymous authors
Paper under double-blind review

ABSTRACT

A GRPO update can improve its sampled objective while reducing the likelihood of an individual successful response. Repeated, directionally similar successes can also dominate the positive update through multiplicity alone. We introduce Nash Credit Routing (NCR), a constrained correction to GRPO that reallocates a fixed positive-credit budget across successful responses and selectively refunds penalties on failed-response segments. The correction remains near GRPO, preserves at least its first-order sampled-objective improvement, and causes no additional first-order likelihood harm to any sampled success. NCR assigns positive bargaining powers by minimizing directional collision under a multiplicity-calibrated ratio constraint, then maximizes weighted Nash welfare over normalized likelihood gains. At the KL relevance radius induced by the Nash solution, NCR maximizes worst-case additional alignment over the same feasible correction set. Under explicit assumptions, we further derive a bound on NCR's worst-case held-out accuracy change relative to GRPO. Across the evaluated models and mathematics benchmarks, NCR achieves higher macro-average pass@1 and pass@128 than the compared baselines.

1 INTRODUCTION

Reasoning models achieve stronger performance by learning from outcome rewards. However, a single correct answer does not reveal which step to be reinforced, meanwhile, an incorrect answer offers little guidance on how the trajectory should be corrected. To this end, reinforcement learning methods, notably GRPO, assign advantages by comparing sampled responses to the same prompt, then combine their token-level policy gradients (Shao et al., 2024). Within a mixed response group, a successful response receives a positive advantage, and a failed one gets a negative advantage. Nevertheless, the combined update may decrease the likelihood of a successful response while still improving the overall training objective. In particular, the gradients from failed responses could oppose direction from successful responses, consequently weakening the learning signal for effective reasoning (Deng et al., 2025; 2026; Zhang et al., 2026c).

The second challenge lies in the question of credit allocation among successful trajectories. In GRPO, successful responses with shared gradient directions collectively reinforce each other, hence increasing the influence of these directions in proportion to their sampling frequency (Lochab et al., 2026a). However, we argue that frequency is *not* necessarily evidence for usefulness on future prompts. We therefore not only investigate coefficient of gradient of positive samples but also from negative samples (Zhai et al., 2026; Zhang et al., 2026b), and they can be used to construct a more informative credit-allocation mechanism for reasoning strategies.

Our resulting framework is Nash credit routing (NCR) - a joint positive-credit and negative-refund rule, resembling GRPO in spirit but go beyond to reallocate a fixed effective positive budget among the whole batch of successful responses and reduce selected penalties on failed-response segments, achieving a dual effect where a negative contribution can make another allocation of positive credit possible. We further introduce constraints to limit the correction magnitudes and guarantee no additional first-order harm relative to the original GRPO update and the successful trajectories.

An important decision in our proposed NCR is how to distribute permitted credit among competing successful trajectories. We address this problem using weighted Nash welfare, which, in principle,

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.

discounts trajectories that already receive substantial support and allocates more credit to under-supported ones. As a result, the allocation is driven by marginal utility rather than by sample frequency alone. Since stronger support under the training objective does not necessarily imply greater benefit on unseen prompts we analyze the allocation from a worst-case generalization perspective and show that Nash welfare produces stronger guarantees on test-time improvement than both linear allocation and standard GRPO. Our contributions are threefold:

- **Joint credit routing.** We introduce NCR, which jointly reallocates a fixed positive-credit budget and selectively refunds failed-segment penalties within each prompt group. Its positive bargaining powers minimize directional collision under a multiplicity-calibrated ratio cap, while Nash welfare allocates feasible likelihood improvements.
- **Protection and robust optimality.** We prove first-order protection relative to GRPO for the sampled objective and every sampled successful trajectories. At the KL relevance radius induced by the Nash solution, NCR maximizes worst-case additional alignment over the same feasible correction set, including matched linear routing and GRPO. Under explicit assumptions, we further bound NCR's worst-case held-out accuracy change relative to GRPO.
- **Empirical evaluation.** Across the evaluated models, NCR achieves higher macro-average pass@1 and pass@128 over the mathematics benchmarks than the compared baselines.

2 BACKGROUND

This section introduces the two ingredients used by NCR. We first derive the current GRPO update and express its gradient coefficients as positive response credit and negative segment credit. We then review weighted Nash bargaining, which NCR later uses to compare feasible likelihood improvements across successful responses. We use bold lowercase letters for vectors, bold uppercase letters for matrices, and non-bold symbols for scalars unless stated otherwise.

2.1 LOCAL GRPO UPDATE

In this section, we will recall formulation of GRPO method and present its gradient update as a combination of updates from positive and negative samples. For prompt x, the rollout policy $\pi_{\boldsymbol{\theta}_{\text{old}}}$ samples $G \geq 2$ responses $y_j = (y_{j,1}, \ldots, y_{j,T_j})$ for $j = 1, \ldots, G$ with binary rewards $R_j = R(x, y_j)$. Let $y_{j,<t}$ be the prefix precedes token t in y_j. Define the advantage as $\widehat{A}_j = \big(R_j - \text{mean}(\{R_k\}_{k=1}^G)\big)/\big(\text{std}(\{R_k\}_{k=1}^G) + \varepsilon\big)$, and the token-level importance ratio as $\rho_{j,t}(\boldsymbol{\theta}) = \pi_{\boldsymbol{\theta}}(y_{j,t} \mid x, y_{j,<t})/\pi_{\boldsymbol{\theta}_{\text{old}}}(y_{j,t} \mid x, y_{j,<t})$, where std denotes the population standard deviation and $\varepsilon > 0$ prevents division by zero. The GRPO loss (Shao et al., 2024; Schulman et al., 2017) is

$$\mathcal{L}_{\text{GRPO}}(\boldsymbol{\theta}) = -\frac{1}{G}\sum_{j=1}^{G}\frac{1}{T_j}\sum_{t=1}^{T_j}\min\Big\{\rho_{j,t}(\boldsymbol{\theta})\widehat{A}_j,\ \text{clip}(\rho_{j,t}(\boldsymbol{\theta}), 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}})\,\widehat{A}_j\Big\}. \quad (1)$$

where $\epsilon_{\text{low}} < 1$ and ϵ_{high} control clipping. Following DAPO (Yu et al., 2025), we omit the reference-policy KL penalty and allow asymmetric clipping, but retain GRPO's response normalization. Let $\mathbf{u}_{j,t} = \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(y_{j,t} \mid x, y_{j,<t})$ be the token score, and let $\gamma_{j,t}$ equal $\rho_{j,t}(\boldsymbol{\theta})$ when token t is unclipped and 0 otherwise (exact form and boundary conventions in Appendix F). Grouping tokens by response, the ascent direction $\mathbf{g} := -\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\text{GRPO}}(\boldsymbol{\theta})$ is

$$\mathbf{g} = \sum_{j=1}^{G} a_j \mathbf{s}_j, \qquad a_j = \frac{\widehat{A}_j \bar{\gamma}_j}{G}, \quad \bar{\gamma}_j = \frac{1}{T_j}\sum_t \gamma_{j,t}, \quad \mathbf{s}_j = \frac{\sum_t \gamma_{j,t}\mathbf{u}_{j,t}}{\sum_t \gamma_{j,t}}. \quad (2)$$

At $\boldsymbol{\theta} = \boldsymbol{\theta}_{\text{old}}$ in a mixed group, $\gamma_{j,t} = 1$, $a_j = \widehat{A}_j/G$, and $\mathbf{s}_j = \nabla \ell_j(\boldsymbol{\theta})$. Next, we will decompose the gradient $\mathbf{g}$ with respect to positive and negative responses. Index successes by $i \leq m$ and failures by $j > m$. Because parts of a failed response can act differently on the update (Section 3.1), we keep each success whole but partition each failed response into disjoint token segments, enumerated globally by $s = 1, \ldots, S$, with j_s the response containing segment s and $\mathcal{I}_s$ its positions. For segments with $\sum_{t\in\mathcal{I}_s} \gamma_{j_s,t} > 0$,

$$\widetilde{\mathbf{n}}_s = \frac{\sum_{t\in\mathcal{I}_s}\gamma_{j_s,t}\mathbf{u}_{j_s,t}}{\sum_{t\in\mathcal{I}_s}\gamma_{j_s,t}}, \qquad \beta_s = \frac{-\widehat{A}_{j_s}}{G\,T_{j_s}}\sum_{t\in\mathcal{I}_s}\gamma_{j_s,t}, \qquad \boxed{\mathbf{g} = \sum_{i=1}^{m} a_i \mathbf{s}_i - \sum_{s=1}^{S}\beta_s\widetilde{\mathbf{n}}_s,} \quad (3)$$

Here, $\widetilde{\mathbf{n}}_s$ is the clipping-weighted average token gradient of segment s, and $-\beta_s$ is its coefficient in the GRPO update. We refer to the scalar coefficient multiplying each gradient term as its *credit*: $a_i \geq 0$ is the positive credit of successful response i, whereas $-\beta_s \leq 0$ is the negative credit of failed-response segment s. Segmentation only regroups the failed-response gradient, each token score keeping its original prefix (proof in Appendix F).

2.2 Weighted Nash welfare

Consider m players with a feasible set of positive utility vectors $\mathcal{U} \subset \mathbb{R}^m_{++}$. Let $\mathbf{U} = (U_1, \ldots, U_m)^\top \in \mathcal{U}$, and let $q_i > 0$, $\sum_i q_i = 1$, denote the bargaining power of player i.

Since the logarithm is strictly increasing, the product and log forms have the same maximizers:

$$\arg\max_{\mathbf{U}\in\mathcal{U}} \prod_{i=1}^{m} U_i^{q_i} = \arg\max_{\mathbf{U}\in\mathcal{U}} \sum_{i=1}^{m} q_i \log U_i. \tag{4}$$

Its marginal value for player i is $\partial(\sum_j q_j \log U_j)/\partial U_i = q_i/U_i$. Hence, for fixed q_i, the same absolute increase in utility is worth less when that player's utility is already larger. The objective therefore favors allocating further improvement toward players that have received less utility so far.

3 What response-level rewards do not guarantee

Using the GRPO decomposition in Section 2.1, we study two questions: whether a GRPO step protects successful responses in its group, and whether support it assigns is useful on new prompts.

3.1 A lower group loss can suppress a successful response

For successful response i, let $\mathbf{p}_i = \nabla \ell_i(\boldsymbol{\theta})$. The full GRPO update lowers the sampled loss to first order, $\frac{d}{d\eta}\mathcal{L}_{\mathrm{GRPO}}(\boldsymbol{\theta} + \eta\mathbf{g})\big|_{\eta=0} = -\|\mathbf{g}\|_2^2$, yet the full GRPO updates could decrease the influence of positive sample i because: at $\boldsymbol{\theta} = \boldsymbol{\theta}_{\mathrm{old}}$, where $\mathbf{s}_k = \mathbf{p}_k$,

$$\frac{d}{d\eta}\ell_i(\boldsymbol{\theta} + \eta\mathbf{g})\Big|_{\eta=0} = \mathbf{p}_i^\top \mathbf{g} = a_i\|\mathbf{p}_i\|_2^2 + \sum_{k\leq m}^{k\neq i} a_k \mathbf{p}_i^\top \mathbf{p}_k - \sum_s \beta_s \mathbf{p}_i^\top \widetilde{\mathbf{n}}_s \tag{5}$$

could be negative. The reason could be that the term $\mathbf{p}_i^\top \widetilde{\mathbf{n}}_s$ are positive. Such interference can arise in several ways: a failed response may compose of a useful or nearly correct reasoning path before a later mistake, reach the correct answer but fail the verifier because of formatting, or contain shared structural and problem-specific tokens that also occur in successful responses. More generally, token-level contributions inside a failed response are heterogeneous: some suppress correct-segment likelihood, whereas others provide useful negative signal (Deng et al., 2025; 2026). This observation motivates *selective* refunds rather than equally penalizing the whole negative side. Thus, a refund means only weakening the penalty on negative segments, not treating them as semantically correct.

3.2 Repeated support need not be useful support

When several successful responses in a mixed group have near-aligned gradients, their GRPO contributions repeatedly reinforce essentially the same direction, whereas a direction represented once contributes once. This difference reflects sampling multiplicity, not known held-out usefulness, which is unavailable during training. For a held-out prompt distribution Q, define expected sampled accuracy and its gradient by

$$\mathrm{Acc}_Q(\boldsymbol{\theta}) = \mathbb{E}_{z\sim Q}\mathbb{E}_{Y\sim\pi_{\boldsymbol{\theta}}(\cdot|z)} R(z, Y), \qquad \mathbf{h}_Q = \nabla_{\boldsymbol{\theta}}\,\mathrm{Acc}_Q(\boldsymbol{\theta}). \tag{6}$$

For each correct training response i with $\mathbf{p}_i \neq \mathbf{0}$, let $\mathbf{v}_i = \mathbf{p}_i/\|\mathbf{p}_i\|_2$. For an allocation strategy $A \in \{\mathrm{GRPO}, \mathrm{NCR}\}$, let $\mathbf{d}_A$ denote its local update and define its support for direction i as $S_i^{(A)} = \mathbf{v}_i^\top \mathbf{d}_A$. Let $\widehat{\mathbf{h}}_Q$ denote the independent held-out reward-weighted estimate of $\mathbf{h}_Q$, and define the usefulness of the same direction as $B_i = \mathbf{v}_i^\top \widehat{\mathbf{h}}_Q$. Figure 1 compares these quantities on the

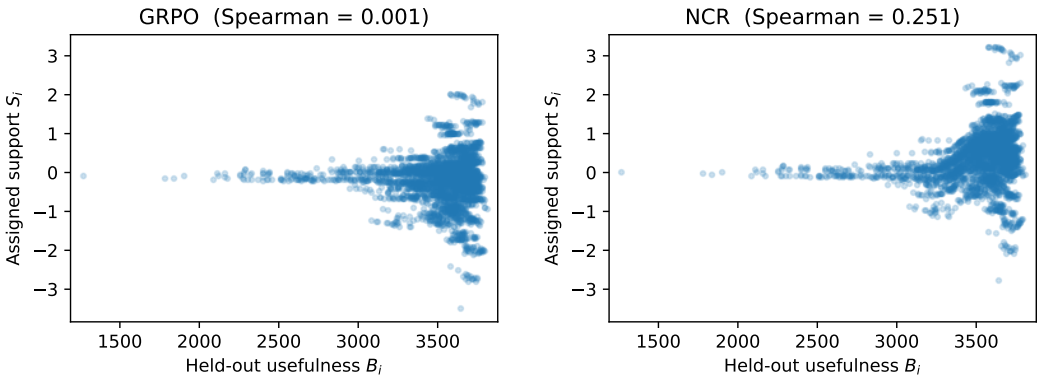

Figure 1: Support $S_i^{(A)}$ versus held-out usefulness B_i for the same 4,340 correct training responses. GRPO support is essentially uncorrelated with B_i (Spearman 0.001), whereas NCR has substantially stronger positive rank alignment (Spearman 0.251).

public Qwen3-4B checkpoint. GRPO support thus provides almost no ranking of held-out usefulness, whereas NCR substantially improves this alignment. This diagnostic neither shows that rare directions are inherently better nor measures post-training accuracy; it shows that sampled multiplicity is a poor proxy for held-out usefulness here, while NCR mitigates the mismatch.

4 Nash credit routing with collision-minimized powers

Motivated by the above observations, we propose our method Nash Credit Routing (NCR), that preserves the total positive credit; redistribute the credit among positive response and may refund part of the negative credit on selected failed segments.

4.1 Feasible credit changes

Let $\Gamma = \sum_{i \le m} a_i > 0$ be the total positive credit and let $\bar{w}_i = a_i/\Gamma$ be its GRPO's allocation, so $\mathbf{g} = \Gamma \sum_i \bar{w}_i \mathbf{s}_i - \sum_s \beta_s \widetilde{\mathbf{n}}_s$. NCR replaces $\bar{\mathbf{w}}$ by $\mathbf{w} \in \Delta_m = \{\mathbf{w} \ge \mathbf{0} : \sum_i w_i = 1\}$ and refunds a fraction $\xi_s \in [0,1]$ of each negative credit. The NCR ascent directions is: $\mathbf{d}(\mathbf{w}, \boldsymbol{\xi}) = \Gamma \sum_{i=1}^m w_i \mathbf{s}_i - \sum_{s=1}^S \beta_s (1-\xi_s) \widetilde{\mathbf{n}}_s = \mathbf{g} + \boldsymbol{\delta}(\mathbf{w}, \boldsymbol{\xi})$. Thus, NCR changes the GRPO direction by a direction vector

$$\boldsymbol{\delta}(\mathbf{w}, \boldsymbol{\xi}) = \Gamma \sum_{i=1}^m (w_i - \bar{w}_i)\mathbf{s}_i + \sum_{s=1}^S \beta_s \xi_s \widetilde{\mathbf{n}}_s. \tag{7}$$

We view $\boldsymbol{\delta}(\mathbf{v}, \boldsymbol{\xi})$ as a correction vector and allow only correction satisfying

$$\boxed{\|\boldsymbol{\delta}\|_2 \le c\|\mathbf{g}\|_2, \qquad \mathbf{g}^\top \boldsymbol{\delta} \ge 0, \qquad \mathbf{p}_i^\top \boldsymbol{\delta} \ge 0 \quad i = 1, \dots, m,} \tag{8}$$

where $c > 0$ limits the departure from GRPO. The second constraint gives $\mathbf{g}^\top \mathbf{d} \ge \|\mathbf{g}\|_2^2$, so NCR preserves at least GRPO's first-order progress on the sampled objective. The third constraint gives $\mathbf{p}_i^\top \mathbf{d} \ge \mathbf{p}_i^\top \mathbf{g}$ for every sampled success, so NCR does not worsen its first-order log-likelihood change relative to GRPO. Finally, $\mathbf{w} = \bar{\mathbf{w}}$ and $\boldsymbol{\xi} = \mathbf{0}$ recover $\mathbf{d} = \mathbf{g}$, so GRPO is always feasible.

4.2 Nash allocation of likelihood improvement

For successful response i, $\mathbf{p}_i^\top \boldsymbol{\delta}$ is the additional first-order log-likelihood slope contributed by the NCR correction. With correction radius $R = c\|\mathbf{g}\|_2$, define $U_i(\boldsymbol{\delta}) = 1 + \mathbf{p}_i^\top \boldsymbol{\delta}/(R\|\mathbf{p}_i\|_2)$. Under (8), $1 \le U_i \le 2$. Thus $U_i = 1$ means that NCR gives response i no additional support beyond GRPO. Using the bargaining powers q_i defined in Section 4.3, NCR selects

$$(\mathbf{w}_\mathsf{N}, \boldsymbol{\xi}_\mathsf{N}) \in \underset{\mathbf{w} \in \Delta_m,\ \boldsymbol{\xi} \in [0,1]^S}{\arg\max} \sum_{i=1}^m q_i \log U_i\big(\boldsymbol{\delta}(\mathbf{w}, \boldsymbol{\xi})\big) \quad \text{subject to (8)}, \tag{9}$$

Among welfare maximizers, we choose the correction with minimum norm. The role of the logarithm is seen from its marginal value: $\partial(\sum_j q_j \log U_j)/\partial U_i = q_i/U_i$. Nash gives less value to further improving a response that has already received more normalized support, creating pressure to allocate the remaining correction toward less-supported responses.

4.3 Powers from redundant directional representation

Repeated successful responses with similar likelihood-gradient directions should not receive additional bargaining power solely because they are sampled multiple times. For each successful response, let $\mathbf{v}_i = \mathbf{p}_i/\|\mathbf{p}_i\|_2$ and define

$$B_{ij} = \left(\frac{1 + \mathbf{v}_i^\top \mathbf{v}_j}{2}\right)^2, \qquad \mathbf{B} = [B_{ij}].$$

The shifted cosine maps aligned, orthogonal, and opposite directions to 1, $1/2$, and 0, respectively, before squaring. For $\mathbf{q} \in \Delta_m$, define the collision objective

$$C(\mathbf{B}, \mathbf{q}) = \sum_{i,j} q_i q_j B_{ij} = \mathbf{q}^\top \mathbf{B} \mathbf{q}. \tag{10}$$

Thus, minimizing C discourages concentrating bargaining power on highly similar directions. This objective is closely related to similarity-sensitive order-2 diversity measures (Leinster & Cobbold, 2012; Friedman & Dieng, 2023; Pasarkar & Dieng, 2024). Appendix M shows that exact duplicate directions affect C only through their aggregate power.

To keep every sampled success represented, for $m \geq 2$ we restrict the powers to

$$\mathcal{Q}_m = \{\mathbf{q} \in \Delta_m : q_i \leq (m-1) q_j \text{ for all } i, j\}, \qquad C_{\min} = \min_{\mathbf{q} \in \mathcal{Q}_m} C(\mathbf{B}, \mathbf{q}), \tag{11}$$

and select

$$\mathbf{q}_* = \operatorname*{arg\,min}_{\substack{\mathbf{q} \in \mathcal{Q}_m \\ C(\mathbf{B}, \mathbf{q}) = C_{\min}}} \operatorname{Variance}(\mathbf{q}). \tag{12}$$

For $m = 1$, we set $q_1 = 1$. The first stage minimizes directional collision, while the second only resolves ties. Appendix M gives the duplicate-aggregation result, properties of the ratio constraint, and the existence and uniqueness of the selected powers.

5 Theoretical analysis

A credit-allocation strategy A chooses $(\mathbf{w}_A, \boldsymbol{\xi}_A)$ and produces a correction $\boldsymbol{\delta}_A = \boldsymbol{\delta}(\mathbf{w}_A, \boldsymbol{\xi}_A)$ through (7), with update direction $\mathbf{d}_A = \mathbf{g} + \boldsymbol{\delta}_A$. All strategies in this section use the same feasible set

$$\mathcal{C} = \left\{\boldsymbol{\delta}(\mathbf{w}, \boldsymbol{\xi}) : \mathbf{w} \in \Delta_m,\ \boldsymbol{\xi} \in [0,1]^S,\ \|\boldsymbol{\delta}\|_2 \leq R,\ \mathbf{g}^\top \boldsymbol{\delta} \geq 0,\ \mathbf{p}_i^\top \boldsymbol{\delta} \geq 0\ \forall i\right\}. \tag{13}$$

We write $\boldsymbol{\delta}_{\mathsf{N}}$ for NCR, which maximizes (9), and $\boldsymbol{\delta}_{\mathsf{L}}$ for the matched linear strategy, which replaces only the Nash objective by $\sum_i q_i U_i$ while keeping the same feasible set and minimum-norm tie-break. GRPO makes no credit change, so $\boldsymbol{\delta}_{\mathsf{G}} = \mathbf{0}$.

5.1 Comparison with the current GRPO step

The constraints in (8) retain GRPO's first-order loss decrease and prevent any positive response from having a worse log-likelihood slope. These requirements apply to every feasible allocation.

Theorem 1 (Sampled-loss decrease and successful-response likelihoods)**.** *The maximization in* (9) *attains its optimum, and its norm tie-break selects a unique correction* $\boldsymbol{\delta}_{\mathsf{N}}$*. Every feasible allocation* A *satisfies*

$$-\nabla \mathcal{L}_{\mathrm{GRPO}}(\boldsymbol{\theta})^\top \mathbf{d}_A \geq \|\mathbf{g}\|_2^2, \qquad \mathbf{p}_i^\top \mathbf{d}_A \geq \mathbf{p}_i^\top \mathbf{g} \quad (i = 1, \ldots, m). \tag{14}$$

A response whose likelihood decreases under GRPO has a nonnegative first-order change under A *exactly when* $\mathbf{p}_i^\top \boldsymbol{\delta}_A \geq -\mathbf{p}_i^\top \mathbf{g}$*.*

Proof: Appendix G.

5.2 THE ADVANTAGE ATTRIBUTABLE TO NASH

Because the test-accuracy gradient is unavailable during training, we cannot directly choose the credit allocation that best improves held-out performance. We therefore model the test gradient by a declared family of weighted combinations of the observed successful directions and compare feasible corrections by their worst first-order test gain over this family. We then show that, at the uncertainty radius induced by the Nash solution, the Nash correction maximizes this worst-case gain. Let $\mathfrak{Q}$ be the declared family of test-prompt distributions and $\mathbf{h}_Q = \nabla \operatorname{Acc}_Q(\boldsymbol{\theta})$ for $Q \in \mathfrak{Q}$. We model a possible test gradient using the observed successful directions $\mathbf{V}$ and an unknown relevance vector $\mathbf{r} \in \Delta_m$. The vector $\mathbf{Vr}$ is therefore a weighted combination of the successful directions. Without further clarification, $\mathbf{r}$ is any point in the simplex. Our theory considers the transfer regime in which the relevant test gradients are represented by weights that remain close to the training powers $\mathbf{q}$. For a radius $\rho \geq 0$, define $\mathcal{R}_\rho(\mathbf{q}) = \{\mathbf{r} \in \Delta_m : \mathrm{KL}(\mathbf{q}\|\mathbf{r}) \leq \rho\}$.

Assumption 1 (Representation of the test family)**.** *There exist constants $\alpha > 0$ and $e_{\mathrm{tr}} \geq 0$ such that (1) for every $Q \in \mathfrak{Q}$, there exists $\mathbf{r} \in \mathcal{R}_\rho(\mathbf{q})$ satisfying $\|\mathbf{h}_Q - \alpha \mathbf{Vr}\|_2 \leq e_{\mathrm{tr}}$; (2) for every $\mathbf{r} \in \mathcal{R}_\rho(\mathbf{q})$, there exists $Q \in \mathfrak{Q}$ satisfying the same bound.*

This assumption gives the neighborhood $\mathcal{R}_\rho(\mathbf{q})$ its test-time meaning: the first condition requires every test gradient in the declared family to be represented, up to scale and error, by relevance weights near $\mathbf{q}$; the converse ensures that the worst-case relevance weights considered below also correspond to gradients in that family. For a feasible correction $\boldsymbol{\delta}$, the first-order test-accuracy gain relative to GRPO under Q is proportional to $\mathbf{h}_Q^\top \boldsymbol{\delta}$. Under Assumption 1, its modeled counterpart for relevance weights $\mathbf{r}$ is $\alpha(\mathbf{Vr})^\top \boldsymbol{\delta}$. We therefore define

$$\mathcal{A}_\rho(\boldsymbol{\delta}) := \min_{\mathbf{r} \in \mathcal{R}_\rho(\mathbf{q})} (\mathbf{Vr})^\top \boldsymbol{\delta}. \tag{15}$$

Thus $\alpha \mathcal{A}_\rho(\boldsymbol{\delta})$ is the worst first-order test gain predicted by this representation model, before accounting for the approximation error e_{tr}. The next subsection converts this quantity into a held-out accuracy bound. We now compare credit-allocation strategies on $\mathcal{A}_\rho$. For fixed powers $\mathbf{q}$, Nash maximizes $F(\boldsymbol{\delta}) = \sum_{i=1}^m q_i \log U_i(\boldsymbol{\delta})$, with $\nabla F(\boldsymbol{\delta}) = \frac{1}{R} \sum_{i=1}^m \frac{q_i}{U_i(\boldsymbol{\delta})} \mathbf{v}_i$. Hence q_i / U_i is the marginal objective value assigned by Nash to further improvement of response i.

Theorem 2 (Nash maximizes worst-case gain under relevance uncertainty)**.** *Define*

$$U_{N,i} := U_i(\boldsymbol{\delta}_N), \qquad r_{N,i} = \frac{q_i / U_{N,i}}{\sum_j q_j / U_{N,j}}, \qquad \rho_N = \mathrm{KL}(\mathbf{q}\|\mathbf{r}_N). \tag{16}$$

$$\textit{Then} \quad \boldsymbol{\delta}_N \in \arg\max_{\boldsymbol{\delta} \in \mathcal{C}} \mathcal{A}_{\rho_N}(\boldsymbol{\delta}), \qquad \mathcal{A}_{\rho_N}(\boldsymbol{\delta}_N) = R \left[\left(\sum_i \frac{q_i}{U_{N,i}} \right)^{-1} - 1 \right]. \tag{17}$$

Consequently, every feasible credit-allocation strategy A satisfies

$$\mathcal{A}_{\rho_N}(\boldsymbol{\delta}_N) \geq \mathcal{A}_{\rho_N}(\boldsymbol{\delta}_A). \tag{18}$$

Proof: Appendix H. Theorem 2 shows that, at the corresponding uncertainty radius ρ_N, the Nash correction gives the largest guaranteed modeled test gain among all feasible credit allocations.

For GRPO, $\boldsymbol{\delta}_G = \mathbf{0}$, so $\mathcal{A}_{\rho_N}(\boldsymbol{\delta}_G) = 0$. Nash therefore has the exact gain

$$\mathcal{A}_{\rho_N}(\boldsymbol{\delta}_N) - \mathcal{A}_{\rho_N}(\boldsymbol{\delta}_G) = R \left[\left(\sum_i \frac{q_i}{U_{N,i}} \right)^{-1} - 1 \right]. \tag{19}$$

Because $U_{N,i} \geq 1$ for every i, this gain is strictly positive whenever at least one successful response receives strict additional improvement. The matched linear strategy is also feasible, hence

$$\mathcal{A}_{\rho_N}(\boldsymbol{\delta}_N) \geq \mathcal{A}_{\rho_N}(\boldsymbol{\delta}_L). \tag{20}$$

Nash and linear use the same bargaining powers and the same feasible credit changes, so this comparison isolates the effect of the Nash log-welfare objective. ρ_N is induced by the Nash solution. The test-time interpretation of Theorem 2 applies when Assumption 1 holds for $\rho = \rho_N$.

5.3 FROM ADDITIONAL ALIGNMENT TO HELD-OUT ACCURACY

For any two feasible allocations A and B and a common step size $\eta > 0$, define the worst-case accuracy advantage of A over B as

$$\mathcal{I}_\eta(A; B) = \inf_{Q \in \mathfrak{Q}} \left[\mathrm{Acc}_Q(\boldsymbol{\theta} + \eta \mathbf{d}_A) - \mathrm{Acc}_Q(\boldsymbol{\theta} + \eta \mathbf{d}_B) \right]. \tag{21}$$

The two accuracies in each difference are evaluated under the same prompt distribution Q. We use N, L, and G to denote the Nash, linear, and GRPO allocations, respectively. Thus, $\mathcal{I}_\eta(A; \mathsf{G})$ is the worst-case accuracy advantage of allocation A over GRPO. To extend the first-order comparison to a finite update, we require the test-accuracy functions to be locally smooth.

Assumption 2 (Smooth test accuracy)**.** *For every $Q \in \mathfrak{Q}$, Acc_Q is differentiable and its gradient is L_{acc}-Lipschitz on a common convex neighborhood containing $\boldsymbol{\theta}, \boldsymbol{\theta} + \eta \mathbf{g}, \boldsymbol{\theta} + \eta \mathbf{d}_{\mathsf{N}}, \boldsymbol{\theta} + \eta \mathbf{d}_{\mathsf{L}}$. That is, for all points $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ in this neighborhood,*

$$\|\nabla \mathrm{Acc}_Q(\boldsymbol{\theta}_1) - \nabla \mathrm{Acc}_Q(\boldsymbol{\theta}_2)\|_2 \le L_{\mathrm{acc}} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2.$$

Theorem 3 (Worst-case accuracy comparison)**.** *Under Assumptions 1 and 2, ... $M \in \{\mathsf{N}, \mathsf{L}\}$, define Then*

$$\mathcal{I}_\eta(\mathsf{N}; \mathsf{G}) \ge \eta\alpha\, \mathcal{A}_\rho(\boldsymbol{\delta}_{\mathsf{N}}) - E_{\mathsf{N}}, \tag{22}$$

$$\mathcal{I}_\eta(\mathsf{N}; \mathsf{G}) - \mathcal{I}_\eta(\mathsf{L}; \mathsf{G}) \ge \eta\alpha \left[\mathcal{A}_\rho(\boldsymbol{\delta}_{\mathsf{N}}) - \mathcal{A}_\rho(\boldsymbol{\delta}_{\mathsf{L}}) \right] - E_{\mathsf{N}} - E_{\mathsf{L}} \textit{.where,} \tag{23}$$

$$E_M = \eta e_{\mathrm{tr}} \|\boldsymbol{\delta}_M\|_2 + L_{\mathrm{acc}} \eta^2 \|\boldsymbol{\delta}_M\|_2 \left(\|\mathbf{g}\|_2 + \frac{1}{2} \|\boldsymbol{\delta}_M\|_2 \right). \tag{24}$$

The proof of this theorem is in Appendix I. This theorem establishes a lower bound on the worst-case test-accuracy advantage of the Nash-welfare allocation over both linear allocation and the standard GRPO allocation. The bound consists of a first-order approximation term together with a second-order correction term. When the step size η is sufficiently small and α is sufficiently larger than e_{tr} , the right-hand sides of(22) and (23) become positive. Consequently, the Nash-welfare allocation is guaranteed to achieve a positive improvement under the worst-case test scenario.

6 EMPIRICAL EVALUATION

We evaluate NCR across three model scales: Qwen3-1.7B, Qwen3-4B (Yang et al., 2025), and Qwen2.5-Math-7B (Yang et al., 2024). We consider five mathematical reasoning benchmarks: AIME24 (MAA Committees, 2024), AIME25 (MAA Committees, 2025), AIME26 (MAA Committees, 2026), BeyondAIME (ByteDance-Seed, 2025), and BRUMO25 (Dekoninck et al., 2026). We compare NCR against the base checkpoint, GRPO (Shao et al., 2024), NTHR (Deng et al., 2025), and Div-R1 (Yao et al., 2025). We report Pass@1 and Pass@128, averaged across the five benchmarks. To evaluate generalization beyond mathematical reasoning, we additionally report ARC-C (Clark et al., 2018) and the MMLU-Pro-NoMath-Sml subset of MMLU-Pro (Wang et al., 2024), summarized by their average as *OOD Avg*. ARC-C is averaged across eight evaluation seeds.

NCR is particularly consistent under multi-sample evaluation: it obtains the best or tied-best Pass@128 on all 15 model–benchmark pairs and the highest Math Avg Pass@128 for every model. Relative to GRPO, the macro-average Pass@1 / Pass@128 gains are respectively $+1.01/+9.13$ points on Qwen3-1.7B, $+0.17/+5.93$ on Qwen3-4B, and $+0.35/+5.86$ on Qwen2.5-Math-7B. The Pass@1 gains are therefore smaller and more benchmark-dependent than the Pass@128 gains, but the macro-average remains higher for NCR at all three scales. NCR also gives the highest OOD Avg in each model block: 62.50, 76.28, and 62.85, corresponding to $+0.85$, $+0.55$, and $+0.04$ points over GRPO, respectively. Individual OOD sub-benchmarks need not be best simultaneously, so we interpret this as a modest aggregate transfer result rather than uniform dominance.

Figure 2 complements the endpoint statistics in Table 1. Across the three model scales and five benchmarks, NCR's advantage is most systematic at larger k, consistent with the strong Pass@128 results. This is operationally useful for test-time sampling: for a fixed answer budget, the NCR checkpoints more often include at least one correct response. We avoid interpreting these curves as direct evidence of semantic response diversity.

Table 1: **Main results (%).** Mathematics performance is reported as Pass@1 and Pass@128. Math Avg is the unweighted macro-average over the five mathematics benchmarks; OOD Avg averages ARC-C and MMLU-Pro-NoMath-Sml (MMLU-Pro in the table). Bold marks the best result and underlining the second-best distinct result within each model block; ties share a rank.

Method	In-Domain Math												Out-of-Domain		
	AIME24		AIME25		AIME26		BeyondAIME		BRUMO25		Math Avg		ARC-C	MMLU-Pro	OOD Avg
	P@1	P@128	P@1	P@128	P@1	P@128	P@1	P@128	P@1	P@128	P@1	P@128	Avg	Avg	Avg
Qwen3-1.7B															
Base	3.67	40.00	3.12	50.00	2.40	40.00	1.30	31.00	7.45	46.67	3.59	41.53	69.36	28.57	48.97
GRPO	14.84	43.33	9.60	46.70	7.58	40.00	2.37	32.00	14.92	53.33	9.86	43.07	79.95	43.35	61.65
NTHR	16.28	43.33	9.30	43.33	8.36	36.67	2.50	34.00	11.56	50.00	9.60	41.47	79.05	42.33	60.69
Div-R1	13.67	46.67	9.69	46.67	**8.59**	40.00	1.94	37.00	12.70	56.67	9.32	45.40	78.67	42.86	60.77
NCR	**16.41**	**50.00**	**10.05**	**63.33**	7.53	**43.33**	**2.85**	**41.00**	**17.50**	**63.33**	**10.87**	**52.20**	**80.43**	**44.56**	**62.50**
Qwen3-4B															
Base	11.04	60.00	7.89	50.00	7.11	46.67	4.86	39.00	19.01	56.67	9.98	50.47	82.60	48.88	65.74
GRPO	**28.23**	73.33	**27.40**	70.00	21.35	70.00	**10.77**	55.00	33.07	80.00	24.16	69.67	**91.01**	60.44	75.73
NTHR	28.15	73.33	25.99	66.67	21.74	70.00	10.22	53.00	33.67	80.00	23.95	68.60	90.47	**62.03**	76.25
Div-R1	27.71	70.00	24.87	73.33	21.17	73.33	9.42	55.00	**34.27**	80.00	23.49	70.33	88.98	60.70	74.84
NCR	28.20	**83.33**	26.54	**76.67**	**22.40**	**76.67**	**10.77**	**58.00**	33.75	**83.33**	**24.33**	**75.60**	90.87	61.69	**76.28**
Qwen2.5-Math-7B															
Base	15.60	70.00	6.28	53.33	7.58	**46.67**	3.53	41.00	13.46	60.00	9.29	54.20	66.06	26.53	46.30
GRPO	36.15	73.33	**19.61**	56.67	12.89	43.33	7.51	42.00	22.97	60.00	19.83	55.07	**79.50**	46.12	62.81
NTHR	34.27	70.00	17.58	60.00	**15.26**	36.67	**7.57**	40.00	23.18	56.67	19.57	52.67	78.89	45.21	62.05
Div-R1	35.76	76.67	17.03	63.33	11.12	43.33	7.52	44.00	21.41	60.00	18.57	57.47	77.73	44.45	61.09
NCR	**37.32**	**80.00**	18.83	**66.67**	14.27	**46.67**	7.29	**48.00**	**23.20**	**63.33**	**20.18**	**60.93**	78.90	**46.80**	**62.85**

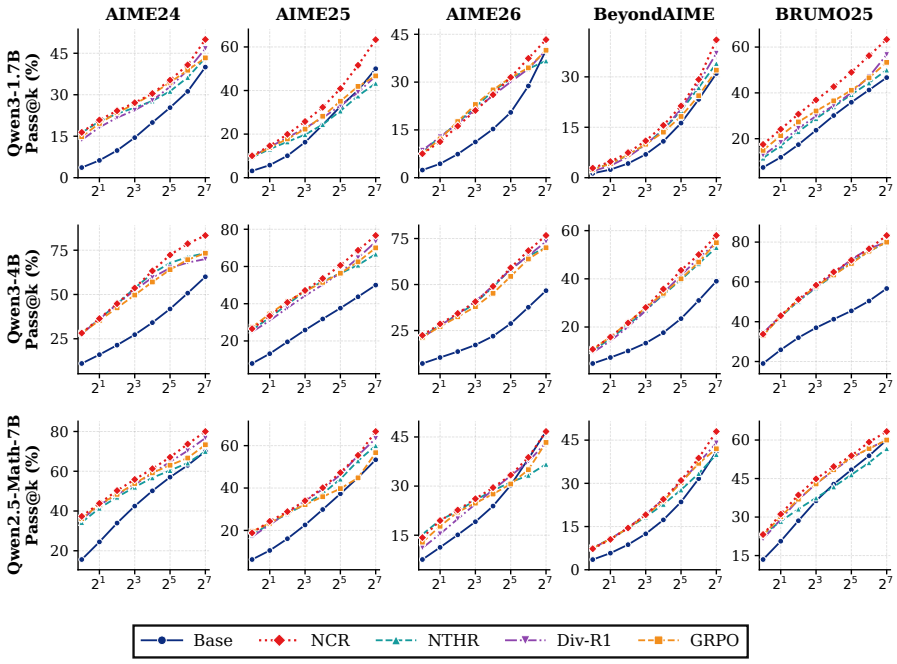

Figure 2: **Pass@k across models and mathematics benchmarks.** NCR generally stays on or near the upper envelope as the sampling budget increases, showing that the gains in Table 1 persist across a range of inference budgets.

Across the three timing runs, the NCR module requires 29, 81, and 121 seconds per step, with corresponding total wall-clock times of 144, 396, and 867 seconds Table 3. Despite the additional NCR computation, its overhead remains modest relative to the overall training time, making NCR practical to incorporate into the training pipeline. These timing measurements are descriptive; the available package does not provide run-level uncertainty estimates or matched checkpoint-selection records.

Table 2: **Component ablation of NCR on Qwen3-1.7B.** Linear NCR replaces the Nash objective with the matched linear objective; Uniform-power NCR uses $q_i = 1/m$; and Response-level NCR ties refund fractions within each failed response. Math Avg denotes the macro-average across the five benchmarks.

Variant	NCR components			Benchmark performance										Math Avg	
	Powers	Objective	Refund	AIME24		AIME25		AIME26		BeyondAIME		BRUMO25		P@1	P@128
				P@1	P@128	P@1	P@128	P@1	P@128	P@1	P@128	P@1	P@128		
GRPO	–	–	None	14.84	43.33	9.60	46.70	7.58	40.00	2.37	32.00	14.92	53.33	9.86	43.07
Linear NCR	Collision	Linear	Segment	17.30	56.67	8.60	40.00	8.45	33.33	1.95	39.00	13.70	56.67	10.00	45.13
Uniform-power NCR	Uniform	Nash	Segment	14.90	43.33	11.20	63.33	6.65	36.67	3.30	45.00	15.55	50.00	10.32	47.67
Response-level NCR	Collision	Nash	Response	17.05	53.33	9.15	56.67	8.20	46.67	2.20	35.00	16.15	60.00	10.55	50.33
Full NCR	Collision	Nash	Segment	16.41	50.00	10.05	63.33	7.53	43.33	2.85	41.00	17.50	63.33	**10.87**	**52.20**

Table 3: **Average per-step runtime (seconds).** Module times and total wall-clock times are reported as recorded in the supplied experiment log.

Model + dataset	Data Gen	Model Upd	NCR	Ref	Old Prob	Total
Qwen2.5-Math-7B (DAPO-Math-17k)	51	79	29	–	16	144
Qwen3-1.7B (DAPO-Math-17k)	198	170	81	–	28	396
Qwen3-4B (DAPO-Math-17k)	504	314	121	–	52	867

The sweep in Figure 3 exhibits a clear stability–freedom trade-off. Increasing c from 0.2 to 0.5 raises Pass@1 monotonically from 41.83 to 44.18. Pass@128 rises from 75.00 at $c = 0.2$ to a peak of 77.57 at $c = 0.4$, then decreases to 76.47 at $c = 0.5$. Thus the intermediate regime around $c = 0.4$ captures most of the single-sample gain without sacrificing the multi-sample success rate observed at the largest radius.

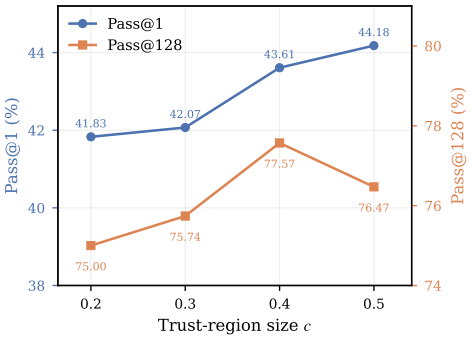

Figure 3: **Sensitivity to the trust-region size c.** Pass@1 increases with c, while Pass@128 peaks at an intermediate value.

We ablate the three components of NCR on Qwen3-1.7B while keeping the training data, number of updates, correction radius c, and evaluation protocol fixed. Linear NCR replaces the weighted Nash objective $\sum_i q_i \log U_i$ with the matched linear objective $\sum_i q_i U_i$; Uniform-power NCR uses $q_i = 1/m$; and Positive-only NCR disables negative-credit refunds by fixing $\boldsymbol{\xi} = \mathbf{0}$.

Table 2 isolates the three design choices in NCR. Removing collision-aware powers, replacing the Nash objective, or tying refunds at the response level each reduces the overall Math Avg, while Full NCR achieves the strongest aggregate Pass@1 and Pass@128.

7 Conclusion

We introduced NCR, a credit-routing framework that refines the GRPO update by jointly reallocating positive credit among successful responses and refunding selected negative contributions. Collision-aware powers reduce redundant representation, while Nash welfare favors trajectories with higher marginal value under explicit trust-region and protection constraints. Our analysis characterizes when this local allocation yields a stronger robust improvement criterion, and the empirical results show consistent gains across model scales, sampling budgets, and out-of-domain evaluation. Together, these results suggest that reasoning RL can benefit not only from identifying which responses succeed or fail, but also from allocating their learning signal more carefully. Limitations are discussed in Appendix N.

AI USE STATEMENT

Generative AI tools were used to review the manuscript for grammatical errors and clarity of exposition, and to assist in checking the mathematical proofs for logical and algebraic consistency. All AI-assisted feedback was independently assessed by the authors, who verified the final text and proofs and take full responsibility for the manuscript.

REPRODUCIBILITY STATEMENT

Appendix E specifies outer rollout stages and multiple inner optimizer steps, the Nash optimization (9), the unprojected head-feature construction, failure handling, and the matched linear control. All main proofs are in Appendices G–I; Appendix J treats finite groups and reused batches, and Appendix M proves the new collision-power statements. The source package includes executable finite-dimensional checks, not a production LLM trainer. Section 6 reports recorded checkpoint scores; the matching run records and mechanism diagnostics remain unavailable.

ETHICS STATEMENT

Binary final-answer verification does not establish the validity of every reasoning step. Gradient-based refunds should not be interpreted as process supervision or error localization. Any released policy must be evaluated for its intended use and comply with the licenses and privacy requirements of its data and base model. Fairness between gradient utilities is not a demographic fairness guarantee. Data- and deployment-specific considerations must be completed with the experimental study.

REFERENCES

ByteDance-Seed. Beyondaime: Advancing math reasoning evaluation beyond high school olympiads. https://huggingface.co/datasets/ByteDance-Seed/BeyondAIME, 2025.

Haoxuan Chen, Tianming Liang, Wei-Shi Zheng, and Jian-Fang Hu. Breaking winner-takes-all: Cooperative policy optimization improves diverse LLM reasoning. *arXiv preprint arXiv:2605.11461*, 2026. URL https://arxiv.org/abs/2605.11461.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*, 2018.

Jasper Dekoninck, Nikola Jovanović, Tim Gehrunger, Kári Rögnvaldsson, Ivo Petrov, Chenhao Sun, and Martin Vechev. Beyond benchmarks: Matharena as an evaluation platform for mathematics with llms, 2026. URL https://arxiv.org/abs/2605.00674.

Wenlong Deng, Yi Ren, Muchen Li, Danica J. Sutherland, Xiaoxiao Li, and Christos Thrampoulidis. On the effect of negative gradient in group relative deep reinforcement optimization. *arXiv preprint arXiv:2505.18830*, 2025. URL https://arxiv.org/abs/2505.18830.

Wenlong Deng, Yi Ren, Yushu Li, Boying Gong, Danica Sutherland, Xiaoxiao Li, and Christos Thrampoulidis. Token hidden reward: Steering exploration-exploitation in group relative deep reinforcement learning. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.03669v4.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023. URL https://arxiv.org/abs/2210.02410.

Tom Leinster and Christina A. Cobbold. Measuring diversity: The importance of species similarity. *Ecology*, 93(3):477–489, 2012. doi: 10.1890/10-2402.1.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://arxiv.org/abs/2110.14048.

Maoqi Liu, Qingchun Meng, Guodong Yu, and Zhi-Hai Zhang. Fairness as a robust utilitarianism. *Production and Operations Management*, 34(4), 2025. doi: 10.1177/10591478241262285. URL https://doi.org/10.1177/10591478241262285. First published online in 2024. See also the supplementary proof of Proposition 7.

Anamika Lochab, Bolian Li, and Ruqi Zhang. Uniform-correct policy optimization: Breaking rlvr's indifference to diversity, 2026a. URL https://arxiv.org/abs/2605.00365.

Anamika Lochab, Bolian Li, and Ruqi Zhang. Uniform-correct policy optimization: Breaking RLVR's indifference to diversity. *arXiv preprint arXiv:2605.00365*, 2026b. URL https://arxiv.org/abs/2605.00365.

MAA Committees. Aime problems and solutions. https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions, 2024. Art of Problem Solving.

MAA Committees. Aime problems and solutions. https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions, 2025. Art of Problem Solving.

MAA Committees. Aime problems and solutions. https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions, 2026. Art of Problem Solving.

Aviv Navon, Aviv Shamsian, Idan Achituve, Haggai Maron, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. Multi-task learning as a bargaining game. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16428–16446, 2022. URL https://proceedings.mlr.press/v162/navon22a.html.

Kishan Panaganti, Zhenwen Liang, Wenhao Yu, Haitao Mi, and Dong Yu. Group distributionally robust optimization-driven reinforcement learning for LLM reasoning. *arXiv preprint arXiv:2601.19280*, 2026. URL https://arxiv.org/abs/2601.19280.

Amey P. Pasarkar and Adji Bousso Dieng. Cousins of the Vendi score: A family of similarity-based diversity metrics for science and machine learning. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 3808–3816. PMLR, 2024.

Shyam Sundhar Ramesh, Xiaotong Ji, Matthieu Zimmer, Sangwoong Yoon, Zhiyong Wang, Haitham Bou Ammar, Aurelien Lucchi, and Ilija Bogunovic. Multi-task GRPO: Reliable LLM reasoning across tasks. *arXiv preprint arXiv:2602.05547*, 2026. URL https://arxiv.org/abs/2602.05547. Version 3; authors report acceptance at ICML 2026.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://arxiv.org/abs/1810.04650.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Bart P. G. Van Parys, Peyman Mohajerin Esfahani, and Daniel Kuhn. From data to decisions: Distributionally robust optimization is optimal. *arXiv preprint arXiv:1704.04118*, 2017. URL https://arxiv.org/abs/1704.04118. Version 3, revised December 2019.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 95266–95290. Curran Associates, Inc., 2024. doi: 10.52202/079017-3018. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ad236edc564f3e3156e1b2feafb99a24-Paper-Datasets_and_Benchmarks_Track.pdf.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992. doi: 10.1007/BF00992696.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement, 2024. URL https://arxiv.org/abs/2409.12122.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Jian Yao, Ran Cheng, Xingyu Wu, Jibin Wu, and Kay Chen Tan. Diversity-aware policy optimization for large language model reasoning. *arXiv preprint arXiv:2505.23433*, 2025. URL https://arxiv.org/abs/2505.23433.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025. URL https://arxiv.org/abs/2503.14476.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://arxiv.org/abs/2001.06782.

Zepeng Zhai, Meilin Chen, Jiaxuan Zhao, Junlang Qian, Lei Shen, and Yuan Lu. Rewards as labels: Revisiting rlvr from a classification perspective, 2026. URL https://arxiv.org/abs/2602.05630.

Feng Zhang, Xinhong Ma, Ziqiang Dong, Xi Leng, Jianfei Zhao, Xin Sun, Yang Yang, and Guanjun Jiang. Revisiting reinforcement learning with verifiable rewards from a contrastive perspective. *arXiv preprint arXiv:2605.12969*, 2026a. URL https://arxiv.org/abs/2605.12969v4. Version 4.

Feng Zhang, Xinhong Ma, Ziqiang Dong, Xi Leng, Jianfei Zhao, Xin Sun, Yang Yang, and Guanjun Jiang. Revisiting reinforcement learning with verifiable rewards from a contrastive perspective, 2026b. URL https://arxiv.org/abs/2605.12969.

Kaiyi Zhang, Wei Wu, and Yankai Lin. Delta: Discriminative token credit assignment for reinforcement learning from verifiable rewards. *arXiv preprint arXiv:2605.21467*, 2026c. URL https://arxiv.org/abs/2605.21467.

APPENDIX

A RELATED WORK

RLVR credit assignment. GRPO uses group-relative response advantages and token-level clipped ratios; DAPO changes clipping, sampling, and token normalization (Shao et al., 2024; Yu et al., 2025). We adopt asymmetric clipping and omit reference KL, but keep response normalization. NTHR studies negative-gradient interference, THR uses token influence to steer learning, and DelTA reweights discriminative token gradients (Deng et al., 2025; 2026; Zhang et al., 2026c). Our response-level positive control is optimized jointly with negative-segment refunds, rather than assigning semantic correctness to those segments. ConSPO instead derives sequence-level contrastive scores for positive and negative responses (Zhang et al., 2026a). UCPO targets conditional-correct uniformity, while GCPO redistributes semantic-coverage credit (Lochab et al., 2026b; Chen et al., 2026). Our utility is additional likelihood support, not semantic coverage or a target response distribution.

Gradient allocation and robustness. MGDA, PCGrad, and CAGrad provide multi-objective aggregation methods; Nash-MTL uses bargaining to select task-gradient combinations (Sener & Koltun, 2018; Yu et al., 2020; Liu et al., 2021; Navon et al., 2022). Here the players are sampled successes, the feasible directions are generated by response coefficients and segment refunds, and the reference is the current clipped GRPO gradient. The Nash–robustness relation is prior work (Liu et al., 2025; Van Parys et al., 2017); the comparison here uses the actual coefficient image and a separately stated accuracy condition. Task- and prompt-level robust reasoning methods (Ramesh et al., 2026; Panaganti et al., 2026) operate at a different allocation level.

B TRAINING AND EVALUATION DETAILS

Table 4: Full Training Hyperparameters

Hyperparameter	**Value**
Advantage Estimator	GRPO
Reward Function	DAPO
Trust-region size c	0.4
Use KL Loss	No
Use Entropy Regularization	No
Train Batch Size	256
Max Prompt Length	2048
Max Response Length (Qwen3 1.7B base)	8192
Max Response Length (Qwen3 4B base)	16384
Max Response Length (Qwen2.5-Math-7B)	4096
PPO Mini Batch Size	64
PPO Micro Batch Size/GPU	1
Clip Ratio Range	$[0.8, 1.28]$
Grad Clip	1.0
Learning Rate	10^{-6}
Warm-up Step	10
Training Temperature	1.0
Training Top-p	1.0
Validation Temperature	1.0
Validation Top-p	1.0
Rollout N	8

Table 5: Evaluation Settings

Model	Temperature	Top-p	Max Response Length
Qwen3 1.7B base	0.7	0.9	25000
Qwen3 4B base	0.7	0.9	30000
Qwen2.5-Math-7B	0.7	0.9	4096

C Evaluation Prompts

C.1 Math

Math Prompt Template

```
<|im_start|>user
{problem}

Please reason step by step, and put your final answer within
\boxed{}<|im_end|>
<|im_start|>assistant
```

C.2 ARC-Challenge

ARC-C Prompt Template

```
<|im_start|>user
{question}

A) {option_a}
B) {option_b}
C) {option_c}
D) {option_d}

Please reason step by step, and put your final answer within
\boxed{}<|im_end|>
<|im_start|>assistant
```

C.3 MMLU-Pro-NoMath-Sml

MMLU-Pro-NoMath-Sml Prompt Template

```
<|im_start|>user
The following are multiple choice questions (with answers) about
{category}.  Think step by step and then finish your answer with
"\boxed{X}" where X is the correct letter choice.
Question:  {question}
Options:
A. {option_a}
B. {option_b}
...

Please reason step by step, and put your final answer within
\boxed{}<|im_end|>
<|im_start|>assistant
```

D SELECTED-TOKEN GRADIENT PROXY

The full-gradient definitions in Section 4 distinguish the likelihood gradient $\mathbf{p}_i$ from the clipping-weighted score $\mathbf{s}_i$. Their proxy counterparts must retain this distinction. For token $y_{j,t}$, let $\mathbf{h}_{j,t}$ be its final hidden state and let $p_{j,t} = \pi_{\boldsymbol{\theta}}(y_{j,t} \mid x, y_{j,<t})$. The selected-token proxy is

$$\mathbf{u}_{j,t}^{\mathrm{prx}} = (1 - p_{j,t})\mathbf{h}_{j,t}. \tag{D.1}$$

For unit-temperature logits $\mathbf{W}\mathbf{h}_{j,t}$ with the hidden state held fixed, this is the derivative of the token log probability with respect to its selected output row (Zhang et al., 2026c). It is not the full output-matrix gradient $(\mathbf{e}_{y_{j,t}} - \boldsymbol{\pi}_{j,t})\mathbf{h}_{j,t}^{\top}$, where $\boldsymbol{\pi}_{j,t}$ is the vocabulary probability vector, or a full-policy gradient. In particular, placing selected rows for different tokens in one hidden-state space does not preserve full-gradient inner products.

The corresponding response and segment vectors are

$$\mathbf{p}_i^{\mathrm{prx}} = \frac{1}{T_i}\sum_t \mathbf{u}_{i,t}^{\mathrm{prx}}, \qquad \mathbf{s}_i^{\mathrm{prx}} = \frac{\sum_t \gamma_{i,t}\mathbf{u}_{i,t}^{\mathrm{prx}}}{\sum_t \gamma_{i,t}},$$
$$\widetilde{\mathbf{n}}_s^{\mathrm{prx}} = \frac{\sum_{t\in\mathcal{I}_s} \gamma_{j_s,t}\mathbf{u}_{j_s,t}^{\mathrm{prx}}}{\sum_{t\in\mathcal{I}_s} \gamma_{j_s,t}}. \tag{D.2}$$

A zero denominator defines the corresponding weighted vector to be zero. The scalar credits remain those in (2) and (3). At the rollout policy, these are $a_i = \widehat{A}_i/G$ and $\beta_s = (-\widehat{A}_{j_s}/G)|\mathcal{I}_s|/T_{j_s}$. Partitioning a failed response preserves its baseline contribution; merging segments need not preserve the available corrections when their refunds can differ.

For proxy-based coefficient selection, replace every vector in the power and routing programs by its proxy counterpart, including $\mathbf{g}^{\mathrm{prx}} = \sum_i a_i\mathbf{s}_i^{\mathrm{prx}} - \sum_s \beta_s\widetilde{\mathbf{n}}_s^{\mathrm{prx}}$ and $R^{\mathrm{prx}} = c\|\mathbf{g}^{\mathrm{prx}}\|_2$. The raw vectors $\mathbf{p}_i^{\mathrm{prx}}$ determine similarities, utilities, and likelihood constraints; the weighted vectors determine the correction. The resulting coefficients are detached before full-policy backpropagation (Appendix E). This realizes their full-policy linear combination, but proxy feasibility alone does not certify the full-gradient protections or radius constraint.

E NCR UPDATE AND TOKEN-LEVEL REALIZATION

Algorithm 1 specifies the current-group optimization in Section 4. It is written in the full-gradient geometry; proxy-based selection makes the substitutions in Appendix D. A rollout stage caches the sampled responses, rewards, advantages, and old-policy log probabilities. Every inner update recomputes the current ratios, clipping factors, vectors, powers, and allocations. Held-out data are not inputs.

Algorithm 1 Nash Credit Routing for one cached prompt group

Require: Current parameters $\boldsymbol{\theta}$; cached group of G responses and its rollout probabilities; fixed advantages $\widehat{A}_j$; failed-response token partition; radius factor $c \geq 0$.

Failure rule: If any solve or feasibility check fails, return the GRPO coefficients $\widehat{A}_j \gamma_{j,t}/(GT_j)$.

Compute $\rho_{j,t}(\boldsymbol{\theta})$ and $\gamma_{j,t}$ by (F.1), and the current token scores $\mathbf{u}_{j,t}$.
Compute $a_i, \mathbf{s}_i, \beta_s, \widetilde{\mathbf{n}}_s$ by (2)–(3), and $\mathbf{p}_i = T_i^{-1} \sum_t \mathbf{u}_{i,t}$.
Set $\mathbf{g} = \sum_i a_i \mathbf{s}_i - \sum_s \beta_s \widetilde{\mathbf{n}}_s$, $\Gamma = \sum_i a_i$, and $R = c\|\mathbf{g}\|_2$.
if $m = 0$, $\Gamma = 0$, $R = 0$, or some $\|\mathbf{p}_i\|_2 = 0$ **then**
    return the GRPO token coefficients $\widehat{A}_j \gamma_{j,t}/(GT_j)$.
end if
Set $\bar{w}_i = a_i/\Gamma$ and $\mathbf{v}_i = \mathbf{p}_i/\|\mathbf{p}_i\|_2$.
if $m = 1$ **then**
    Set $\mathbf{q}_* = (1)$.
else
    Form $B_{ik} = ((1 + \mathbf{v}_i^\top \mathbf{v}_k)/2)^2$.
    Solve $\mathbf{q}_0 \in \arg\min_{\mathbf{q} \in \mathcal{Q}_m} \mathbf{q}^\top \mathbf{B} \mathbf{q}$.
    Minimize $\|\mathbf{q} - \mathbf{1}/m\|_2^2$ over $\mathbf{q} \in \mathcal{Q}_m$ with $\mathbf{B}\mathbf{q} = \mathbf{B}\mathbf{q}_0$; retain the result as $\mathbf{q}_*$.
end if
With $\mathbf{q}_*$ fixed, solve (9) over $\mathbf{w} \in \Delta_m$, $\boldsymbol{\xi} \in [0,1]^S$ and the three constraints (8).
Among its maximizers, minimize $\|\boldsymbol{\delta}(\mathbf{w}, \boldsymbol{\xi})\|_2^2$; retain a representing pair $(\mathbf{w}_\mathsf{N}, \boldsymbol{\xi}_\mathsf{N})$.
return the detached token coefficients in (E.1) for $(\mathbf{w}_\mathsf{N}, \boldsymbol{\xi}_\mathsf{N})$.

The power tie-break is equivalent to minimizing variance because $\sum_i q_i = 1$; its linear equality follows from Lemma M.6. The routing tie-break fixes the optimal utility vector and minimizes correction norm. Neither tie-break adds a penalty to the first-stage objective. The ratio cap constrains powers $\mathbf{q}$, not credit shares $\mathbf{w}$. No post-solve clustering or radial projection is part of the stated method. Fallback requires a finite baseline; nonfinite model probabilities or gradients require rejecting the step rather than claiming a valid GRPO update.

E.1 Applying the selected coefficients

Let $Z_j = \sum_t \gamma_{j,t}$ and $Z_s = \sum_{t \in \mathcal{I}_s} \gamma_{j_s,t}$. For the selected allocation, define

$$b_{j,t} = \begin{cases} \Gamma w_j \gamma_{j,t}/Z_j, & j \leq m,\ Z_j > 0, \\ -\beta_s(1-\xi_s)\gamma_{j,t}/Z_s, & j = j_s,\ t \in \mathcal{I}_s,\ Z_s > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{E.1}$$

This also covers inactive responses and segments without dividing by their zero credit. With all $b_{j,t}$ detached at $\boldsymbol{\theta}$, use the local loss

$$\mathcal{L}_{\mathrm{NCR}}(\boldsymbol{\vartheta}; \boldsymbol{\theta}) = -\sum_{j=1}^{G} \sum_{t=1}^{T_j} \mathrm{sg}[b_{j,t}(\boldsymbol{\theta})] \log \pi_{\boldsymbol{\vartheta}}(y_{j,t} \mid x, y_{j,<t}). \tag{E.2}$$

Its gradient at the coefficient-evaluation point satisfies

$$-\nabla_{\boldsymbol{\vartheta}} \mathcal{L}_{\mathrm{NCR}}(\boldsymbol{\vartheta}; \boldsymbol{\theta})|_{\boldsymbol{\vartheta}=\boldsymbol{\theta}} = \Gamma \sum_i w_i \mathbf{s}_i - \sum_s \beta_s(1-\xi_s)\widetilde{\mathbf{n}}_s = \mathbf{g} + \boldsymbol{\delta}(\mathbf{w}, \boldsymbol{\xi}). \tag{E.3}$$

Indeed, substitution of (E.1) reconstructs each response or segment term, including the zero-activity convention. For an active success this is also the gradient obtained by multiplying its clipped-surrogate terms by $\Gamma w_i/a_i$; a failed segment uses $1 - \xi_s$. The local loss is refreshed after every parameter change, not optimized for multiple steps with stale $b_{j,t}$.

For several prompt groups, average their local losses with the same group normalization as GRPO. The exact per-group guarantees concern direct steps along the corresponding gradients; minibatch and optimizer qualifications are given in Appendix J.

F EXACT CURRENT-ITERATE GRPO CALCULUS

Fix the sampled responses, rewards, advantages, and rollout probabilities. Only the current parameters $\boldsymbol{\theta}$ are differentiated. Write $\ell_j(\boldsymbol{\theta}) = T_j^{-1} \log \pi_{\boldsymbol{\theta}}(y_j \mid x)$, so $\nabla_{\boldsymbol{\theta}} \ell_j = T_j^{-1} \sum_t \mathbf{u}_{j,t}$.

F.1 CLIPPING FACTORS AND TOKEN DERIVATIVES

Assume $0 < \epsilon_{\text{low}} < 1$ and $\epsilon_{\text{high}} > 0$. For the clipped token surrogate $\phi(\rho, A) = \min\{\rho A, \text{clip}(\rho, 1 - \epsilon_{\text{low}}, 1 + \epsilon_{\text{high}})A\}$, it equals $A \min\{\rho, 1 + \epsilon_{\text{high}}\}$ for $A > 0$ and $A \max\{\rho, 1 - \epsilon_{\text{low}}\}$ for $A < 0$; it is zero for $A = 0$. Define

$$\gamma_{j,t} = \rho_{j,t}(\boldsymbol{\theta}) \begin{cases} \mathbb{I}\{\rho_{j,t}(\boldsymbol{\theta}) < 1 + \epsilon_{\text{high}}\}, & \widehat{A}_j > 0, \\ \mathbb{I}\{\rho_{j,t}(\boldsymbol{\theta}) > 1 - \epsilon_{\text{low}}\}, & \widehat{A}_j < 0, \\ 0, & \widehat{A}_j = 0. \end{cases} \tag{F.1}$$

Since $\nabla_{\boldsymbol{\theta}} \rho_{j,t} = \rho_{j,t} \mathbf{u}_{j,t}$, differentiation away from the sign-dependent clipping thresholds gives

$$\mathbf{g} = -\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{GRPO}}(\boldsymbol{\theta}) = \frac{1}{G} \sum_{j=1}^{G} \frac{\widehat{A}_j}{T_j} \sum_t \gamma_{j,t} \mathbf{u}_{j,t}. \tag{F.2}$$

At an active threshold, the strict inequalities select the zero derivative of the flat branch. This defines the boundary update, not an ordinary derivative at a kink; differentiable-loss guarantees apply away from those thresholds. A positive-advantage token is not suppressed merely for falling below the lower bound, nor a negative-advantage token for exceeding the upper bound.

F.2 RESPONSE AND SEGMENT FACTORIZATION

Let $Z_j = \sum_t \gamma_{j,t}$. For $Z_j > 0$,

$$\frac{\widehat{A}_j}{G T_j} \sum_t \gamma_{j,t} \mathbf{u}_{j,t} = \underbrace{\frac{\widehat{A}_j Z_j}{G T_j}}_{a_j} \underbrace{\frac{\sum_t \gamma_{j,t} \mathbf{u}_{j,t}}{Z_j}}_{\mathbf{s}_j},$$

which proves (2). If $Z_j = 0$, set $a_j = 0$ and $\mathbf{s}_j = \mathbf{0}$. For a segment with $Z_s = \sum_{t \in \mathcal{I}_s} \gamma_{j_s,t} = 0$, likewise set $\beta_s = 0$ and $\widetilde{\mathbf{n}}_s = \mathbf{0}$. These cases differ from cancellation of a weighted vector with positive denominator. In particular, $a_i = 0$ need not imply $\mathbf{p}_i = \nabla \ell_i = \mathbf{0}$.

For each failed response j, its segments partition its original token positions, hence

$$-\sum_{s: j_s = j} \beta_s \widetilde{\mathbf{n}}_s = \frac{\widehat{A}_j}{G T_j} \sum_{s: j_s = j} \sum_{t \in \mathcal{I}_s} \gamma_{j,t} \mathbf{u}_{j,t} = a_j \mathbf{s}_j. \tag{F.3}$$

Summing over responses proves (3). Every token retains its original autoregressive prefix. At $\boldsymbol{\theta} = \boldsymbol{\theta}_{\text{old}}$ in a mixed group, $\gamma_{j,t} = 1$, $\mathbf{s}_i = \mathbf{p}_i$, and $\beta_s = (-\widehat{A}_{j_s}/G)|\mathcal{I}_s|/T_{j_s}$. During rollout reuse, nonuniform $\gamma_{j,t}$ generally make $\mathbf{s}_i$ and $\mathbf{p}_i$ nonparallel, so changing a scalar credit alone cannot replace the weighted score by the raw likelihood gradient. Homogeneous binary-reward groups have zero advantages and contribute no reward gradient.

F.3 INDIVIDUAL SUPPORT AND REPEATED CONTRIBUTIONS

The chain rule gives

$$\frac{d}{d\eta} \mathcal{L}_{\text{GRPO}}(\boldsymbol{\theta} + \eta \mathbf{g}) \Big|_{\eta = 0} = -\|\mathbf{g}\|_2^2, \qquad \frac{d}{d\eta} \ell_i(\boldsymbol{\theta} + \eta \mathbf{g}) \Big|_{\eta = 0} = \mathbf{p}_i^\top \mathbf{g}. \tag{F.4}$$

At the current iterate,

$$\mathbf{p}_i^\top \mathbf{g} = a_i \mathbf{p}_i^\top \mathbf{s}_i + \sum_{\substack{k \le m \\ k \ne i}} a_k \mathbf{p}_i^\top \mathbf{s}_k - \sum_s \beta_s \mathbf{p}_i^\top \widetilde{\mathbf{n}}_s, \tag{F.5}$$

which reduces to (5) at the rollout policy. A negative slope decreases the response likelihood for sufficiently small positive steps, despite the group-loss decrease. At a fresh mixed group, h identical successful directions contribute $ha^+\mathbf{p}$ because all successes share the same positive advantage. This identity concerns sampled multiplicity; it contains no held-out relevance information.

G FEASIBLE CORRECTIONS AND COMPARISON WITH GRPO

Work at a differentiable iterate with $\Gamma > 0$, $R > 0$, and nonzero successful likelihood gradients. Degenerate cases use the baseline as specified in Appendix E. Throughout this appendix, $\mathcal{C}$ is exactly (13); no additional bounds on $\mathbf{w}$ are imposed.

G.1 PROOF OF THEOREM 1

Proof. The coefficient domain $\Delta_m \times [0,1]^S$ is compact and convex. Its affine image under (7), intersected with the closed radius ball and protection halfspaces, is therefore compact and convex. It contains $\mathbf{0}$, represented by $(\bar{\mathbf{w}}, \mathbf{0})$. Since $\mathbf{p}_i^\top \boldsymbol{\delta} \geq 0$ and $\mathbf{p}_i^\top \boldsymbol{\delta} \leq \|\mathbf{p}_i\|_2 R$, each utility lies in $[1,2]$. The continuous concave Nash objective attains its maximum on $\mathcal{C}$. Its maximizer set is compact and convex, so minimizing squared norm on that set selects a unique correction, although its coefficient representation need not be unique.

For every feasible allocation,

$$\mathbf{g}^\top \mathbf{d}_A = \|\mathbf{g}\|_2^2 + \mathbf{g}^\top \boldsymbol{\delta}_A \geq \|\mathbf{g}\|_2^2, \qquad \mathbf{p}_i^\top \mathbf{d}_A = \mathbf{p}_i^\top \mathbf{g} + \mathbf{p}_i^\top \boldsymbol{\delta}_A \geq \mathbf{p}_i^\top \mathbf{g}.$$

Substituting $\mathbf{g} = -\nabla \mathcal{L}_{\mathrm{GRPO}}$ proves (14). A harmed response recovers a nonnegative slope exactly when $\mathbf{p}_i^\top \boldsymbol{\delta}_A \geq -\mathbf{p}_i^\top \mathbf{g}$. □

The likelihood comparison uses $\mathbf{p}_i$, including when $\mathbf{s}_i \neq \mathbf{p}_i$ or the response is fully clipped. It is relative to GRPO, not a guarantee that every successful-response likelihood increases.

G.2 PROPORTIONAL FAIRNESS AND THE NORM TIE-BREAK

Proposition G.1 (Proportional comparison of feasible utilities)**.** *For positive powers, the Nash utility vector $\mathbf{U}_{\mathsf{N}}$ is unique, and every feasible utility vector $\mathbf{U}$ satisfies*

$$\sum_i q_i \frac{U_i - U_{\mathsf{N},i}}{U_{\mathsf{N},i}} \leq 0. \tag{G.1}$$

No feasible vector weakly improves all its coordinates and strictly improves one. If any feasible correction improves a utility above the baseline $\mathbf{1}$, the Nash value is positive.

Proof. The attainable utility set is convex, and weighted log welfare is strictly concave on positive utilities. This gives uniqueness. Its directional derivative toward a feasible $\mathbf{U}$ gives (G.1); strict coordinatewise monotonicity excludes a dominating vector. Finally, a feasible strict improvement over $\mathbf{1}$ has positive log welfare. □

Conversely, $\log x \leq x - 1$ shows that (G.1) characterizes a maximizer. NCR optimizes $\log U_i$, not $\log(U_i - 1)$.

All Nash maximizers have the same utility vector. The norm tie-break can therefore fix that vector and minimize $\|\boldsymbol{\delta}\|_2^2$ under the original constraints. For the matched linear allocation, it instead fixes only the optimal scalar $\mathbf{q}^\top \mathbf{U}$; fixing a particular linear-optimal utility vector would unnecessarily restrict its tie-break.

H NASH OPTIMALITY AT THE INDUCED RELEVANCE RADIUS

We use the exact feasible set $\mathcal{C}$ in (13), with fixed positive powers $\mathbf{q}$, $R > 0$, and $\mathbf{p}_i \neq \mathbf{0}$. Set $\mathbf{v}_i = \mathbf{p}_i / \|\mathbf{p}_i\|_2$ and $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_m]$, so $\mathbf{U}(\boldsymbol{\delta}) = \mathbf{1} + \mathbf{V}^\top \boldsymbol{\delta}/R$ and $F(\boldsymbol{\delta}) = \sum_i q_i \log U_i(\boldsymbol{\delta})$. Appendix G establishes compactness and convexity of $\mathcal{C}$ and $1 \leq U_i \leq 2$. The arguments below require no assumption about held-out gradients.

H.1 From log welfare to the constrained worst weighting

Lemma H.1 (Weighted log-utility identity)**.** *For positive utilities and powers,*

$$\sum_i q_i \log U_i = \min_{\mathbf{r}\in\Delta_m} \left\{\log(\mathbf{r}^\top \mathbf{U}) + \mathrm{KL}(\mathbf{q}\|\mathbf{r})\right\}. \tag{H.1}$$

The unique minimizer is $r_i = (q_i/U_i)/(\sum_j q_j/U_j)$.

Proof. For positive $\mathbf{r}$, set $b_i = r_i U_i/(\mathbf{r}^\top \mathbf{U})$. Then

$$\log(\mathbf{r}^\top \mathbf{U}) + \mathrm{KL}(\mathbf{q}\|\mathbf{r}) - \sum_i q_i \log U_i = \mathrm{KL}(\mathbf{q}\|\mathbf{b}) \geq 0.$$

Equality holds exactly at $\mathbf{b} = \mathbf{q}$, which determines the stated normalized weights. A zero coordinate of $\mathbf{r}$ gives infinite divergence. □

For the Nash utility vector, define

$$H_{\mathsf{N}} = \left(\sum_i q_i/U_{\mathsf{N},i}\right)^{-1}, \qquad r_{\mathsf{N},i} = H_{\mathsf{N}} q_i/U_{\mathsf{N},i}, \qquad \rho_{\mathsf{N}} = \mathrm{KL}(\mathbf{q}\|\mathbf{r}_{\mathsf{N}}). \tag{H.2}$$

These definitions give

$$\mathbf{r}_{\mathsf{N}}^\top \mathbf{U}_{\mathsf{N}} = H_{\mathsf{N}}, \qquad \rho_{\mathsf{N}} = F(\boldsymbol{\delta}_{\mathsf{N}}) - \log H_{\mathsf{N}}, \qquad \mathbf{V}\mathbf{r}_{\mathsf{N}} = R H_{\mathsf{N}} \nabla F(\boldsymbol{\delta}_{\mathsf{N}}). \tag{H.3}$$

For any $\mathbf{r} \in \mathcal{R}_{\rho_{\mathsf{N}}}(\mathbf{q})$, the lemma implies

$$\log(\mathbf{r}^\top \mathbf{U}_{\mathsf{N}}) \geq F(\boldsymbol{\delta}_{\mathsf{N}}) - \mathrm{KL}(\mathbf{q}\|\mathbf{r}) \geq F(\boldsymbol{\delta}_{\mathsf{N}}) - \rho_{\mathsf{N}} = \log H_{\mathsf{N}}.$$

The vector $\mathbf{r}_{\mathsf{N}}$ is feasible and attains equality. Moreover, equality forces the unique equality case in Lemma H.1. Thus

$$\min_{\mathbf{r}\in\mathcal{R}_{\rho_{\mathsf{N}}}(\mathbf{q})} \mathbf{r}^\top \mathbf{U}_{\mathsf{N}} = H_{\mathsf{N}}, \tag{H.4}$$

with unique minimizer $\mathbf{r}_{\mathsf{N}}$. For finite radii, the KL sublevel set is compact and excludes zero coordinates, so all these minima exist. When $\rho_{\mathsf{N}} = 0$, the set is $\{\mathbf{q}\}$ and the Nash utilities are equal.

H.2 Proof of Theorem 2

Proof. Since $(\mathbf{V}\mathbf{r})^\top \boldsymbol{\delta} = R(\mathbf{r}^\top \mathbf{U}(\boldsymbol{\delta}) - 1)$, (H.4) gives

$$\mathcal{A}_{\rho_{\mathsf{N}}}(\boldsymbol{\delta}_{\mathsf{N}}) = (\mathbf{V}\mathbf{r}_{\mathsf{N}})^\top \boldsymbol{\delta}_{\mathsf{N}} = R(H_{\mathsf{N}} - 1). \tag{H.5}$$

For every $\boldsymbol{\delta}_A \in \mathcal{C}$, first-order optimality on the convex set $\mathcal{C}$ yields $\nabla F(\boldsymbol{\delta}_{\mathsf{N}})^\top (\boldsymbol{\delta}_A - \boldsymbol{\delta}_{\mathsf{N}}) \leq 0$. The same $\mathbf{r}_{\mathsf{N}}$ is admissible when evaluating $\boldsymbol{\delta}_A$, hence

$$\begin{aligned} \mathcal{A}_{\rho_{\mathsf{N}}}(\boldsymbol{\delta}_{\mathsf{N}}) - \mathcal{A}_{\rho_{\mathsf{N}}}(\boldsymbol{\delta}_A) &\geq (\mathbf{V}\mathbf{r}_{\mathsf{N}})^\top (\boldsymbol{\delta}_{\mathsf{N}} - \boldsymbol{\delta}_A) \\ &= R H_{\mathsf{N}} \nabla F(\boldsymbol{\delta}_{\mathsf{N}})^\top (\boldsymbol{\delta}_{\mathsf{N}} - \boldsymbol{\delta}_A) \geq 0. \end{aligned} \tag{H.6}$$

This proves (17) and (18), including the matched linear allocation. For GRPO, $\boldsymbol{\delta}_{\mathsf{G}} = \mathbf{0}$, so its alignment is zero. Since all $q_i > 0$ and $U_{\mathsf{N},i} \geq 1$, $H_{\mathsf{N}} > 1$ exactly when at least one Nash utility exceeds one, proving (19). □

Distinct corrections need not have distinct worst-case values. Also, the radius is solution-induced: at an externally fixed $\rho = 0$, $\mathcal{A}_0(\boldsymbol{\delta}) = (\mathbf{V}\mathbf{q})^\top \boldsymbol{\delta}$, which is maximized by the matched linear allocation.

H.3 A SUFFICIENT CONDITION FOR A STRICT GAP

Let $\mathcal{C}_0$ be the correction set with only the radius constraint removed. A strict comparison requires an additional condition.

Assumption H.1 (Positive normalized-radius multiplier)**.** *There exist $\lambda_{\mathsf{N}} > 0$ and $\mathbf{z}$ such that*

$$\nabla F(\boldsymbol{\delta}_{\mathsf{N}}) = \frac{\lambda_{\mathsf{N}}}{R^2}\boldsymbol{\delta}_{\mathsf{N}} + \mathbf{z}, \qquad \|\boldsymbol{\delta}_{\mathsf{N}}\|_2 = R, \qquad \mathbf{z}^\top(\boldsymbol{\delta} - \boldsymbol{\delta}_{\mathsf{N}}) \le 0 \quad (\boldsymbol{\delta} \in \mathcal{C}_0). \tag{H.7}$$

The radius constraint is normalized as $(\|\boldsymbol{\delta}\|_2^2/R^2 - 1)/2 \le 0$.

Proposition H.2 (Strict separation)**.** *Under Assumption H.1, every $\boldsymbol{\delta}_A \in \mathcal{C}$ satisfies*

$$\mathcal{A}_{\rho_{\mathsf{N}}}(\boldsymbol{\delta}_{\mathsf{N}}) - \mathcal{A}_{\rho_{\mathsf{N}}}(\boldsymbol{\delta}_A) \ge \frac{\lambda_{\mathsf{N}} H_{\mathsf{N}}}{2R}\|\boldsymbol{\delta}_{\mathsf{N}} - \boldsymbol{\delta}_A\|_2^2. \tag{H.8}$$

Proof. The normal inequality in (H.7) and $\|\boldsymbol{\delta}_A\|_2 \le R$ give

$$\begin{aligned}\nabla F(\boldsymbol{\delta}_{\mathsf{N}})^\top(\boldsymbol{\delta}_{\mathsf{N}} - \boldsymbol{\delta}_A) &\ge \frac{\lambda_{\mathsf{N}}}{2R^2}\left(\|\boldsymbol{\delta}_{\mathsf{N}} - \boldsymbol{\delta}_A\|_2^2 + R^2 - \|\boldsymbol{\delta}_A\|_2^2\right) \\ &\ge \frac{\lambda_{\mathsf{N}}}{2R^2}\|\boldsymbol{\delta}_{\mathsf{N}} - \boldsymbol{\delta}_A\|_2^2. \end{aligned} \tag{H.9}$$

Substitution in (H.6) proves the result. □

The condition is not implied merely by an active radius constraint and is not needed for the non-strict main theorem.

I FROM ADDITIONAL ALIGNMENT TO HELD-OUT ACCURACY

We prove Theorem 3 for the expected sampled accuracy in (6). All comparisons use the same test family $\mathfrak{Q}$, radius ρ, scale α, and representation error e_{tr}.

I.1 SMOOTHNESS AND TEST-FAMILY APPROXIMATION

Lemma I.1 (Two-update smoothness bound)**.** *Suppose f has L_f-Lipschitz gradient on a convex neighborhood containing $\boldsymbol{\theta}$, $\boldsymbol{\theta} + \eta\mathbf{d}_0$, and $\boldsymbol{\theta} + \eta(\mathbf{d}_0 + \boldsymbol{\Delta})$. Then*

$$\left|f(\boldsymbol{\theta} + \eta(\mathbf{d}_0 + \boldsymbol{\Delta})) - f(\boldsymbol{\theta} + \eta\mathbf{d}_0) - \eta\nabla f(\boldsymbol{\theta})^\top\boldsymbol{\Delta}\right| \le L_f\eta^2\|\boldsymbol{\Delta}\|_2\left(\|\mathbf{d}_0\|_2 + \tfrac{1}{2}\|\boldsymbol{\Delta}\|_2\right). \tag{I.1}$$

Proof. Integrating along the segment between the updated points gives the remainder

$$\eta\int_0^1 [\nabla f(\boldsymbol{\theta} + \eta\mathbf{d}_0 + t\eta\boldsymbol{\Delta}) - \nabla f(\boldsymbol{\theta})]^\top\boldsymbol{\Delta}\, dt.$$

Its absolute value is bounded by $L_f\eta^2\|\boldsymbol{\Delta}\|_2\int_0^1(\|\mathbf{d}_0\|_2 + t\|\boldsymbol{\Delta}\|_2)\,dt$, proving the claim. □

The two parts of Assumption 1 give, for every vector $\boldsymbol{\Delta}$,

$$\left|\inf_{Q\in\mathfrak{Q}}\mathbf{h}_Q^\top\boldsymbol{\Delta} - \alpha\mathcal{A}_\rho(\boldsymbol{\Delta})\right| \le e_{\mathrm{tr}}\|\boldsymbol{\Delta}\|_2. \tag{I.2}$$

For the lower bound, approximate each test gradient by a relevance mixture and apply Cauchy–Schwarz. For the upper bound, approximate a minimizing mixture by a test gradient. The definition of $\mathcal{A}_\rho$ extends to arbitrary $\boldsymbol{\Delta}$ by the same linear minimization. Only the first coverage condition is needed for the lower bound; both are needed to compare two worst-case gains using matching lower and upper bounds.

I.2 Proof of Theorem 3

Proof. Apply Lemma I.1 to $f = \mathrm{Acc}_Q$, $\mathbf{d}_0 = \mathbf{g}$, and $\mathbf{\Delta} = \boldsymbol{\delta}_M$ for $M \in \{\mathsf{N}, \mathsf{L}\}$. Assumption 2 makes the remainder bound uniform in Q. Taking an infimum preserves this error bound, and (I.2) therefore gives

$$|\mathcal{I}_\eta(M; \mathsf{G}) - \eta\alpha\mathcal{A}_\rho(\boldsymbol{\delta}_M)| \le E_M, \tag{I.3}$$

where E_M is (24). Its lower bound for Nash proves (22). Subtracting its upper bound for linear gives

$$\mathcal{I}_\eta(\mathsf{N}; \mathsf{G}) - \mathcal{I}_\eta(\mathsf{L}; \mathsf{G}) \ge \eta\alpha\left[\mathcal{A}_\rho(\boldsymbol{\delta}_\mathsf{N}) - \mathcal{A}_\rho(\boldsymbol{\delta}_\mathsf{L})\right] - E_\mathsf{N} - E_\mathsf{L},$$

which is (23). □

At $\rho = \rho_\mathsf{N}$, Theorem 2 supplies the nonnegative alignment comparisons. Positivity of an accuracy lower bound still requires its leading term to exceed the representation and smoothness errors. In particular, sufficiently small η yields a positive Nash–linear bound only if

$$\alpha\left[\mathcal{A}_\rho(\boldsymbol{\delta}_\mathsf{N}) - \mathcal{A}_\rho(\boldsymbol{\delta}_\mathsf{L})\right] > e_{\mathrm{tr}}\left(\|\boldsymbol{\delta}_\mathsf{N}\|_2 + \|\boldsymbol{\delta}_\mathsf{L}\|_2\right).$$

Small step size alone does not make a zero alignment gap strict.

The theorem compares two worst-case advantages over GRPO, not a worst-case direct advantage over linear:

$$\mathcal{I}_\eta(\mathsf{N}; \mathsf{L}) \ne \mathcal{I}_\eta(\mathsf{N}; \mathsf{G}) - \mathcal{I}_\eta(\mathsf{L}; \mathsf{G}) \quad \text{in general.}$$

The minimizing test distributions can differ, so this distinction cannot be removed when interpreting the bound.

J Rollout reuse, minibatches, and optimizer scope

The calculus in Appendix F conditions on the sampled group; it does not require a fresh sample at each inner iterate. During rollout reuse, the old probabilities and advantages stay fixed, while the current ratios, clipping factors, gradients, and NCR coefficients are refreshed. The resulting identities remain exact for that cached group. They do not make a later iterate independent of the responses that produced it or supply a population or finite-horizon guarantee.

Minibatch aggregation. For K_b prompt groups, write their baseline gradients and corrections as $\mathbf{g}_b, \boldsymbol{\delta}_b$. The averaged gradient and correction satisfy

$$\left(\frac{1}{K_b}\sum_b \mathbf{g}_b\right)^\top \left(\frac{1}{K_b}\sum_k \boldsymbol{\delta}_k\right) = \frac{1}{K_b^2}\sum_{b,k} \mathbf{g}_b^\top \boldsymbol{\delta}_k.$$

Per-group anchor constraints control only the terms with $b = k$. Likewise, protecting a response against its own group's correction does not control its inner products with other groups' corrections. Thus the per-group theorems are not automatically guarantees for the averaged update.

Applied parameter updates. The theory concerns direct steps along $\mathbf{g} + \boldsymbol{\delta}$. Momentum, adaptive preconditioning, weight decay, and gradient clipping can change the applied displacement. Any guarantee for that displacement requires checking its relevant margins, not just the input gradient. Similarly, smooth dependence on a finite sampled group cannot be assumed at clipping or power-selection degeneracies (Appendix M.5). Local smoothness arguments require regularity of both coefficient optimizations; no such global assumption is used in the main fixed-iterate results.

K Bound on the Nash-induced relevance radius

Proposition K.1 (Radius bound from shifted utilities)**.** *For* $1 \le U_{\mathsf{N},i} \le 2$*, the radius in* (16) *satisfies*

$$0 \le \rho_\mathsf{N} \le 2\log 2 - 1 - \log(2\log 2). \tag{K.1}$$

Proof. Let $X = 1/U_{\mathsf{N},i} \in [1/2, 1]$ with probability q_i, and set $z = \mathbb{E}X$. Equation (H.2) gives $\rho_{\mathsf{N}} = \log z - \mathbb{E}\log X$. The convex function $-\log X$ lies below the chord between its endpoint values, so $-\mathbb{E}\log X \le 2(1-z)\log 2$. Maximizing $\log z + 2(1-z)\log 2$ over $z \in [1/2, 1]$ gives $z = 1/(2\log 2)$ and the upper bound. Nonnegativity follows from KL divergence. □

This bounds the solution-induced relevance neighborhood; it does not establish Nash optimality at an arbitrary prescribed radius.

L Evaluation measurements and support diagnostic

The trained-checkpoint benchmark results appear in Table 1, with settings in Appendices B and C. The local diagnostic below uses a separate public checkpoint and does not apply a training update.

L.1 Expected accuracy and its score estimator

For the target in (6), take Q uniform over held-out prompts $z_1, \dots, z_M$, and draw K_e independent responses per prompt from the current policy. The sample estimators are

$$\widehat{\mathrm{Acc}}_Q = \frac{1}{MK_e}\sum_{b,k} R(z_b, Y_{b,k}), \qquad \widehat{\mathbf{h}}_Q = \frac{1}{MK_e}\sum_{b,k} R(z_b, Y_{b,k})\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(Y_{b,k} \mid z_b). \tag{L.1}$$

Lemma L.1 (Accuracy-gradient identity)**.** *If the response support is locally fixed, R is parameter-independent, and differentiation can be interchanged with expectation, then*

$$\mathbf{h}_Q = \mathbb{E}_{z\sim Q, Y\sim\pi_{\boldsymbol{\theta}}(\cdot|z)}[R(z,Y)\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(Y \mid z)]. \tag{L.2}$$

Proof. Differentiate $\sum_Y R(z,Y)\pi_{\boldsymbol{\theta}}(Y \mid z)$, substitute $\nabla\pi_{\boldsymbol{\theta}} = \pi_{\boldsymbol{\theta}}\nabla\log\pi_{\boldsymbol{\theta}}$, and average over Q. □

The sequence score is a token sum, not the response mean used by GRPO (Williams, 1992). Thus (L.1) uses unstandardized rewards and all draws, including failures, in the denominator. The sampled distribution must match the differentiated policy; this identity does not justify mixing a temperature-modified sampler with unmodified policy scores.

L.2 Diagnostic B: support versus usefulness

Checkpoint and samples. Figure 1 uses the public `Qwen/Qwen3-4B` checkpoint in non-thinking mode (`enable_thinking=false`). We select 500 training prompts before observing rewards and use the 500-prompt MATH500 test split of `ShuoZheLi/MATH-train-MATH500-test`.[1] The selection excludes four invalid gold answers; normalized exact-text train/test overlap is zero. Each split has 32 responses per prompt, sampled with temperature 1, top-$p = 1$, no top-k or minimum-p truncation, a 30,000-token response cap, and master seed 42. All recorded responses terminate before the cap. Boxed answers are checked with `math_verify`. The diagnostic uses the system instruction "Please reason step by step, and put your final answer within `\boxed{}`", followed by the problem as the user message.

Of the 500 training groups, 266 are all-correct, 23 all-incorrect, and 211 mixed. Homogeneous groups have zero advantages and are excluded. The shared analysis population is all 4,340 correct responses from the mixed groups; there are no zero-norm or nonfinite-feature exclusions.

Directions and measurements. Use the selected-token proxy (D.1) in $\mathbb{R}^{2560}$. At this checkpoint $\rho_{j,t} = 1$, so raw and clipping-weighted means agree for responses in mixed groups. In this diagnostic, the main-text symbols $\mathbf{p}_i, \mathbf{v}_i, \mathbf{d}_A$ denote their proxy counterparts, with

$$\mathbf{p}_i = \frac{1}{T_i}\sum_t \mathbf{u}^{\mathrm{prx}}_{i,t}, \qquad \mathbf{v}_i = \frac{\mathbf{p}_i}{\|\mathbf{p}_i\|_2}. \tag{L.3}$$

[1] Model/tokenizer revision: `1cfa9a7208912126459214e8b04321603b3df60c`; dataset revision: `09ad2162c6f3a07ea71e48d7af02922b093df041`.

Table 6: Support–usefulness correlations on the shared Qwen3-4B responses. The within-prompt column uses the 4,335 responses from 206 groups with at least two successes.

Allocation strategy	Spearman	Pearson	Within-prompt rank
GRPO	0.0007	−0.0229	−0.2332
Symmetric bargaining	0.1062	0.0869	−0.2010
NCR	0.2512	0.2437	−0.1049
Linear routing	0.1322	0.1233	−0.1518

The common held-out reward-weighted reference is

$$\widehat{\mathbf{h}}_Q = \frac{1}{500 \times 32} \sum_{b=1}^{500} \sum_{k=1}^{32} R_{b,k} \sum_t \mathbf{u}_{b,k,t}^{\mathrm{prx}}. \tag{L.4}$$

This replaces the full score in (L.1) by the proxy; it is not an unbiased estimator of the full-policy gradient. There are 13,427 successful held-out draws; all 16,000 draws remain in the denominator. Held-out usefulness and allocation support are

$$B_i = \mathbf{v}_i^\top \widehat{\mathbf{h}}_Q, \qquad S_i^{(A)} = \mathbf{v}_i^\top \mathbf{d}_A. \tag{L.5}$$

The same B_i is used for every allocation strategy. Held-out data enter only this measurement, never power selection or credit allocation.

Allocations and solver. GRPO uses $\mathbf{d}_{\mathsf{G}} = \mathbf{g}$. NCR uses (9) and the collision-minimized powers. Symmetric bargaining changes only the powers to $q_i = 1/m$; matched linear routing instead maximizes $\sum_i q_i U_i$ with NCR's powers. All routed strategies use the same feasible set and minimum-correction-norm tie-break. Here $R = 0.4\|\mathbf{g}\|_2$, $\varepsilon = 10^{-6}$, the reward standard deviation uses divisor G, and there is no reference-policy KL term. Failed responses are split at blank lines and merged adjacently to at most 16 segments per response, retaining all tokens and their original prefixes.

The recorded solves use CLARABEL with internal tolerance 10^{-9} and at most 500 iterations. External feasibility and objective tolerances are 5×10^{-6} and 10^{-8}; the power-gap tolerance is 2×10^{-7}. Feature accumulation is FP32, with 128-token chunks and an 8,192-token vocabulary-logit chunk. Failed solves retain their observations using GRPO fallback rather than removing them.

Descriptive results and controls. Table 6 reports pooled Spearman and Pearson correlations, giving each response equal weight and averaging ranks for ties. The final column takes within-prompt average ranks, transforms each by $(\mathrm{rank} - (m+1)/2)/m$, and reports the pooled Pearson correlation of transformed support and usefulness for groups with $m \geq 2$. It remains negative for every strategy: the pooled result does not establish positive within-prompt alignment.

Restricting the pooled Spearman calculation to those same 4,335 responses gives 0.0019, 0.1075, 0.2526, and 0.1335, respectively. Every routed strategy satisfies $S_i^{(A)} - S_i^{(\mathsf{G})} \geq 0$ for all 4,340 responses. This verifies only the recorded scalar proxy-support check, not the full group-level constraints, whose audit requires the per-group correction and solver records.

No policy update is applied. These measurements concern local proxy alignment, not post-training accuracy or the test-family representation assumption in Theorem 3.

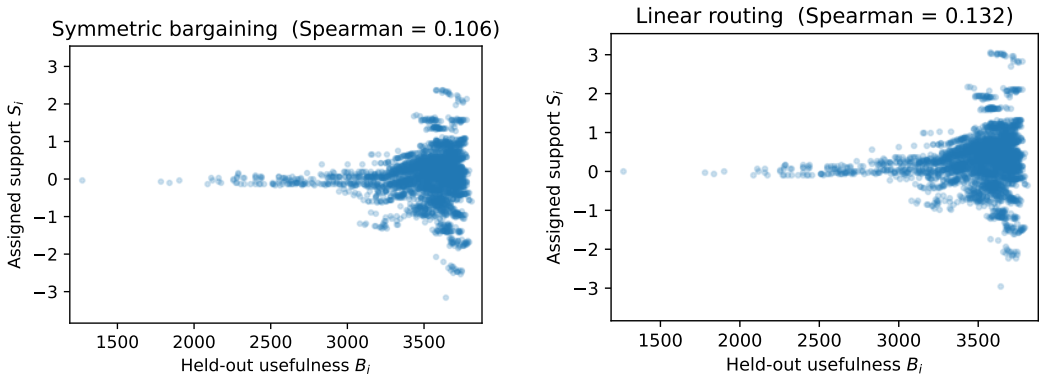

Figure 4: Auxiliary allocation controls on the same 4,340 responses: symmetric bargaining (left) and matched linear routing (right). Axes match the GRPO and NCR panels in Fig. 1.

M COLLISION-MINIMIZED BARGAINING POWERS

This appendix proves the power-selection claims in Section 4.3. Fix the normalized directions $\mathbf{v}_i$, let $K_{ij} = (1 + \mathbf{v}_i^\top \mathbf{v}_j)/2$, and write $C(\mathbf{q}) = C(\mathbf{B}, \mathbf{q})$ for $\mathbf{B} = \mathbf{K} \odot \mathbf{K}$.

M.1 CONVEXITY, POSITIVE POWERS, AND DUPLICATE AGGREGATION

The vectors $\boldsymbol{\phi}_i = (1, \mathbf{v}_i^\top)^\top/\sqrt{2}$ satisfy $K_{ij} = \boldsymbol{\phi}_i^\top \boldsymbol{\phi}_j$ and $B_{ij} = \langle \boldsymbol{\phi}_i \otimes \boldsymbol{\phi}_i, \boldsymbol{\phi}_j \otimes \boldsymbol{\phi}_j \rangle$. Thus both matrices are positive semidefinite, have unit diagonal, and have entries in $[0, 1]$. The probability-weighted kernel $\widetilde{\mathbf{K}}_{\mathbf{q}} = \operatorname{diag}(\sqrt{\mathbf{q}})\mathbf{K}\operatorname{diag}(\sqrt{\mathbf{q}})$ has nonnegative eigenvalues summing to one, with

$$\operatorname{tr}(\widetilde{\mathbf{K}}_{\mathbf{q}}^2) = \sum_{i,j} q_i q_j K_{ij}^2 = C(\mathbf{q}). \tag{M.1}$$

Consequently, $1/C(\mathbf{q})$ is its order-2 effective number (Leinster & Cobbold, 2012; Friedman & Dieng, 2023; Pasarkar & Dieng, 2024).

Proposition M.1 (Existence and uniqueness of the selected powers)**.** *For $m \geq 2$, collision minimization over $\mathcal{Q}_m$ has a nonempty compact convex minimizer set. The minimum-variance tie-break selects a unique strictly positive vector.*

Proof. The quadratic is convex because $\mathbf{B} \succeq 0$. The set $\mathcal{Q}_m$ is a compact convex polytope containing $\mathbf{1}/m$. If a feasible $q_j = 0$, the ratio constraints force every $q_i = 0$, a contradiction. Continuity gives a first-stage minimum. On its convex minimizer set, minimizing variance is equivalent to minimizing $\|\mathbf{q} - \mathbf{1}/m\|_2^2$, which is strictly convex. □

Duplicate aggregation. For $i \in G_a$ and $j \in G_b$, identical directions give $B_{ij} = B_{ab}^d$. Collecting the corresponding terms yields

$$C(\mathbf{B}, \mathbf{q}) = \sum_{a,b} B_{ab}^d \left(\sum_{i \in G_a} q_i \right) \left(\sum_{j \in G_b} q_j \right) = \sum_{a,b} B_{ab}^d t_a t_b = C(\mathbf{B}^d, \mathbf{t}).$$

□

This identity concerns the objective at fixed aggregate powers. Duplicating a response can still change the per-response feasible set and tie-break.

M.2 RATIO-CAP CALIBRATION AND COORDINATE BOUNDS

Proposition M.2 (Smallest cap for two-direction balancing)**.** *For $m \geq 2$, $L = m - 1$ is the smallest pairwise cap $q_i \leq L q_j$ that permits the collision-minimizing equal aggregate powers for every pair of distinct directions with positive multiplicities summing to m.*

Proof. Let the multiplicities be n_1, n_2, and let $b < 1$ be their squared cross-similarity. For aggregate power t on the first direction,

$$C(t) = t^2 + (1 - t)^2 + 2bt(1 - t) = \frac{1 + b}{2} + 2(1 - b)(t - 1/2)^2.$$

The minimizer is $t = 1/2$. Equal sharing within each direction gives per-copy powers $1/(2n_1)$ and $1/(2n_2)$, whose ratio is at most $m - 1$. Conversely, for multiplicities $(m - 1, 1)$ the singleton needs power $1/2$; the cap requires every other copy to have at least $1/(2L)$. Their total can be $1/2$ only if $L \geq m - 1$. □

Proposition M.3 (Coordinate bounds)**.** *For $m \geq 2$, every $\mathbf{q} \in \mathcal{Q}_m$ satisfies*

$$\frac{1}{1 + (m - 1)^2} \leq q_i \leq \frac{1}{2}. \tag{M.2}$$

Proof. Summing $q_j \le (m-1)q_i$ over $j \ne i$ gives $1 \le [1+(m-1)^2]q_i$. Summing the reverse bounds $q_j \ge q_i/(m-1)$ gives $1 \ge 2q_i$. □

For $m = 2$ the powers are uniform; for $m = 1$ the separate convention is $q_1 = 1$ rather than imposing the ratio cap.

M.3 Comparison with uniform and inverse-collision powers

Define the comparators

$$\chi_i = \frac{1}{m}\sum_j B_{ij}, \qquad q_i^{\mathrm{inv}} = \frac{\chi_i^{-1}}{\sum_j \chi_j^{-1}}, \qquad \mathbf{q}^{\mathrm{unif}} = \mathbf{1}/m. \tag{M.3}$$

Proposition M.4 (Collision dominance)**.** *For $m \ge 2$, $\mathbf{q}^{\mathrm{inv}} \in \mathcal{Q}_m$ and*

$$C(\mathbf{q}_*) \le C(\mathbf{q}^{\mathrm{inv}}) \le C(\mathbf{q}^{\mathrm{unif}}). \tag{M.4}$$

The first inequality is strict exactly when $\mathbf{q}^{\mathrm{inv}}$ is not a collision minimizer; the second is strict when the χ_i are not all equal.

Proof. For $i \ne j$, symmetry and the entry bounds imply $\chi_j/\chi_i \le (m-1+B_{ij})/(1+B_{ij}) \le m-1$. This proves feasibility of the inverse powers. Also,

$$m\sum_i \chi_i q_i^2 - C(\mathbf{q}) = \sum_{i<j} B_{ij}(q_i - q_j)^2 \ge 0. \tag{M.5}$$

The diagonal quadratic has unique simplex minimizer $\mathbf{q}^{\mathrm{inv}}$ and equals C at uniform powers. Minimizing it gives the second inequality, strictly when its minimizer is not uniform. Minimizing C over $\mathcal{Q}_m$ gives the first inequality and its equality condition. □

This is a comparison of collision values, not held-out accuracy.

M.4 Duplicate-direction powers

Proposition M.5 (Aggregate power on repeated directions)**.** *For two distinct directions, each receives aggregate power $1/2$ and each copy of direction c receives $1/(2n_c)$. For $K \ge 2$ mutually orthogonal directions, the corresponding powers are $1/K$ and $1/(Kn_c)$.*

Proof. The two-direction result follows from Proposition M.2. Permuting identical copies preserves collision and feasibility, so uniqueness of the minimum-variance solution forces equal within-direction powers. For orthogonal directions, cross-direction entries of $\mathbf{B}$ equal $1/4$. Writing their aggregate powers as Q_c gives $C(\mathbf{q}) = 1/4 + (3/4)\sum_c Q_c^2$, minimized by $Q_c = 1/K$. Equal sharing satisfies the cap because $n_{\max}/n_{\min} \le m-1$. □

M.5 Exact tie-break and stability

Lemma M.6 (First-stage minimizer set)**.** *If $\mathbf{q}_0$ minimizes collision on $\mathcal{Q}_m$, its complete minimizer set is $\{\mathbf{q} \in \mathcal{Q}_m : \mathbf{B}\mathbf{q} = \mathbf{B}\mathbf{q}_0\}$.*

Proof. For two minimizers, the quadratic is constant along their connecting segment. Hence $(\mathbf{q}-\mathbf{q}_0)^\top \mathbf{B}(\mathbf{q}-\mathbf{q}_0) = 0$; positive semidefiniteness gives $\mathbf{B}(\mathbf{q}-\mathbf{q}_0) = \mathbf{0}$. The converse follows by substitution. □

Thus the second stage is a strictly convex quadratic program with linear constraints. A feasible approximate $\widehat{\mathbf{q}}$ has the certificate

$$0 \le C(\widehat{\mathbf{q}}) - C_{\min} \le \nabla C(\widehat{\mathbf{q}})^\top \widehat{\mathbf{q}} - \min_{\mathbf{q}\in\mathcal{Q}_m} \nabla C(\widehat{\mathbf{q}})^\top \mathbf{q}.$$

This controls objective error, not distance to the selected powers.

Degenerate directions. Two distinct directions with unequal multiplicities n_1, n_2 have per-copy powers $1/(2n_c)$. As the directions merge, these powers remain unchanged until exact coincidence, where collision is constant and the tie-break selects $1/m$ for every copy. The power map is therefore not globally continuous. Nevertheless,

$$|C_{\min}(\mathbf{B}) - C_{\min}(\mathbf{B}')| \leq \max_{i,j} |B_{ij} - B'_{ij}|,$$

because the same bound holds for each simplex vector before minimization. A jump in powers alone does not prove a jump in the final correction.

N LIMITATIONS

NCR operates on the response groups observed during training, and therefore its credit allocation is limited to the reasoning directions represented in those samples. For efficiency, we construct the routing geometry using the last-layer proxy described in Appendix D, rather than explicit full-parameter gradients. Moreover, the theoretical connection from local gradient support to held-out accuracy relies on the transfer condition stated in Section 5. Our experiments focus primarily on mathematical reasoning with Qwen models; evaluating the same mechanism across broader tasks and model families is a natural direction for future work.